# Generalization & Rademacher complexities

**Thm** Let $g$ be a function satisfying (one coordinate doesn't change fn too much)

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

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)

**Proof**

**Recap** We call $\{D_i := M_i - M_{i-1}\}_{i=1}^n$ a martingale difference sequence w.r.t. $Z_1^n$ if $\mathbb{E}[D_i | X_1^{i-1}] = 0$

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

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

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

Define the Doob martingale $M_i = \mathbb{E}[g(Z_1^n) | Z_1^i]$ $\quad (M_0 = \mathbb{E}g(Z_1^n),\ M_n = g(Z_1^n))$

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

Note that

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

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

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

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

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

$$\hat\theta_n = \operatorname{argmin}_{\theta\in\Theta} \frac{1}{n}\sum \ell(\theta; Z_i)$$

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

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

are small w.h.p.

why is this useful?

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

Taking inf over $\theta$, we get

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

So if $\varepsilon_n$ 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.

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

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

**Part I** We show $\varepsilon_n$ 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 $g(Z_1^n) = \varepsilon_n$. 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 $Q\ell(\theta; z) = \mathbb{E}_{z \sim Q}\ell(\theta; z)$.

$$\begin{aligned} &|g(z_1, \cdots, z_i, \cdots, z_n) - g(z_1, \cdots, z_i', \cdots, z_n)| \\ &= \left| \sup_{\theta \in \Theta} \left\{ \frac{1}{n}\sum \ell(\theta; z_i) - \mathbb{E}\ell(\theta; Z) \right\} - \sup_{\theta \in \Theta} \left\{ \frac{1}{n}\sum \ell(\theta; z_i) - \mathbb{E}\ell(\theta; Z) - \frac{1}{n}\ell(\theta; z_i) + \frac{1}{n}\ell(\theta; z_i') \right\} \right| \le \frac{2M}{n} \end{aligned}$$

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

$$\varepsilon_n \le \mathbb{E}\varepsilon_n + M\sqrt{\frac{2t}{n}} \quad \text{w.p.} \ge 1 - e^{-t}$$

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

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

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

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

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

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

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

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

$$\mathcal{R}_n \mathcal{H} := \mathbb{E}\left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(z_i) \,\middle|\, z_1^n \right]$$

↳ Interpretation: How well can $\mathcal{H}$ fit random noise $\sigma_i$'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 $\bar{\varepsilon}_n$ is symmetric.

Collecting bounds in Parts I & II, we arrive at

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

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 HW1.)

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

2) $\mathcal{R}_n(\text{convex-hull}(\mathcal{H})) = \mathcal{R}_n(\mathcal{H})$ for finite $\mathcal{H}$ (Think LP, sup obtained at vertices.)

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 HW1.)

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, $= \mathcal{R}_n\{Z \mapsto \theta^T Z : \theta \in \Theta\}$.

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) $\mathbb{E}\mathcal{R}_n\mathcal{H} = \frac{1}{n}\mathbb{E}\sup_{\|\theta\|_2 \le r} \theta^T\left(\sum_i \sigma_i Z_i\right) \le \frac{r}{n}\mathbb{E}\left\|\sum \sigma_i Z_i\right\|_2$ by Cauchy-Schwarz

$\le \frac{r}{n}\sqrt{\mathbb{E}\left\|\sum_{i=1}^n \sigma_i Z_i\right\|_2^2}$ by Jensen's inequality

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

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

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.

☆ $\log d$ vs. $d$ 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 for 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
$\sqrt{n}\left(\frac{1}{n}\sum h(Z_i) - \mathbb{E}Z\right) \Rightarrow G(h)$ where $G$ is a Gaussian process indexed by $h \in \mathcal{H}$

## Covering

We begin with notions of packing & covering numbers.

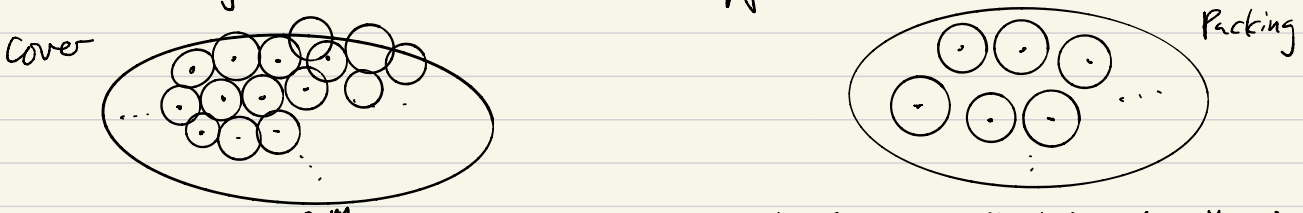
Consider a metric space $(T, d)$ ($T$: any nonempty set; $d$: metric on $T$)

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

**Def** The $\varepsilon$-covering number of $T$ is the size of the smallest $\varepsilon$-cover of $T$

$$N(T, d, \varepsilon) := \inf\{N \ge 0 : \exists \varepsilon\text{-cover } \{h_i\}_{i=1}^N \text{ of } T\}.$$

We call $\log N(T, d, \varepsilon)$ the metric entropy.




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

The $\delta$-packing number of $T$ is the size of the largest $\delta$-packing

$$M(T, d, \delta) := \sup\{M \ge 0 : \exists \delta\text{-packing } \{h_i\}_{i=1}^M \text{ of } T\}$$

**Lemma**

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

Pf)

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

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

**Lemma** Consider $\|\cdot\|, \|\cdot\|'$ on $\mathbb{R}^d$. Let $B, B'$ be corresponding unit balls. Then,

$$\left(\frac{1}{\delta}\right)^d \frac{\mathrm{Vol}(B)}{\mathrm{Vol}(B')} \overset{①}{\le} N(B, \|\cdot\|', \delta) \overset{②}{\le} \frac{\mathrm{Vol}\left(\frac{2}{\delta}B + B'\right)}{\mathrm{Vol}(B')}$$

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

②: Let $\{h_i\}_{i=1}^m$ be a maximal $\frac{\delta}{2}$-packing of $B$ (in $\|\cdot\|'$). By def of packing, $\{h_j + \frac{\delta}{2}B'\}_{j=1}^m$ are disjoint and contained in $B + \frac{\delta}{2}B'$.

$$\mathrm{Vol}\left(\bigcup_{j=1}^m \{h_j + \tfrac{\delta}{2}B'\}\right) = M \,\mathrm{Vol}\left(\tfrac{\delta}{2}B'\right) = M \cdot \left(\tfrac{\delta}{2}\right)^d \mathrm{Vol}(B') \le \mathrm{Vol}\left(B + \tfrac{\delta}{2}B'\right) = \left(\tfrac{\delta}{2}\right)^d \mathrm{Vol}\left(\tfrac{2}{\delta}B + B'\right)$$

☒

**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 $\varepsilon$-cover of $\Theta$ 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 $\varepsilon$-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 $R_{n,h} := \frac{1}{\sqrt{n}}\sum \sigma_i h(Z_i)$ where $\sigma_i$ : iid. 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)
$R_{n,h} - R_{n,h'} = \frac{1}{\sqrt{n}}\sum \sigma_i (h-h')(Z_i)$. Recalling that $\sigma_i$'s are 1-sub-Gaussian,

$$\mathbb{E}\left[\exp\left(\lambda(R_{n,h}-R_{n,h'})\right) \mid Z_1^n\right] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{\lambda\sigma_i}{\sqrt{n}}(h-h')(Z_i)\right) \mid Z_i\right] \le \prod_{i=1}^n \exp\left(\frac{\lambda^2}{2n}(h-h')^2(Z_i)\right)$$
$$= \exp\left(\frac{\lambda^2}{2}\cdot\frac{1}{n}\sum_{i=1}^n (h(Z_i)-h'(Z_i))^2\right) = \exp\left(\frac{\lambda^2}{2}\|h-h'\|^2_{L^2(\hat{P}_n)}\right) \quad \square.$$

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

**Key Lemma** Let $X_j$ be $\sigma_i^2$-sub-G RVs, $j=1,\cdots,N$. 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 $D := \sup_{h,h'\in T} d(h,h')$ (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 $\delta$-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, $\mathbb{E}\max_{1\le j\le N}|V_{h_j} - V_{h_1}| \le 2D\sqrt{\log N}$. $\square$.

**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,

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

$$\mathcal{R}_n \mathcal{H} = \mathbb{E}\left[ \sup_{\theta \in [0,1]} \frac{1}{n} \sum \sigma_i (1 - e^{-\theta z_i}) \,\middle|\, Z_1^n \right] = \frac{1}{\sqrt{n}} \mathbb{E} \sup_{h \in \mathcal{H}} R_n h$$
$$\le \frac{1}{\sqrt{n}} \cdot \left( 2\delta\sqrt{n} + 4\sqrt{\log\left(\frac{1}{\delta} + 1\right)} \right) \quad \text{for any } \delta$$
$$= \frac{2}{\sqrt{n}} \, 2 \inf_{\delta \in (0, \frac{1}{4})} \left( \sqrt{n}\delta + 2\sqrt{\log\left(\frac{1}{\delta} + 1\right)} \right)$$

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

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 $\delta \in [0, D]$,

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

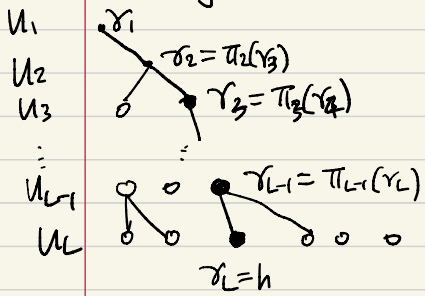
Rmk: Setting $\delta = 0$, $\mathbb{E} \sup_{h \in T} V_h \le 32 \int_0^\infty \sqrt{\log N(T, d, \varepsilon)} \, d\varepsilon$, $N(T, d, \delta) = 0$ $\forall \delta \ge D$.

Pf) We start with ineq pwn before: $\sup_{h, h' \in T} V_h - V_{h'} \overset{*}{\le} 2 \sup_{\substack{r, r' \in T \\ d(r, r') \le \delta}} (V_r - V_{r'}) + 2 \max_{1 \le j \le N} |V_{h_j} - V_{h_1}|$ (keep this written)

Instead of bounding the last term via Lemma, we use a chaining argument.
Recall that $U := \{h_j\}_{j=1}^N$ was a $\delta$-cover of $T$.
For each $m$, define $U_m :=$ minimal $(D \cdot 2^{-m})$-cover of $U_{m-1}$ (we allow elements of $T$).
For $L = \lceil \log_2 D/\delta \rceil$, $2^{-L} \le \delta$, so set $U_L = U$.

Best approx of $h \in U$ in $U_m$.

By def, $|U_m| \le N(T, d, D \cdot 2^{-m})$.
For each $m$, define $\pi_m : U \to U_m$, $\pi_m(h) = \operatorname{argmin}_{\tilde{h} \in U_m} d(h, \tilde{h})$
Using this, we construct a chain pwn any $h \in U$. $\gamma_{m-1} = \pi_{m-1}(\gamma_m)$




$V_h - V_{\gamma_1} = \sum_{m=2}^L V_{\gamma_m} - V_{\gamma_{m-1}}$ and
$\mathbb{E}|V_h - V_{\gamma_1}| \le \sum_{m=2}^L \sup_{r \in U_m} |V_r - V_{\pi_{m-1}(r)}|$ //

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

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

$\le |V_{r_1} - V_{\tilde r_1}| + |V_h - V_{r_1}| + |V_{\tilde r_1} - V_{\tilde h}|$ — bound via chaining

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

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

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

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

Conclude that

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

Since $\delta \mapsto \log N(T, d, \delta)$ is dec, $D \cdot 2^{-m} \sqrt{\log N(T, d, D \cdot 2^{-m})} \le 2 \int_{D2^{-(m+1)}}^{D 2^{-m}} \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon$.

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

Combining with $\star$, we get the result. ☒

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

Let's use Dudley's entropy integral.

$$\mathcal{R}_n \mathcal{H} \le \frac{32}{\sqrt{n}} \int_0^1 \sqrt{\log\left(1 + \frac{1}{\varepsilon}\right)}\, d\varepsilon \le \frac{32}{\sqrt{n}} \int_0^1 \sqrt{\log \frac{2}{\varepsilon}}\, d\varepsilon, \quad u = \sqrt{\log \frac{2}{\varepsilon}} \Rightarrow \varepsilon = 2e^{-u^2}, \; d\varepsilon = -4u e^{-u^2} du$$

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

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

**Example** Lipschitz functions $|\ell(\theta; z) - \ell(\theta'; z)| \le L \|\theta - \theta'\|$, $\mathcal{H} = \{\ell(\theta; \cdot) : \theta \in \Theta\}$

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

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

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

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

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

Comment on measurability issues. Outer measures.

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

**Theorem** 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, $\mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum h(Z_i) - \mathbb{E}h(Z) \le 2\mathbb{E}\mathcal{R}_n\mathcal{H}$

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

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

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

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

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

Same bound holds for $\sup_{h\in\mathcal{H}} \mathbb{E}h(Z) - \frac{1}{n}\sum h(Z_i)$.

So $\mathbb{E}\sup_{h\in\mathcal{H}} \left| \frac{1}{n}\sum h(Z_i) - \mathbb{E}h(Z) \right| \le 4\mathbb{E}\, h(Z)\mathbb{1}\{H(Z)>M\} + 4\mathbb{E}\mathcal{R}_n\mathcal{H}_{M,\varepsilon} + \varepsilon$

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