B9145: Reliable Statistical Learning **Spring 2023**

Lecture 1: Generalization

Lecturer: Hongseok Namkoong *Scribe: Yuanzhe Ma*

1.1 Generalization

Notation:

$$\widehat{P}_n \ell(\theta; Z) := \mathbb{E}_{\widehat{P}_n} \ell(\theta; Z) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i)$$

We want to show that

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i)$$

achieves near-optimal population loss.

Now, we will use bounded difference inequality to show the following uniform concentration result:

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

are small w.h.p.

Why is this useful?

$$\begin{aligned}
\mathbb{E}\ell(\hat{\theta}_n; Z) &\le \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{\theta}_n; Z) + \overline{\Delta}_n \quad \text{by def of } \overline{\Delta}_n \\
&\le \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z) + \overline{\Delta}_n, \quad \forall \theta \in \Theta \quad \text{by def of } \hat{\theta}_n \\
&\le \mathbb{E}\ell(\theta; Z) + \overline{\Delta}_n + \Delta_n \quad \text{by def of } \Delta_n
\end{aligned}$$

Taking infimum over θ, we get

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

so if $\overline{\Delta}_n + \Delta_n$ 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 Δ_n & $\overline{\Delta}_n$. As we'll see, the case for In is symmetric, so we focus on Δ_n below.

1.2 Bounded differences

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

Theorem 1. *Let g be a function satisfying*

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

(one coordinate doesn't change the function too much), then for independent random variable Z_i's,

$$\mathbb{P}\left(g(Z_1^n) - \mathbb{E}g(Z_1^n) \ge t\right) \le \exp\left(-\frac{2t}{\sum_{i=1}^n c_i^2}\right).$$

Assumption A. *We assume $\ell(\theta; Z) \in [0, M]$ in this lecture note.*

1.2.1 Part 1

We can use bounded differences to show that Δ_n is concentrated around its mean w.h.p.

Define $g(z_1, \cdots, z_n) := \sup_{\theta\in\Theta}\left\{\frac{1}{n}\sum_{i=1}^n \ell(\theta; z_i) - \mathbb{E}\ell(\theta; Z_i)\right\}$ so that $g(Z_1^n) = \Delta_n$. We will apply bounded differences.

As a notational shorthand, we use $\widehat{P}_n(\cdot) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Z_i \in \cdot\}$ and $Q\ell(\theta; Z) := \mathbb{E}_{Z\sim Q}\ell(\theta; Z)$.

Then

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

From bounded differences, $\mathbb{P}\left(\Delta_n - \mathbb{E}\Delta_n \ge t\right) \le \exp\left(-\frac{nt^2}{M}\right)$. Equivalently, $\Delta_n \le \mathbb{E}\Delta_n + M\sqrt{\frac{2t}{n}}$ w.p. $\ge 1 - e^{-t}$. So now, it suffices to control $\mathbb{E}\Delta_n$!

We begin with concentration results for light-tailed RVs.

Definition 1. *A RV X is σ^2-subGaussian if $\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \le \exp\left(\frac{\sigma^2}{2}\lambda^2\right), \forall\lambda \in \mathbb{R}$.*

From Markov inequality, for any $\lambda \ge 0$,

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

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

Example 1. ε: random signs (Rademacher) is 1-subGaussian.

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

◇

Example 2. If $X \in [a, b], \mathbb{E}X = 0$, then X is $(b-a)^2$-subGaussian.
By Jensen inequality, we have $\mathbb{E}_X e^{\lambda X} = \mathbb{E}_X e^{\lambda(X - \mathbb{E}_{X'} X')} \leq \mathbb{E} e^{\lambda(X-X')}$ where X' is an independent copy of X. Let ε be random signs independent of everything so that $X - X' \overset{d}{=} \varepsilon(X - X')$ (verify by MGF) so

$$\mathbb{E} e^{\lambda(X-X')} = \mathbb{E}_{X,X'} \mathbb{E}_\varepsilon e^{\varepsilon \lambda(X-X')} \overset{\text{Example 1}}{\leq} \mathbb{E}_{X,X'} e^{\frac{\lambda^2}{2}(X-X')^2} \leq e^{\frac{\lambda^2}{2}(b-a)^2}$$

◇

Actually, we can show a stronger result.

Lemma 1. *If $X \in [a, b], \mathbb{E}X = 0$, then X is $\frac{(b-a)^2}{4}$-subGaussian.*

Proof. By convexity of $x \mapsto e^{\lambda x}, e^{\lambda x} \leq \frac{b-x}{b-a} e^{\lambda a} + \frac{x-a}{b-a} e^{\lambda b}$. Take expectations on both sides. For $h = \lambda(b-a), p = \frac{-a}{b-a}, L(h) = -hp + \log(1 - p + pe^h), \mathbb{E}e^{\lambda X} \leq e^{L(h)}, L(0) = L'(0) = 0, L''(h) \leq \frac{1}{4}, \forall h$, so $L(h) \leq \frac{1}{8} h^2$ by Taylor. □

We are ready to show bounded differences inequality (Theorem 1) now.

Definition 2. *$\{M_i\}_{i=1}^n$ is a martingale sequence 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 $\mathbb{E}[M_i | Z_1, \cdots, Z_{i-1}] = M_{i-1}$. We call $\{D_i = M_i - M_{i-1}\}_{i=1}^n$ a martingale difference sequence w.r.t. Z_1^n ($\mathbb{E}[D_i | Z_1^{i-1}] = 0$)*

Lemma 2. *Let D_i be a martingale difference sequence w.r.t. Z_1^n s.t. $\exists \sigma_i^2$ with $\mathbb{E}[e^{\lambda D_i} | Z_1^{i-1}] \leq \exp\left(\frac{\sigma_i^2 t^2}{2}\right) \forall i$. Then, $M_n - M_0 = \sum_{i=1}^n D_i$ is $(\sum_{i=1}^n \sigma_i^2)$-subGaussian.*

Proof.

$$\mathbb{E} e^{\lambda \sum_{i=1}^n D_i} = \mathbb{E} e^{\lambda D_n} \cdot e^{\lambda \sum_{i=1}^{n-1} D_i} = \mathbb{E}\Big[\mathbb{E}[e^{\lambda D_n} \cdot e^{\lambda \sum_{i=1}^{n-1}} D_i | Z_1^{n-1}]\Big] \leq \exp\left(\frac{\sigma_n^2 t^2}{2}\right) \cdot \mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i}]$$

By induction, we get the result. □

Proof of bounded differences or Theorem 1. Define the Doob martingale $M_i = \mathbb{E}[g(Z_1^n) | Z_1^i]$ ($M_0 = \mathbb{E}g(Z_1^n)$ and $M_n = g(Z_1^n)$). So we we can bound $\mathbb{P}(M_n - M_0 \geq t)$.

Note that

$$|D_i| = |\mathbb{E}[g(Z_1^n)|Z_1^i] - \mathbb{E}[g(Z_1^n)|Z_1^{i-1}]| \leq \sup_{z,z'} |\mathbb{E}_{Z_{i+1}^n}[g(Z_1^{i-1}, z, Z_{i+1}^n)] - \mathbb{E}_{Z_{i+1}^n}[g(Z_1^{i-1}, z', Z_{i+1}^n)] \leq c_i,$$

so $\mathbb{E}[e^{\lambda D_i} | Z_1^{i-1}] = \mathbb{E}[e^{\lambda(D_i - \mathbb{E}[D_i | Z_1^{i-1}])} | Z_1^{i-1}] \leq \exp\left(\frac{\lambda^2}{2} \frac{c_i^2}{4}\right)$. From the previous lemma, and tail inequality for subGaussian RVs, we have the result. □

1.2.2 Part 2

We bound $\mathbb{E}\Delta_n$ via symmetrization.

Let $Z_1', \cdots, Z_n'$ be indepenent copies of $Z_1, \cdots, Z_n$,

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

Let ε_i be i.i.d. random signs (Rademacher RVs), independent of everything else. From $\varepsilon_i(\ell(\theta; Z_i) - \ell(\theta; Z_i')) \overset{d}{=} (\ell(\theta; Z_i) - \ell(\theta; Z_i'))$,

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

Definition 3. *The (empirical) Rademacher complexity of a class $\mathcal{H}$ of functions $h : \mathcal{Z} \to \mathbb{R}$ is*

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

Interpretation: how well can $\mathcal{H}$ fit random noise ε_i's? (where $\varepsilon_i h(Z_i)$ is the margin).

Note that $\mathfrak{R}_n(\mathcal{H}) = \mathfrak{R}_n(-\mathcal{H})$, so the case for $\overline{\Delta}_n$ is symmetric.

Collecting bounds in Part 1 and 2, we arrive at

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

so we conclude

$$\mathbb{E}\ell(\hat{\theta}_n; Z) \le \inf_{\theta\in\Theta}\mathbb{E}\ell(\theta; Z) + 4\mathbb{E}\mathfrak{R}_n(\mathcal{H}) + 2M\sqrt{\frac{t}{2n}} \quad w.p. \ge 1 - 2e^{-t}, \tag{1.1}$$

Basic properties of Rademacher complexity:

1. Contraction principle: Let ϕ be a C_ϕ-Lipschitz function with $\phi(0) = 0$, then $\mathfrak{R}_n(\phi \circ \mathcal{H}) \le C_\phi\mathfrak{R}_n(\mathcal{H})$.
2. $\mathfrak{R}_n(\text{convex-hull}(\mathcal{H})) = \mathfrak{R}_n(\mathcal{H})$ for finite $\mathcal{H}$. (Think LP, sup obtained at vertices)
3. Consider any finite $\mathcal{H}$, then $\mathfrak{R}_n(\mathcal{H}) \le \sqrt{\frac{2\log|\mathcal{H}|}{n}}\sqrt{\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}h(Z_i)^2}$.

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

Example 3. $\ell(\theta; X, Y) = (1 - Y\theta^T X)_+ = \phi(Y\theta^T X), \Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_p \le r\right\}$. Then

$$\begin{aligned}
\mathfrak{R}_n((X, Y) \mapsto \ell(\theta; X, Y) : \theta \in \Theta) &= \mathfrak{R}_n\left((X, Y) \mapsto \phi(Y\theta^T X) - \phi(0) : \theta \in \Theta\right) \\
&\le \mathfrak{R}_n\left((X, Y) \mapsto Y\theta^T X : \theta \in \Theta\right) \quad \text{by contraction principle} \\
&= \mathfrak{R}_n\left(Z \mapsto \theta^T Z : \theta \in \Theta\right) \quad \text{define } \ Z = Y \cdot X
\end{aligned}$$

We now derive scale-sensitive bounds on this quantity. ◇

Theorem 2. *Let $\mathcal{H}_r := \left\{\theta^T Z : \|\theta\|_2 \le r\right\}$. If $\mathbb{E}\|Z\|_2^2 \le C_2^2$, then $\mathbb{E}\mathfrak{R}_n(\mathcal{H}_r) \le \frac{C_2}{\sqrt{n}}r$.*

Proof.

$$\begin{aligned}
\mathbb{E}\mathfrak{R}_n\left(\mathcal{H}_r\right) &= \frac{1}{n}\mathbb{E}\sup_{\|\theta\|_2\le r}\theta^T\left(\sum_{i=1}^n \varepsilon_i Z_i\right) \\
&\le \frac{r}{n}\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i Z_i\right\|_2 \quad \text{by Cauchy-Schwarz inequality} \\
&\le \frac{r}{n}\sqrt{\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i Z_i\right\|_2^2} \quad \text{by Jensen's inequality}
\end{aligned}$$

Write out $\|\sum_{i=1}^n \varepsilon_i Z_i\|_2^2$ and note that across terms have mean zero, we have

$$\frac{r}{n}\sqrt{\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i Z_i\right\|_2^2} = \frac{r}{n}\sqrt{\mathbb{E}\sum_{i=1}^n \|\varepsilon_i Z_i\|_2^2} = \frac{r}{n}\sqrt{\mathbb{E}\sum_{i=1}^n \|Z_i\|_2^2} \le \frac{r}{\sqrt{n}}C_2.$$

□

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

Theorem 3. *Let $\mathcal{H} = \left\{Z \mapsto \theta^T Z : \|\theta\|_1 \le s\right\}$, if $\|Z\|_\infty \le C_\infty$ a.s., then $\mathbb{E}\mathfrak{R}_n(\mathcal{H}) \le \frac{C_\infty}{\sqrt{n}} s\sqrt{2\log 2d}$.*

Proof. See HW 1. □

$\log d$ vs d

Remark 1. *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 of whether that model class is good.*

1.3 Chaining and Dudley's entropy integral

We now give more sophisticated bounds on the Rademacher complexity. These bounds we develop play a key role in empirical process theory, e.g. uniform CLT: $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n h(Z_i) - \mathbb{E}h(Z)\right) \Rightarrow \mathbb{G}(h)$ where $\mathbb{G}$ is a Gaussian process indexed by $h \in \mathcal{H}$.

1.3.1 Covering

We begin with notations of packing & covering numbers.

Consider a metric space $(\mathcal{T}, d)$, $\mathcal{T}$ is any nonempty set and d is a metric on $\mathcal{T}$.

Definition 4. *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. $d(h, h_i) \le \varepsilon$.*

Definition 5. *The ε-cover number of $\mathcal{T}$ is the size of the smallest ε-cover of $\mathcal{T}$*

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

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

Definition 6. *For any $\delta > 0, \{h_i\}_{i=1}^N \subset \mathcal{T}$ is a δ-packing of $\mathcal{T}$ if $d(h_i, h_j) > \delta, \forall i \neq j$.*

Definition 7. *The δ-packing number of $\mathcal{T}$ of the size of the largest δ-packing of $\mathcal{T}$:*

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

Lemma 3. $M(\mathcal{T}, d, 2\delta) \overset{(1)}{\leq} N(\mathcal{T}, d, \delta) \overset{(2)}{\leq} M(\mathcal{T}, d, \delta)$.

Proof. (1): Suppose there exists 2δ-packing $\{h_1, \cdots, h_M\}$ and δ-cover $\{h_1, \cdots, h_N\}$ with $M \geq N+1$. Then, $\exists 1 \leq i < j \leq M$ and $1 \leq k \leq N$ s.t. $d(h_i, h_k) \leq \delta, d(h_j, h_k) \leq \delta$, so $d(h_i, h_j) \leq 2\delta$ which is a contradiction.

(2): Let $\{h_i\}_{i=1}^M$ be the maximal δ-packing. Then for any $h \in \mathcal{T}, \exists i = 1, \cdots, M$ s.t. $d(h, h_i) \leq \delta$, (if this is not true, then we can create a packing of size $M+1$), so this is a δ-cover of $\mathcal{T}$. □

Lemma 4. *Consider two norms $\|\cdot\|, \|\cdot\|'$ on $\mathbb{R}^d$. Let $\mathbb{B}, \mathbb{B}'$ be the corresponding unit balls. Then*

$$\left(\frac{1}{\delta}\right)^d \frac{\mathrm{vol}(\mathbb{B})}{\mathrm{vol}(\mathbb{B}')} \leq N(\mathbb{B}, \|\cdot\|', \delta) \leq \frac{\mathrm{vol}\left(\frac{2}{\delta}\mathbb{B} + \mathbb{B}'\right)}{\mathrm{vol}(\mathbb{B}')},$$

where $+$ is the Minkovski sum.

Proof. (1): Let $\{h_j\}_{j=1}^N$ be a δ-cover (in $\|\cdot\|'$) of $\mathbb{B}$, so $\mathbb{B} \subset \cup_{j=1}^N \{h_j + \delta\mathbb{B}'\}$. This implies $\mathrm{vol}(\mathbb{B}) \leq N\mathrm{vol}(\delta\mathbb{B}') = N\delta^d\mathrm{vol}(\mathbb{B}')$.
(2): Let $\{h_i\}_{i=1}^M$ be a maximal $\frac{\delta}{2}$-packing of $\mathbb{B}$ (in $\|\cdot\|'$). By definition of packing, $\left\{h_j + \frac{\delta}{2}\mathbb{B}'\right\}_{j=1}^M$ are disjoint and contained in $\mathbb{B} + \frac{\delta}{2}\mathbb{B}'$. And $\mathrm{vol}\left(\cup_{j=1}^M \left\{h_j + \frac{\delta}{2}\mathbb{B}'\right\}\right) = M\mathrm{vol}(\frac{\delta}{2}\mathbb{B}') = M\left(\frac{\delta}{2}\right)^d \mathrm{vol}(\mathbb{B}') \leq \mathrm{vol}(\mathbb{B} + \frac{\delta}{2}\mathbb{B}') = \left(\frac{\delta}{2}\right)^d \mathrm{vol}\left(\frac{2}{\delta}\mathbb{B} + \mathbb{B}'\right)$. □

Example 4. Consider $\mathcal{H} = \{\ell(\theta; \cdot) : \theta \in \Theta\}$. Let $\|h\|_{L^2(\hat{P}_n)} := \sqrt{\frac{1}{n}\sum_{i=1}^n h(Z_i)^2}$. Assume $|\ell(\theta; Z) - \ell(\theta'; Z)| \leq L(Z)\|\theta - \theta'\|$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$. Then, any ε-cover of Θ induces a $\|L\|_{L^2(\hat{P}_n)} \cdot \varepsilon$-cover on $\mathcal{H}$ in $\|\cdot\|_{L^2(\hat{P}_n)}$: (Let $\{\theta_j\}_{j=1}^N$ be a ε-cover. Then, consider $\{\ell(\theta_j; \cdot)\}_{j=1}^N$, a $\|L\|_{L^2(\hat{P}_n)}\, \varepsilon$-cover of $\mathcal{H}$. $\forall \theta \in \Theta$, let j be s.t. $\|\theta - \theta_j\| \leq \varepsilon$, then $\|\ell(\theta; Z) - \ell(\theta_j; Z)\|_{L^2(\hat{P}_n)} \leq \|L\|_{L^2(\hat{P}_n)}\|\theta - \theta_j\| \leq \|L\|_{L^2(\hat{P}_n)}\, \varepsilon$.) So we conclude

$$N\Big(\mathcal{H}, \|\cdot\|_{L^2(\hat{P}_n)}, \varepsilon\, \|L\|_{L^2(\hat{P}_n)}\, L\Big) \leq N(\Theta, \|\cdot\|, \varepsilon).$$

◇

1.3.2 SubGaussian process

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

Definition 8. *A collection of zero mean RVs $\{V_h : h \in \mathcal{T}\}$ is a sub-Gaussian process w.r.t. d if*

$$\mathbb{E}e^{\lambda(V_h - V_{h'})} \leq \exp\left(\frac{\lambda^2}{2} d(h,, h')^2\right) \quad \forall h, h' \in \mathcal{T}, \forall \lambda \in \mathbb{R}.$$

Example 5. (Rademacher process) Consider $R_{n,h} := \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i h(Z_i)$ 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. $\|\cdot\|_{L^2(\hat{P}_n)}$ on $\mathcal{H}$. ◇

Proof. Note that $R_{n,h} - R_{n,h'} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (h - h')Z_i$. Recalling that ε_i's are 1-subGaussian by Example 1,

$$\begin{aligned}
\mathbb{E}[\exp(\lambda(R_{n,h} - R_{n,h'}))|Z_1^n] &= \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{\lambda\varepsilon_i}{\sqrt{n}}(h-h')Z_i\right)\Big|Z_i\right] \\
&\leq \prod_{i=1}^n \exp\left(\frac{\lambda^2}{2n}(h-h')^2(Z_i)\right) \\
&= \exp\left(\frac{\lambda^2}{2}\frac{1}{n}\sum_{i=1}^n (h(Z_i) - h'(Z_i)^2)\right) \\
&= \exp\left(\frac{\lambda^2}{2}\|h-h'\|_{L^2(\hat{P}_n)}\right).
\end{aligned}$$

□

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

Lemma 5. *Let X_j be ε_i^2-subGaussian RVs, $j = 1, \cdots, N$, then $\mathbb{E}\max_{1\leq j\leq N} X_j \leq \max_{1\leq j\leq N}\sigma_j \cdot \sqrt{2\log N}$ for $N \geq 2$.*

Proposition 4. *Let $\{V_h : h \in \mathcal{T}\}$ be a subGaussian process w.r.t. a metric d on $\mathcal{T}$. Let $D := \sup_{h,h'\in\mathcal{T}} d(h,h')$. Then for any $\delta > 0$,*

$$\mathbb{E}\sup_{h\in\mathcal{T}} V_h \leq 2\mathbb{E}\sup_{d(h,h')\leq\delta, h,h'\in\mathcal{T}}(V_h - V_{h'}) + 4D\sqrt{\log N(\mathcal{T}, d, \delta)} \tag{1.2}$$

Proof. Let $N = N(\mathcal{T}, d, \delta)$ and $\{h_j\}_{j=1}^N$ be a δ-cover of $\mathcal{T}$. Fix an arbitrary $h \in \mathcal{T}$. There exists j s.t. $d(h, h_j) \leq \delta$. Then,

$$V_h - V_{h_1} = V_h - V_{h_j} + V_{h_j} - V_{h_1} \leq \sup_{d(h,h')\leq\delta, h,h'\in\mathcal{T}}(V_h - V_{h'}) + \max_{1\leq j\leq N}|V_{h_j} - V_{h_1}|$$

Given another arbitrary $\tilde{h} \in \mathcal{T}$, the same bound holds for $V_{h_1} - V_{\tilde{h}}$. Adding the two, and taking supremum over $h, \tilde{h} \in \mathcal{T}$,

$$\sup_{h,\tilde{h}\in\mathcal{T}} V_h - V_{\tilde{h}} \leq 2\sup_{d(h,h')\leq\delta, h,h'\in\mathcal{T}}(V_h - V_{h'}) + 2\max_{1\leq j\leq N}|V_{h_j} - V_{h_1}|$$

From Lemma 5, $\mathbb{E}\max_{1\leq j\leq N}|V_{h_j} - V_{h_1}| \leq 2D\sqrt{\log N}$. □

Example 6. (A parameter on $[0,1]$). Define $\ell(\theta; Z) = 1 - e^{-\theta Z}, \theta \in [0,1], Z \in [0,1]$. $\mathcal{H} = \{\ell(\theta;\cdot) : \theta \in [0,1]\} \subset \{h : [0,1] \to \mathbb{R}\}$. The first term of RHS of the bound (1.2) is

$$\mathbb{E}\sup_{\|h-h'\|_{L^2(\hat{P}_n)}\leq\delta} R_{n,h} - R_{n,h'} = \mathbb{E}\sup_{\|h-h'\|_{L^2(\hat{P}_n)}\leq\delta}\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i(h(Z_i) - h'(Z_i)) \leq \sqrt{n}\cdot\delta \quad \text{by Cauchy-Schwarz}$$

To deal with the second term of RHS of the bound (1.2), it's easy to check that $\theta \mapsto \ell(\theta; z)$ is 1-Lipschitz for $\forall z \in [0,1]$. From Example 5,

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

and

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{H}) &= \mathbb{E}\left[\sup_{\theta\in[0,1]} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (1-e^{-\theta Z_i}) | Z_1^n\right] \\
&= \frac{1}{\sqrt{n}} \mathbb{E} \sup_{h\in\mathcal{H}} R_{n,h} \\
&\le \frac{1}{\sqrt{n}}\left(2\delta\sqrt{n} + 4\sqrt{\log\left(\frac{1}{\delta}+1\right)}\right) \qquad \text{for any } \delta > 0 \\
&= \frac{2}{\sqrt{n}} \inf_{\delta\in(0,\frac{1}{4})} \left(\delta\sqrt{n} + 2\sqrt{\log\left(\frac{1}{\delta}+1\right)}\right)
\end{aligned}$$

Setting $\delta = \frac{1}{4\sqrt{n}}$, we get $\mathfrak{R}_n(\mathcal{H}) \lessapprox \sqrt{\frac{\log n}{n}}$. ⋄

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

Theorem 5 (Dudley's entroy integral)**.** *Let $\{V_h : h \in \mathcal{T}\}$ be a sub-Gaussian process w.r.t. d on $\mathcal{T}$. For any $\delta > 0$*

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

Remark 2. *Setting $\delta = 0$ gives $\mathbb{E}\sup_{h\in\mathcal{T}} V_h \le 32\int_0^\infty \sqrt{\log N(\mathcal{T}, d, \varepsilon)} d\varepsilon$. ($N(\mathcal{T}, d, \delta) = 1$ for any $\delta \ge D$)*

Example 7. Recall that $\ell(\theta; Z) = 1 - e^{-\theta Z}, \theta, Z \in [0,1], \mathfrak{R}_n(\mathcal{H}) \le \sqrt{\frac{\log n}{n}}$ from Example 6. Let's use Dudley's entropy integral.

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{H}) &\le \frac{32}{\sqrt{n}}\int_0^1 \sqrt{\log(1+\frac{1}{\varepsilon})} d\varepsilon \le \frac{32}{\sqrt{n}}\int_0^1 \sqrt{\log\frac{2}{\varepsilon}} d\varepsilon, \quad u = \sqrt{\log\frac{2}{\varepsilon}} \\
&= \frac{32}{\sqrt{n}}\int_0^{\sqrt{\log 2}} 4u^2 e^{-u^2} du \\
&= \frac{C}{\sqrt{n}}\left(-ue^{-u^2}\Big|_{\sqrt{\log 2}}^\infty + \int_{\sqrt{\log 2}}^\infty e^{-u^2} du\right) = \frac{C}{\sqrt{n}}
\end{aligned}$$

Compare to Example 6, there's no $\sqrt{\log n}$ factor! ⋄

Example 8. Consider Lipschitz functions $|\ell(\theta; Z) - \ell(\theta'; Z)| \le L(Z)\|\theta - \theta'\|$ and $\mathcal{H} = \{\ell(\theta; \cdot) : \theta \in \Theta\}$. Recall: $N(\mathcal{H}, \|\cdot\|_{L^2(\hat{P}_n)}, \varepsilon \cdot L) \le N(\Theta, \|\cdot\|, \varepsilon)$. If $\Theta \subset r\mathbb{B}, N(\Theta, \|\cdot\|, \varepsilon) \le \left(1 + \frac{2r}{\varepsilon}\right)^d$ so

$$\begin{aligned}
\mathfrak{R}_n(\mathcal{H}) &\le \frac{32}{\sqrt{n}}\int_0^{r\cdot L} \sqrt{\log N(\mathcal{H}, \|\cdot\|_{L^2(\hat{P}_n)}, \varepsilon)} d\varepsilon \\
&\le \frac{32L}{\sqrt{n}}\int_0^r \sqrt{\log N(\Theta, \|\cdot\|, \varepsilon)} d\varepsilon \\
&\le 32L\sqrt{\frac{d}{n}}\int_0^r \sqrt{\log\left(1+\frac{2r}{\varepsilon}\right)} d\varepsilon \lessapprox L\cdot r\cdot\sqrt{\frac{d}{n}}
\end{aligned}$$

⋄

Combining this with previous concentration result (1.1), for $\ell(\theta; Z) \in [0, M]$, we have

$$\mathbb{E}\ell(\hat{\theta}_n; Z) \leq \inf_{\theta \in \Theta} \mathbb{E}\ell(\theta; Z) + CLr\sqrt{\frac{d}{n}} + C\sqrt{\frac{t}{n}} \quad w.p. \geq 1 - 2e^{-t}.$$