17.1 Symmetrization Inequality

Let $\mathcal{F} = \{f : S \to \mathbb{R}\}$. $X_1, \ldots, X_n$ are iid, $X_j \in S$. The empirical process is

$$\mathbb{Z}_n(f) = \sqrt{n}(P_n f - Pf),$$

where $P_n f = \frac{1}{n} \sum f(X_j)$. If $\mathcal{F} = \{I\{(-\infty, t]\} : t \in \mathbb{R}\}$, then

$$P_n I\{(-\infty, t]\} = F_n(t) = \frac{1}{n} \sum I\{X_j \le t\}.$$

17.1.1 Rademacher process

Let $\varepsilon_1, \ldots, \varepsilon_n$ be iid Rademacher random variables, i.e. fair coin flips $\varepsilon_j \in \{\pm 1\}$. Assume that ε_j and X_j are jointly independent. The Rademacher process is

$$R_n(f) = \frac{1}{n} \sum \varepsilon_j f(X_j) = \frac{1}{n} \langle \vec{\varepsilon}, f(\vec{X}) \rangle.$$

It is sub-Gaussian conditioned on the X_j.

Theorem 1. Symmetrization inequality, E. Giné and J. Zinn

$$\begin{aligned}
\mathbb{E}||P_n - P||_{\mathcal{F}} &\le 2\mathbb{E}||R_n||_{\mathcal{F}}, \qquad \text{where} \\
||P_n - P||_{\mathcal{F}} &= \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum f(X_j) - \mathbb{E}f(X_j)| \\
||R_n||_{\mathcal{F}} &= \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum \varepsilon_j f(X_j)| = |\frac{1}{\sqrt{n}} ||\mathbb{Z}_n(f)||.
\end{aligned}$$

Proof: Let $Y_1, \ldots, Y_n$ be iid copies of $X_1, \ldots, X_n$ which are jointly independent of X_j. The X_j live in California, the Y_j are born on Jupiter. Note that

$$Pf = \mathbb{E}f(X_1) = \mathbb{E}f(Y_1) = \frac{1}{n} \sum \mathbb{E}f(Y_j).$$

Then

$$\mathbb{E} \sup_f |P_n f - Pf| = \mathbb{E} \sup_f |P_n f - \tilde{\mathbb{E}} \tilde{P}_n f|,$$

where the $\sim$ denotes expectation with respect to Y_j. By Jensen's inequality,

$$\mathbb{E}\sup_f |P_n f - \tilde{\mathbb{E}}\tilde{P}_n f| \leq \mathbb{E}\tilde{\mathbb{E}}\sup_f |P_n f - \tilde{P}_n f|.$$

To apply Jensen's, define $Q(f) = P_n f - \tilde{P}_n f$ and $G(Q) = \sup_{f\in\mathcal{F}} |Q(f)|$; G is convex. Then

$$\mathbb{E}||P_n - P||_{\mathcal{F}} \leq \mathbb{E}_{X,Y}||P_n f - \tilde{P}_n f||_{\mathcal{F}} = \mathbb{E}\sup_f |\frac{1}{n}\sum (f(X_j) - f(Y_j))|.$$

For any (deterministic) sequence $\sigma_j \in \{\pm 1\}$, $j = 1, \ldots, n$,

$$\mathbb{E}\sup_f |\frac{1}{n}\sum (f(X_j) - f(Y_j)| = \mathbb{E}\sup_f |\frac{1}{n}\sum \sigma_j (f(X_j) - f(Y_j))|.$$

There are 2^n possible such sequences so the average is

$$\begin{aligned}2^{-n}\sum_{\sigma_j}\mathbb{E}\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum \sigma_j(f(X_j) - f(Y_j))| &= \mathbb{E}_\varepsilon \mathbb{E}_{X,Y}\sup_f |\frac{1}{n}\sum \varepsilon_j (f(X_j) - f(Y_j))| \\ &\leq \mathbb{E}_{\varepsilon,X}\sup_f |\frac{1}{n}\sum \varepsilon_j f(X_j) + \mathbb{E}_{\varepsilon,Y}\sup_f |\frac{1}{n}\sum \varepsilon_j f(Y_j)| = 2\mathbb{E}||R_n||_{\mathcal{F}}\end{aligned}$$

Theorem 2. Desymmetrization inequality.

$$\mathbb{E}||P_n - P||_{\mathcal{F}} \geq \frac{1}{2}\mathbb{E}||R_n||_{\mathcal{F}_c} = \frac{1}{2}\mathbb{E}\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum \varepsilon_j (f(X_j) - \mathbb{E}f(X_j))|$$

where $\mathcal{F}_c = \{f - Pf,\ f \in \mathcal{F}\}$.

We also have that $\mathbb{E}||R_n||_{\mathcal{F}_c} \geq \mathbb{E}||R_n||_{\mathcal{F}} - k$. WHAT'S k INDEED

$$|\frac{1}{n}\sum \varepsilon_j (f(X_j) - Pf)| \geq \frac{1}{n}\sum \varepsilon f(X_j)| - |\frac{1}{n}Pf\sum \varepsilon|$$

$$\mathbb{E}\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum \varepsilon_j Pf| \leq \sup_{f\in\mathcal{F}}|Pf|\mathbb{E}|\frac{1}{n}\sum \varepsilon_j| \leq \sup_{f\in\mathcal{F}}|Pf|\mathbb{E}^{1/2}(\frac{1}{n}\sum \varepsilon_j)^2 = \sup_f |P_f|/\sqrt{n}.$$

so we can take $k = \frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}}|Pf|$.

We wish to understand this empirical bound for $\mathcal{F}$ the class of indicator functions. Let $\mathcal{C}$ be a collection of subsets of S. Then

$$||P_n - P||_{\mathcal{C}} := ||P_n - P||_{\mathcal{F}_{\mathcal{C}}},$$

where $\mathcal{F}_{\mathcal{C}}$ is the set of indicator functions on $C \in \mathcal{C}$.

Recall that if x_t is a process indexed by a finite set T, and $x_t \sim SG(\sigma_t^2)$, then

$$\mathbb{E}\sup_t |X_t| \leq \sqrt{2}\max_t \sigma_t \sqrt{\log(2N)}.$$

Consider $X_t = \sum t_j \varepsilon_j$, where $t = (t_1, \ldots, t_n) \in \mathbb{R}^n$ and ε_j are Rademacher r.v.s. Then $X_t \in SG(||t||_2^2)$. Let $T \subset \mathbb{R}^n$ be finite with cardinality N. Then

$$R_n(T) = \sup_{t\in T}|\sum \varepsilon_j t_j|,$$

$$\mathbb{E}R_n(T) \leq \sqrt{2}\max_t ||t||_2 \sqrt{\log(2\,\mathrm{card}(T)}.$$