Generalization

We want to show that

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \frac{1}{n} \sum \ell(\theta; Z_i)$$

achieves near-optimal population loss.

Again, our goal is to show an optimality guarantee for ERM

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum \ell(\theta; Z_i)$$

Now, we will use bdd diff to show the following uniform concentration result:

$$\Delta_n := \sup_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i) - \mathbb{E}_P \ell(\theta; Z), \quad \bar{\Delta}_n := \sup_{\theta \in \Theta} \mathbb{E}_P \ell(\theta; Z) - \frac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i)$$

are small w.h.p.

why is this useful?

$$\begin{aligned}
\mathbb{E}\ell(\hat{\theta}_n; Z) &\le \frac{1}{n}\sum \ell(\hat{\theta}_n; Z_i) + \bar{\Delta}_n && \text{by def of } \bar{\Delta}_n \\
&\le \frac{1}{n}\sum \ell(\theta; Z_i) + \bar{\Delta}_n && \text{by def of } \hat{\theta}_n, \text{ for any arbitrary } \theta \in \Theta \\
&\le \mathbb{E}\ell(\theta; Z) + \bar{\Delta}_n + \Delta_n && \text{by def of } \Delta_n
\end{aligned}$$

Taking inf over θ, we get

$$\mathbb{E}\ell(\hat{\theta}_n; Z) \le \inf_{\theta \in \Theta} \mathbb{E}\ell(\theta; Z) + \bar{\Delta}_n + \Delta_n$$

So if ε_n is small, then $\hat{\theta}_n$ is near-optimal.

We will focus on finite-sample results today. Traditionally, ML guarantees are finite sample since it allows quantifying dimension dependence!

This is useful for high-dim, large-scale models.

We proceed in two parts to bound ε_n & $\bar{\varepsilon}_n$. As we'll see, the case for $\bar{\varepsilon}_n$ is symmetric, so we focus on Δ_n below.

Bounded differences

Bounded differences will play a key role in showing Δ_n is small.

one coordinate doesn't change ftn too much

Thm Let g be a function satisfying

$$|g(z_1, \cdots z_i, \cdots, z_n) - g(z_1, \cdots, z_i', \cdots, z_n)| \leq c_i \qquad \forall\, 1 \leq i \leq n$$

For independent RVs Z_i's,

$$\mathbb{P}\left(g(Z_1^n) - \mathbb{E}g(Z_1^n) \geq t \right) \leq \exp\left(- \frac{2t^2}{\sum_1^n c_i^2}\right).$$

Generalization of Hoeffding bound

Assumption $\ell(\theta; z) \in [0, M]$

Part I We show Σ_n is concentrated around its mean w.h.p.

Define $g(z_1, \cdots, z_n) := \sup_{\theta \in \Theta} \frac{1}{n}\sum \ell(\theta; z_i) - \mathbb{E}\ell(\theta; Z_i)$ so that $g(Z_1^n) = \Delta_n$. We'll apply bdd diff.

As a notational shorthand, we use $\hat{P}_n(\cdot) = \frac{1}{n}\sum \mathbb{1}\{z_i \in \cdot\}$, and write $Q\ell(\theta; z) = \mathbb{E}_{Z \sim Q}\, \ell(\theta; z)$.

$$|g(z_1, \cdots, z_i, \cdots, z_n) - g(z_1, \cdots, z_i', \cdots, z_n)|$$

$$= \left| \sup_{\theta \in \Theta} \left\{ \frac{1}{n}\sum \ell(\theta; z_i) - \mathbb{E}\ell(\theta; Z) \right\} - \sup_{\theta \in \Theta} \left\{ \frac{1}{n}\sum \ell(\theta; z_i) - \mathbb{E}\ell(\theta; Z) - \frac{1}{n}\ell(\theta; z_i) + \frac{1}{n}\ell(\theta; z_i') \right\} \right| \leq \frac{2M}{n}$$

From bdd differences, $\mathbb{P}(\Delta_n - \mathbb{E}\Delta_n \geq t) \leq \exp\left(-\frac{nt^2}{2M^2}\right)$. Equivalently,

$$\Delta_n \leq \mathbb{E}\Delta_n + M\sqrt{\frac{2t}{n}} \qquad \text{w.p.} \geq 1 - e^{-t}$$

So now, it suffices to control $\mathbb{E}\Delta_n$!

We begin with concentration results for light-tailed RVs.

<u>Def</u> A RV X is σ^2-sub Gaussian if $\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \le \underline{\exp\left(\frac{\sigma^2}{2}\lambda^2\right)}$ $\forall \lambda \in \mathbb{R}$

↳ light-tailed RV

↳ MGF of $N(0,\sigma^2)$

From Markov's inequality, $\forall \lambda \ge 0$

$$\mathbb{P}(X-\mathbb{E}X \ge t) = \mathbb{P}(\lambda(X-\mathbb{E}X) \ge \lambda t) = \mathbb{P}(e^{\lambda(X-\mathbb{E}X)} \ge e^{\lambda t}) \le e^{-\lambda t}\mathbb{E}e^{\lambda(X-\mathbb{E}X)}$$
$$\le \exp\left(\frac{\sigma^2}{2}\lambda^2 - \lambda t\right)$$

Taking min over $\lambda \ge 0$, we get $\mathbb{P}(X-\mathbb{E}X \ge t) \le \exp\left(-\frac{t^2}{2\sigma^2}\right)$.

Similarly, we have $\mathbb{P}(X-\mathbb{E}X \le -t) \le \exp\left(-\frac{t^2}{2\sigma^2}\right)$.

<u>Example</u> ε : random signs (Rademacher) is 1-sub Gaussian.

$$\mathbb{E}e^{\lambda\varepsilon} = \frac{1}{2}(e^{-\lambda}+e^{\lambda}) = \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\right) = \sum_{k\ge 0}\frac{\lambda^{2k}}{(2k)!} \le 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{(2k)!}$$
$$\le 1 + \sum_{k=1}^{\infty}\frac{\lambda^{2k}}{2^k k!} = e^{\frac{\lambda^2}{2}}$$

<u>Example</u> $X \in [a,b]$, $\mathbb{E}X = 0$. Then, X is $(b-a)^2$-sub-Gaussian.

Pf) $$\mathbb{E}_X e^{\lambda X} = \mathbb{E}_X e^{\lambda(X-\mathbb{E}_{X'}X')} \underset{\text{Jensen}}{\le} \mathbb{E}e^{\lambda(X-X')}$$ where X' indep copy of X

Let ε be random signs indep of everything so that $X-X' \overset{D}{=} \varepsilon(X-X')$

$$\mathbb{E}e^{\lambda(X-X')} = \mathbb{E}_{X,X'}\mathbb{E}_{\varepsilon}e^{\lambda\varepsilon(X-X')} \le \mathbb{E}_{X,X'}e^{\frac{\lambda^2}{2}(X-X')^2}$$ by Ex 1
$$\le e^{\frac{\lambda^2}{2}(b-a)^2}$$

Actually, we can show X is $\frac{(b-a)^2}{4}$-sub-G.

cf. $X \in [a,b]$ is $\frac{(b-a)^2}{4}$-sub-Gaussian.

By convexity of $x \mapsto e^{\lambda x}$, $e^{\lambda x} \le \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}$

Take expectations on both sides. For $h = \lambda(b-a)$, $p = \frac{-a}{b-a}$, $L(h) = -hp + \log(1-p+pe^h)$,

$\mathbb{E}e^{\lambda X} \le e^{L(h)}$. $L(0) = L'(0) = 0$, $L''(h) \le \frac{1}{4}$ $\forall h$ so $L(h) \le \frac{1}{8}h^2$ by Taylor ☒.

We're ready to show bdd differences now.

Thm Let g be a function satisfying

one coordinate doesn't change ftn too much

$$|g(z_1,\cdots,z_i,\cdots,z_n) - g(z_1,\cdots,z_i',\cdots,z_n)| \le c_i \qquad \forall\, 1\le i\le n$$

For independent RVs Z_i's,

$$\mathbb{P}\left(g(Z_1^n) - \mathbb{E}g(Z_1^n) \ge t \right) \le \exp\left(-\frac{2t^2}{\sum_1^n c_i^2}\right).$$

Generalization of Hoeffding bound

Def $\{M_i\}_{i=0}^n$ is a martingale seq w.r.t. RVs $Z_1,\cdots,Z_n$

if M_i is $(Z_1,\cdots,Z_i)$-measurable, $\mathbb{E}|M_i|<\infty$, and

$\mathbb{E}[M_i \mid Z_{i-1}] = M_{i-1}$.

We call $\{D_i := M_i - M_{i-1}\}_{i=1}^n$ a martingale difference sequence w.r.t. Z_1^n.

$(\mathbb{E}[D_i \mid Z_1^{i-1}] = 0)$

Lemma Let D_i be a martingale difference sequence w.r.t. Z_1^n s.t. $\exists\, \sigma_i^2$

$$\mathbb{E}[e^{\lambda D_i} \mid Z_1^{i-1}] \le \exp\left(\frac{\sigma_i^2 t^2}{2}\right) \qquad \forall i \qquad \cdots (\star)$$

Then, $M_n - M_0 = \sum_{i=1}^n D_i$ is $(\sum \sigma_i^2)$-sub-Gaussian.

Pf

$$\mathbb{E}e^{\lambda \sum D_i} = \mathbb{E}\, e^{\lambda D_n} \cdot e^{\lambda \sum_1^{n-1} D_i} = \mathbb{E}\left[\mathbb{E}\left[e^{\lambda D_n} \cdot \underline{e^{\lambda \sum_1^{n-1} D_i}} \mid Z_1^{n-1} \right]\right]$$

$$\le \exp\left(\frac{\sigma_n^2 t^2}{2}\right) \cdot \mathbb{E}\left[e^{\lambda \sum^{n-1} D_i}\right]$$

$\hookrightarrow Z_1^{n-1}$-measurable

By induction, we get the result. $\boxtimes$.

__Proof of bdd differences__

Define the Doob martingale $M_i = \mathbb{E}[g(Z_1^n) \mid Z_1^i]$ $\quad \left(\begin{matrix} M_0 = \mathbb{E}g(Z_1^n) \\ M_n = g(Z_1^n) \end{matrix} \right)$

So we'd like to bound $\mathbb{P}(M_n - M_0 \ge t)$.

Note that

$$|D_i| = |\mathbb{E}[g(Z_1^n) \mid Z_1^i] - \mathbb{E}[g(Z_1^n) \mid Z_1^{i-1}]|$$

$$\le \sup_{z,z'} |\mathbb{E}_{Z_{i+1}^n} g(Z_1^{i-1}, z, Z_{i+1}^n) - \mathbb{E}_{Z_{i+1}^n} g(Z_1^{i-1}, z', Z_{i+1}^n)| \le c_i \qquad \cdots (\ast)$$

So

$$\mathbb{E}[e^{\lambda D_i} \mid Z_1^{i-1}] = \mathbb{E}[e^{\lambda(D_i - \mathbb{E}[D_i \mid Z_1^{i-1}])} \mid Z_1^{i-1}] \le \exp\left(\frac{\lambda^2}{2} \cdot \frac{c_i^2}{4}\right)$$

From previous lemma, and tail inequality for sub-G RVs, we have the result $\boxtimes$.

Part II We bound $\mathbb{E}\Delta_n$ via symmetrization. cf. See VWW Ch. 2.2-3, 2.14 for more.

↳ This is tricky to bound. Think about how you would approach this.

Let $Z_1', \dots, Z_n'$ be indep copies of $Z_1, \dots, Z_n$.

$$\mathbb{E}\Delta_n = \mathbb{E}\left[\sup_{\theta\in\Theta} \frac{1}{n}\sum \ell(\theta;Z_i) - \mathbb{E}\left[\frac{1}{n}\sum \ell(\theta;Z_i') \,\middle|\, Z_1^n\right]\right] \le \mathbb{E}\sup_{\theta\in\Theta} \frac{1}{n}\sum \left(\ell(\theta;Z_i) - \ell(\theta;Z_i')\right)$$

Let ε_i be i.i.d. random signs (Rademacher RVs), indep of everything else.

From $\varepsilon_i(\ell(\theta;Z_i) - \ell(\theta;Z_i')) \overset{D}{=} \ell(\theta;Z_i) - \ell(\theta;Z_i')$,

$$\begin{aligned}\mathbb{E}\sup_{\theta\in\Theta} \frac{1}{n}\sum \left(\ell(\theta;Z_i) - \ell(\theta;Z_i')\right) &= \mathbb{E}\sup_{\theta\in\Theta} \frac{1}{n}\sum \varepsilon_i\left(\ell(\theta;Z_i) - \ell(\theta;Z_i')\right)\\ &\le \mathbb{E}\sup_{\theta\in\Theta} \frac{1}{n}\sum \varepsilon_i \ell(\theta;Z_i) + \mathbb{E}\sup_{\theta\in\Theta} \frac{1}{n}\sum (-\varepsilon_i)\ell(\theta;Z_i)\\ &= 2\,\mathbb{E}\sup_{\theta\in\Theta} \frac{1}{n}\sum \varepsilon_i \ell(\theta;Z_i)\end{aligned}$$

Def The (empirical) Rademacher complexity of a class $\mathcal{H}$ of functions $h: \mathcal{Z} \to \mathbb{R}$ is

$$\mathcal{R}_n\mathcal{H} := \mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(Z_i) \,\middle|\, Z_1^n\right]$$

↳ Interpretation: How well can $\mathcal{H}$ fit random noise σ_i's? (where $\sigma_i h(Z_i)$ is the margin)

Note that $\mathcal{R}_n\mathcal{H} = \mathcal{R}_n(-\mathcal{H})$. So the case for $\bar{\Delta}_n$ is symmetric.

Collecting bounds in Parts I & II, we arrive at

$$\Delta_n \le 2\mathbb{E}\mathcal{R}_n\mathcal{H} + M\sqrt{\frac{t}{2n}}, \quad \bar{\Delta}_n \le 2\mathbb{E}\mathcal{R}_n\mathcal{H} + M\sqrt{\frac{2t}{n}} \quad \text{w.p.} \ge 1 - 2e^{-t}.$$

So we conclude

$$\mathbb{E}\ell(\hat{\theta}_n;Z) \le \inf_{\theta\in\Theta} \mathbb{E}\ell(\theta;Z) + 4\mathbb{E}\mathcal{R}_n\mathcal{H} + 2M\sqrt{\frac{2t}{n}} \quad \text{w.p.} \ge 1 - 2e^{-t} \;//.$$

Basic properties of Rademacher complexity: This will be useful for HW2.

1) Contraction Principle: Let ϕ be a C_ϕ-Lipschitz function with $\phi(0) = 0$,
$\mathcal{R}_n\, \phi\circ\mathcal{H} \le 2C_\phi\, \mathcal{R}_n\mathcal{H}$

Think LP, sup obtained at vertices.

2) $\mathcal{R}_n(\text{convex-hull}(\mathcal{H})) = \mathcal{R}_n(\mathcal{H})$ for finite $\mathcal{H}$

3) Consider any finite $\mathcal{H}$.

Then, $\mathcal{R}_n\mathcal{H} \le \sqrt{\frac{2\log|\mathcal{H}|}{n}} \cdot \sqrt{\sup_{h\in\mathcal{H}} \frac{1}{n}\sum h(Z_i)^2}$. You'll show this in HW2.

Now, we analyze the Rademacher complexity of regularized linear models.

Example $\ell(\theta; x, y) = (1 - y\theta^T x)_+ = \phi(y\theta^T x)$, $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_p \le r\}$

$$\mathcal{R}_n\{(x,y) \mapsto \ell(\theta; x,y) : \theta \in \Theta\} = \mathcal{R}_n\{(x,y) \mapsto \phi(y\theta^T x) - \phi(0) : \theta \in \Theta\}$$
$$\le \mathcal{R}_n\{(x,y) \mapsto y \cdot \theta^T x : \theta \in \Theta\} \quad \text{by contraction principle}$$

Define $z := y \cdot x$. Then, $\nearrow = \mathcal{R}_n\{z \mapsto \theta^T z : \theta \in \Theta\}$.

We now derive scale-sensitive bounds on this quantity.

Theorem $\mathcal{H} := \{z \mapsto \theta^T z : \|\theta\|_2 \le r\}$ If $\mathbb{E}\|z\|_2^2 \le C_2^2$, then $\mathbb{E}\mathcal{R}_n\mathcal{H} \le \frac{C_2}{\sqrt{n}} r$

Pf)
$$\mathbb{E}\mathcal{R}_n\mathcal{H} = \frac{1}{n}\mathbb{E}\sup_{\|\theta\|_2 \le r} \theta^T\left(\sum_i \sigma_i z_i\right) \le \frac{r}{n}\mathbb{E}\left\|\sum \sigma_i z_i\right\|_2 \quad \text{by Cauchy-Schwarz}$$
$$\le \frac{r}{n}\sqrt{\mathbb{E}\left\|\sum_{i=1}^n \sigma_i z_i\right\|_2^2} \quad \text{by Jensen's inequality}$$

Write out $\|\sum \sigma_i z_i\|_2^2$ and note that cross terms have mean zero.

$$= \frac{r}{n}\sqrt{\mathbb{E}\sum_1^n \|\sigma_i z_i\|_2^2} = \frac{r}{n}\sqrt{\mathbb{E}\sum^n \|z_i\|_2^2} \le \frac{r}{\sqrt{n}} \cdot C_2. \qquad \boxtimes$$

What if you are interested in high-dimensional features, but think the model is sparse?

Theorem $\mathcal{H} := \{z \mapsto \theta^T z : \|\theta\|_1 \le s\}$ If $\|z\|_\infty \le C_\infty$ a.s., then $\mathbb{E}\mathcal{R}_n\mathcal{H} \le \frac{C_\infty}{\sqrt{n}} s \cdot \sqrt{2\log(2d)}$.

↳ You'll show this in HW1.

☆ $\boxed{\log d \quad \text{vs.} \quad d}$ when $s \ll d$ then L_1-regularization is nice.

These theorems say "so long as you regularize properly, your model complexity doesn't grow with problem dimension d"

Of course, all of these results compare performance against best-in-model-class. They don't say anything on whether that model class is good.

Chaining & Dudley's entropy integral

We now give more sophisticated bounds on the Rademacher complexity.
The bounds we develop play a key role in empirical process theory
e.g. uniform CLT
$$\sqrt{n}\left(\frac{1}{n}\sum h(z_i) - \mathbb{E}Z\right) \Rightarrow G(h) \quad \text{where } G \text{ is a Gaussian process indexed by } h\in\mathcal{H}$$

Covering

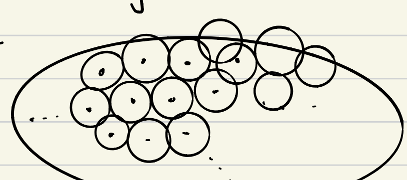
We begin with notions of packing & covering numbers.

Consider a metric space $(\mathcal{T}, d)$ — any nonempty set; metric on $\mathcal{T}$

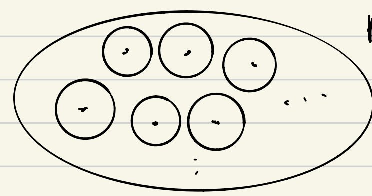
Def For any $\varepsilon>0$, $\{h_i\}_{i=1}^N$ is a ε-cover of $\mathcal{T}$ if $\forall h\in\mathcal{T}\ \exists 1\le i\le n$ s.t. $d(h,h_i)\le\varepsilon$.

Def The ε-covering number of $\mathcal{T}$ is the size of the smallest ε-cover of $\mathcal{T}$
$$N(\mathcal{T},d,\varepsilon) := \inf\{N\ge 0 : \exists\ \varepsilon\text{-cover } \{h_i\}_{i=1}^N \text{ of } \mathcal{T}\}.$$
We call $\log N(\mathcal{T},d,\varepsilon)$ the metric entropy.

Cover



Packing



Def For any $\delta>0$, $\{h_i\}_{i=1}^M\subseteq\mathcal{T}$ is a δ-packing of $\mathcal{T}$ if $d(h_i,h_j)>\delta\ \ \forall i\ne j$.

The δ-packing number of $\mathcal{T}$ is the size of the largest δ-packing
$$M(\mathcal{T},d,\delta) := \sup\{M\ge 0 : \exists\ \delta\text{-packing } \{h_i\}_{i=1}^M \text{ of } \mathcal{T}\}$$

Lemma
$$M(\mathcal{T},d,2\delta) \overset{①}{\le} N(\mathcal{T},d,\delta) \overset{②}{\le} M(\mathcal{T},d,\delta)$$

Pf)

skip

② Let $\{h_i\}_{i=1}^M$ be the maximal δ-packing. Then, $\forall h\in\mathcal{T}$, $d(h,h_i)\le\delta\ \ \forall i=1,\dots,M$.
So this is a δ-cover of $\mathcal{T}$.

① Suppose there exists 2δ-packing $\{h_1,\dots,h_M\}$ and δ-cover $\{\tilde h_1,\dots,\tilde h_N\}$, with $M\ge N+1$.
Then, $\exists\, 1\le i<j\le M$, and $1\le k\le N$ s.t. $d(h_i,\tilde h_k)\le\delta$, $d(h_j,\tilde h_k)\le\delta$. So $d(h_i,h_j)\le 2\delta$ $\times$ ☒.

Lemma Consider $\|\cdot\|, \|\cdot\|'$ on $\mathbb{R}^d$. Let $\mathbb{B}, \mathbb{B}'$ be corresponding unit balls. Then,
$$\left(\frac{1}{\delta}\right)^d\frac{\mathrm{Vol}(\mathbb{B})}{\mathrm{Vol}(\mathbb{B}')} \overset{①}{\le} N(\mathbb{B},\|\cdot\|',\delta) \overset{②}{\le} \frac{\mathrm{Vol}(\frac{2}{\delta}\mathbb{B}+\mathbb{B}')}{\mathrm{Vol}(\mathbb{B}')}$$

Pf) ①: Let $\{h_j\}_{j=1}^N$ be a δ-cover (in $\|\cdot\|'$) of $\mathbb{B}$, so $\mathbb{B}\subseteq\bigcup_{j=1}^N\{h_j+\delta\mathbb{B}'\}$.
This implies $\mathrm{Vol}(\mathbb{B})\le N\,\mathrm{Vol}(\delta\mathbb{B}') = N\delta^d\,\mathrm{Vol}(\mathbb{B}')$.

② : Let $\{h_j\}_{j=1}^M$ be a maximal $\frac{\delta}{2}$-packing of $\mathbb{B}$ (in $\|\cdot\|'$). By def of packing, $\{h_j+\frac{\delta}{2}\mathbb{B}'\}_{j=1}^M$ are disjoint and contained in $\mathbb{B}+\frac{\delta}{2}\mathbb{B}'$.
$$\mathrm{Vol}\left(\bigcup_{j=1}^M\{\theta_j+\tfrac{\delta}{2}\mathbb{B}'\}\right) = M\,\mathrm{Vol}(\tfrac{\delta}{2}\mathbb{B}') = M\cdot\left(\tfrac{\delta}{2}\right)^d\mathrm{Vol}(\mathbb{B}') \le \mathrm{Vol}(\mathbb{B}+\tfrac{\delta}{2}\mathbb{B}') = \left(\tfrac{\delta}{2}\right)^d\mathrm{Vol}(\tfrac{2}{\delta}\mathbb{B}+\mathbb{B}')$$
☒

Example Consider $\mathcal{H} = \{\ell(\theta;\cdot) : \theta \in \Theta\}$. Let $\|h\|_{L^2(\hat{P}_n)} := \sqrt{\frac{1}{n}\sum h(Z_i)^2}$.
Assume $|\ell(\theta;z) - \ell(\theta';z)| \le L(z)\,\|\theta-\theta'\|$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$.

Then, any ε-cover of Θ induces a $\|L\|_n \cdot \varepsilon$ - cover on $\mathcal{H}$ in $\|\cdot\|_{L^2(\hat{P}_n)}$

(Let $\{\theta_j\}_{j=1}^{N}$ be a ε-cover. Then, consider $\{\ell(\theta_j;\cdot)\}_{j=1}^{N}$ is a $\|L\|_{L^2(\hat{P}_n)}\varepsilon$-cover of $\mathcal{H}$.
$\forall \theta \in \Theta$, let j be s.t. $\|\theta-\theta_j\| \le \varepsilon$. Take $\|\ell(\theta;\cdot) - \ell(\theta_j;\cdot)\|_{L^2(\hat{P}_n)} \le \|L\|_{L^2(\hat{P}_n)}\|\theta-\theta_j\| \le \|L\|_{L^2(\hat{P}_n)}\varepsilon$

So we conclude $N(\mathcal{H}, \|\cdot\|_{L^2(\hat{P}_n)}, \varepsilon\|L\|_{L^2(\hat{P}_n)}) \le N(\Theta, \|\cdot\|, \varepsilon)$.

subG processes Instead of the (empirical) Rademacher complexity, we consider more general processes.

Def A collection of zero mean RVs $\{V_h : h \in T\}$ is a sub-Gaussian process w.r.t. d if
$$\mathbb{E}e^{\lambda(V_h - V_{h'})} \le \exp\left(\frac{\lambda^2}{2}d(h,h')^2\right) \quad \forall h,h' \in T,\ \forall \lambda \in \mathbb{R}.$$
↳ tail of $V_h - V_{h'}$ is $d(h,h')^2$-subG.

Example (Rademacher process) Consider $R_{n,h} := \frac{1}{\sqrt{n}}\sum \varepsilon_i h(Z_i)$ where ε_i : i.i.d. random signs, $h \in \mathcal{H}$.
Conditional on Z_1^n, $h \mapsto R_{n,h}$ is a subGaussian process w.r.t. $\|\cdot\|_n$ on $\mathcal{H}$.

Pf) $R_{n,h} - R_{n,h'} = \frac{1}{\sqrt{n}}\sum \varepsilon_i(h-h')(Z_i)$. Recalling that ε_i's are 1-sub-Gaussian,
$$\mathbb{E}[\exp(\lambda(R_{n,h} - R_{n,h'}))\,|\,Z_1^n] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{\lambda\varepsilon_i}{\sqrt{n}}(h-h')(Z_i)\right)\Big|\,Z_i\right] \le \prod_{i=1}^n \exp\left(\frac{\lambda^2}{2n}(h-h')^2(Z_i)\right)$$
$$= \exp\left(\frac{\lambda^2}{2}\cdot\frac{1}{n}\sum_{i=1}^n (h(Z_i) - h'(Z_i))^2\right)$$
$$= \exp\left(\frac{\lambda^2}{2}\|h-h'\|^2_{L^2(\hat{P}_n)}\right) \qquad \boxtimes.$$

So to bound $\mathfrak{R}_n\mathcal{H} = \frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{\sqrt{n}}\sum \varepsilon_i h(Z_i)\,|\,Z_1^n\right] = \frac{1}{\sqrt{n}}\mathbb{E}\left[\sup_{h\in\mathcal{H}} R_{n,h}\,|\,Z_1^n\right]$,
we can bound suprema of sub-Gaussian processes.

Key Lemma Let X_j be σ_j^2-sub-G RVs, $j=1,\dots,N$. Then, $\mathbb{E}\max_{1\le j\le N} X_j \le \max_{1\le j\le N}\sigma_j \cdot 2\sqrt{\log N}$, $N \ge 2$.

Proposition Let $\{V_h : h\in T\}$ be a sub-Gaussian process w.r.t. a metric d on T. Let $D := \sup_{h,h'\in T} d(h,h')$ $\operatorname{diam}(T)$
Then, for any $\delta > 0$, $$\mathbb{E}\sup_{h\in T} V_h \le 2\,\mathbb{E}\sup_{\substack{d(h,h')\le\delta \\ h,h'\in T}}(V_h - V_{h'}) + 4D\sqrt{\log N(T,d,\delta)}$$

Pf) Let $N = N(T,d,\delta)$, and $\{h_j\}_{j=1}^N$ be a δ-cover of T. Fix an arbitrary $h \in T$.
There exists j s.t. $d(h,h_j) \le \delta$. Then,
$$V_h - V_{h_1} = V_h - V_{h_j} + V_{h_j} - V_{h_1} \le \sup_{\substack{r,r'\in T \\ d(r,r')\le\delta}}(V_r - V_{r'}) + \max_{1\le j\le N}|V_{h_j} - V_{h_1}|$$
Given another arbitrary $\tilde{h} \in T$, the same bound holds for $V_{h_1} - V_{\tilde{h}}$.
Adding the two, and taking supremum over $h,\tilde{h} \in T$
$$\sup_{h,\tilde{h}\in T} V_h - V_{\tilde{h}} \le 2\sup_{\substack{r,r'\in T \\ d(r,r')\le\delta}}(V_r - V_{r'}) + 2\max_{1\le j\le N}|V_{h_j} - V_{h_1}|$$
From Lemma, $\mathbb{E}\max_{1\le j\le N}|V_{h_j} - V_{h_1}| \le 2D\sqrt{\log N}$. $\boxtimes$.

Example (A parametric class on $[0,1]$) Define $\ell(\theta;z) = 1 - e^{-\theta z}$, $\theta \in [0,1]$, $z \in [0,1]$.

$\mathcal{H} = \{\ell(\theta,\cdot) : \theta \in [0,1]\} \subseteq \{h : [0,1] \to \mathbb{R}\}$.

First term: $\mathbb{E} \sup_{\|h-h'\|_{L_2(\hat{P}_n)} \le \delta} R_{n,h} - R_{n,h'} = \mathbb{E} \sup_{\|h-h'\|_{L_2(\hat{P}_n)} \le \delta} \frac{1}{\sqrt{n}} \sum \sigma_i (h(z_i) - h'(z_i)) \le \sqrt{n} \cdot \delta$ by Cauchy-Schwarz.

Second term: It's easy to check $\theta \mapsto \ell(\theta;z)$ is 1-Lipschitz $\forall z \in [0,1]$. From above example,

$N(\mathcal{H}, \|\cdot\|_{L_2(\hat{P}_n)}, \delta) \le N([0,1], |\cdot|, \delta) \le \frac{1}{\delta} + 1$. $\quad D = \sup_{\theta \in [0,1]} \frac{1}{n} \sum (1 - e^{-\theta z_i})^2 \le 1$

$$\mathcal{R}_n \mathcal{H} = \mathbb{E}\left[\sup_{\theta \in [0,1]} \frac{1}{n} \sum \sigma_i (1 - e^{-\theta z_i}) \,\Big|\, Z_1^n \right] = \frac{1}{\sqrt{n}} \mathbb{E} \sup_{h \in \mathcal{H}} R_{n,h}$$

$$\le \frac{1}{\sqrt{n}} \cdot \left(2\delta\sqrt{n} + 4\sqrt{\log\left(\tfrac{1}{\delta}+1\right)} \right) \quad \text{for any } \delta$$

$$= \frac{2}{\sqrt{n}}\, 2 \inf_{\delta \in (0,\frac{1}{4})} \left(\sqrt{n}\delta + 2\sqrt{\log\left(\tfrac{1}{\delta}+1\right)} \right)$$

Setting $\delta = \frac{1}{4\sqrt{n}}$, we get $\mathcal{R}_n \mathcal{H} \lesssim \sqrt{\frac{c \log n}{n}}$ $\boxtimes$.

We now use a more refined argument that allows a tighter bound on the supremum.

Theorem (Dudley's entropy integral) Let $\{V_h : h \in T\}$ be a sub-Gaussian process w.r.t. d on T. For any $\delta \in [0, D]$,

$$\mathbb{E} \sup_{h \in T} V_h \le \mathbb{E}\left[\sup_{h,\tilde{h} \in T} V_h - V_{\tilde{h}} \right] \le 2\mathbb{E}\left[\sup_{\substack{\gamma,\gamma' \in T \\ d(\gamma,\gamma') \le \delta}} V_\gamma - V_{\gamma'} \right] + 32 \int_\delta^D \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon.$$

Rmk: setting $\delta = 0$, $\mathbb{E} \sup_{h \in T} V_h \le 32 \int_0^D \sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon$, $\quad N(T,d,\delta) = 0 \ \forall \delta \ge D$.

Pf) We start with ineq from before: keep this written

$$\sup_{h,\tilde{h} \in T} V_h - V_{\tilde{h}} \overset{*}{\le} 2 \sup_{\substack{\gamma,\gamma' \in T \\ d(\gamma,\gamma') \le \delta}} (V_\gamma - V_{\gamma'}) + 2 \max_{1 \le j \le N} |V_{h_j} - V_{h_1}|$$

Instead of bounding the last term via Lemma, we use a chaining argument.

Recall that $\mathcal{U} := \{h_j\}_{j=1}^N$ was a δ-cover of T.

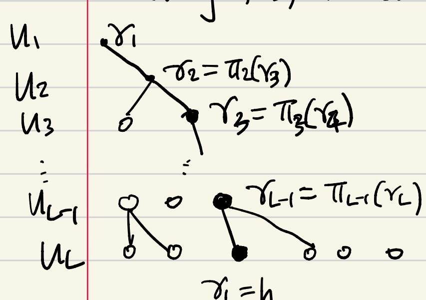
For each m, define $\mathcal{U}_m :=$ minimal $(D \cdot 2^{-m})$-cover of $\mathcal{U}_{m-1}$ (we allow elements of T).

For $L = \lceil \log_2 D/\delta \rceil$, $2^{-L} \le \delta/D$, so set $\mathcal{U}_L = \mathcal{U}$.

By def, $|\mathcal{U}_m| \le N(T, d, D \cdot 2^{-m})$. Best approx of $h \in \mathcal{U}$ in $\mathcal{U}_m$.

For each m, define $\pi_m : \mathcal{U} \to \mathcal{U}_m$, $\quad \pi_m(h) = \operatorname{argmin}_{\tilde{h} \in \mathcal{U}_m} d(h, \tilde{h})$

Using this, we construct a chain from any $h \in \mathcal{U}$. $\quad \gamma_{m-1} = \pi_{m-1}(\gamma_m)$



$$V_h - V_{\gamma_1} = \sum_{m=2}^L V_{\gamma_m} - V_{\gamma_{m-1}} \quad \text{and}$$

$$\mathbb{E}|V_h - V_{\gamma_1}| \le \sum_{m=2}^L \sup_{\gamma \in \mathcal{U}_m} |V_\gamma - V_{\pi_{m-1}(\gamma)}| \quad //$$

Similarly, for any other $\tilde{h} \in T$, we have same bound with $\tilde{\gamma}_m$'s.

We arrive at $|V_h - V_{\tilde h}| = |V_{r_1} - V_{\tilde r_1} + V_h - V_{r_1} + V_{\tilde r_1} - V_{\tilde h}|$

$\leq |V_{r_1} - V_{\tilde r_1}| + \underbrace{|V_h - V_{r_1}| + |V_{\tilde r_1} - V_{\tilde h}|}_{\text{bound via chaining}}$

$\leq \max_{r_1, \tilde r_1 \in U_1} |V_{r_1} - V_{\tilde r_1}| + 2\sum_{m=2}^{L} \max_{r \in U_m} |V_r - V_{\pi_{m-1}(r)}|$

From Lemma, $\mathbb{E} \max_{r_1,\tilde r_1 \in U_1} |V_{r_1} - V_{\tilde r_1}| \leq 2D\sqrt{\log N(T, d, \frac{D}{2})}$ and

since $\max_{r \in U_m} d(r, \pi_{m-1}(r)) \leq D \cdot 2^{-(m-1)}$, and $|U_m| \leq N(T, d, D\cdot 2^{-m})$, we have

$$\mathbb{E} \max_{r \in U_m} |V_r - V_{\pi_{m-1}(r)}| \leq 2D 2^{-(m-1)} \sqrt{\log N(T, d, D\cdot 2^{-m})}.$$

Conclude that

$$\mathbb{E} \sup_{h, \tilde h \in T} |V_h - V_{\tilde h}| \leq 4 \sum_{m=1}^{L} D\cdot 2^{-(m-1)} \sqrt{\log (T, d, D 2^{-m})}.$$

Since $\delta \mapsto \log N(T, d, \delta)$ is dec, $D\cdot 2^{-m} \sqrt{\log N(T, d, D\cdot 2^{-m})} \leq 2\int_{D 2^{-(m+1)}}^{D 2^{-m}} \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon$.

$\Rightarrow 2\,\mathbb{E} \sup_{h,\tilde h \in T} |V_h - V_{\tilde h}| \leq 32 \int_{\delta/4}^{D} \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon$.

Combining with $*$, we get the result. $\boxtimes$.

<u>Example</u> Recall that for $\ell(\theta; z) = 1 - e^{-\theta z}$, $\theta, z \in [0,1]$, $\mathcal{R}_n \mathcal{H} \lesssim \sqrt{\frac{\log n}{n}}$.

Let's use Dudley's entropy integral.

$\mathcal{R}_n \mathcal{H} \leq \frac{32}{\sqrt{n}} \int_0^1 \sqrt{\log(1+\frac{1}{\varepsilon})}\, d\varepsilon \leq \frac{32}{\sqrt{n}} \int_0^1 \sqrt{\log \frac{2}{\varepsilon}}\, d\varepsilon$, $\quad u = \sqrt{\log \frac{2}{\varepsilon}} \Rightarrow \varepsilon = 2e^{-u^2}$

$d\varepsilon = -4u e^{-u^2} du$

$= \frac{32}{\sqrt{n}} \int_0^{\sqrt{\log 2}} 4u^2 e^{-u^2} du$

$= \frac{C}{\sqrt{n}} \cdot \left(-ue^{-u^2}\Big|_{\sqrt{\log 2}}^{\infty} + \int_{\sqrt{\log 2}}^{\infty} e^{-u^2} du\right) = \frac{C}{\sqrt{n}}$, No $\sqrt{\log n}$ factor!

<u>Example</u> Lipschitz functions $|\ell(\theta; z) - \ell(\theta'; z)| \leq L\|\theta - \theta'\|$, $\mathcal{H} = \{\ell(\theta; \cdot) : \theta \in \Theta\}$

Recall: $N(\mathcal{H}, \|\cdot\|_{L^2(\hat P_n)}, \varepsilon\cdot L) \leq N(\Theta, \|\cdot\|, \varepsilon)$. If $\Theta \subseteq r\mathbb{B}$, $N(\Theta, \|\cdot\|, \varepsilon) \leq (1 + \frac{2r}{\varepsilon})^d$

$\mathcal{R}_n \mathcal{H} \leq \frac{32}{\sqrt{n}} \int_0^{r\cdot L} \sqrt{\log N(\mathcal{H}, \|\cdot\|_{L^2(\hat P_n)}, \varepsilon)}\, d\varepsilon \leq \frac{32 L}{\sqrt{n}} \int_0^r \sqrt{\log N(\Theta, \|\cdot\|, \varepsilon)}\, d\varepsilon$

$\leq 32L\sqrt{\frac{d}{n}} \int_0^r \sqrt{\log(1 + \frac{2r}{\varepsilon})}\, d\varepsilon \lesssim L\cdot r \cdot \sqrt{\frac{d}{n}}$.

Combining this with previous concentration result, for $\ell(\theta; z) \in [0, M]$, we have

$$\mathbb{E}\ell(\hat\theta_n; Z) \leq \inf_{\theta \in \Theta} \mathbb{E}\ell(\theta; Z) + CLr\sqrt{\frac{d}{n}} + C\cdot\sqrt{\frac{t}{n}} \quad \text{w.p.} \geq 1 - e^{-t}.$$

Comment on measurability issues. Outer measures.

ULLN what if we just want to show $\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum \ell(\theta; z_i) - \mathbb{E}\ell(\theta; z) \right| \xrightarrow{P} 0$?

<u>Theorem</u> Let H be an envelope function for $\mathcal{H}$: $\forall h\in\mathcal{H}$, $|h|\le H$. Let $\mathbb{E}|H(z)|<\infty$, and define truncated version of $\mathcal{H}$: $\mathcal{H}_M := \{ \underbrace{h\mathbb{1}\{|h|\le M\}}_{=: h_M} : h\in\mathcal{H}\}$.

If $n \cdot \log N(\mathcal{H}_M, \|\cdot\|_{L_1(\hat{P}_n)}, \varepsilon) \xrightarrow{P} 0$ for all fixed $\varepsilon > 0$, $M<\infty$,
then $\sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum h(z_i) - \mathbb{E}h(z)\right| \xrightarrow{P} 0$.

Pf) From symmetrization, $\mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum h(z_i) - \mathbb{E}h(z) \le 2\mathbb{E}\mathcal{R}_n\mathcal{H}$

$$\le 2\mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum \sigma_i\big(h(z_i) - h_M(z_i)\big) + 2\mathbb{E}\mathcal{R}_n\mathcal{H}_M$$

$$\le 2\mathbb{E}\, H(z)\mathbb{1}\{H(z)>M\} + 2\mathbb{E}\mathcal{R}_n\mathcal{H}_M$$

Take a ε-cover $\mathcal{H}_{M,\varepsilon}$ of $\mathcal{H}_M$ in $\|\cdot\|_{L^1(\hat{P}_n)}$. $\mathcal{R}_n\mathcal{H}_M \le \mathcal{R}_n\mathcal{H}_{M,\varepsilon} + \varepsilon$

Now, note that since $\sup_{h\in\mathcal{H}} \|h\|_{L_2(\hat{P}_n)} \le M$, Lemma gives

$$\sqrt{n}\,\mathcal{R}_n\mathcal{H}_{M_\varepsilon} \le 2M\sqrt{\log N(\mathcal{H}_M, \|\cdot\|_{L_1(\hat{P}_n)}, \varepsilon)} \;\Rightarrow\; \mathcal{R}_n\mathcal{H}_{M_\varepsilon} \xrightarrow{P} 0$$

So $\mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum h(z_i) - \mathbb{E}h(z) \le 4\,\mathbb{E}\,h(z)\mathbb{1}\{H(z)>M\} + \underbrace{4\mathbb{E}\mathcal{R}_n\mathcal{H}_{M,\varepsilon}}_{\to 0} + \varepsilon$.

Take $n\to\infty$, then let $\varepsilon\downarrow 0$, $M\uparrow\infty$. MCT gives the result.