

Lecture 6: Small risk bound, empirical processes, lower bracketing (Sept 11)

Lecturer: Csaba Szepesvári

Scribes: Shuai Liu

Note: \LaTeX template courtesy of UC Berkeley EECS dept. ([link](#) to directory)

Disclaimer: These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.

This lecture's notes illustrate some uses of various \LaTeX macros. Take a look at this and imitate.

6.1 Recap

Recall the setting of ERM introduced in the previous lectures. We have a dataset (or datalist) $D_n = \{(X_i, f_*(X_i))\}_{i=1}^n$ where $X_i \sim P \in \mathcal{M}_1(\mathcal{X})$ are independent and $f_* \in C_d \subset \mathbb{R}^{2^d}$. Let $|C_d| = N < \infty$. For a fixed function $f \in \mathbb{R}^{2^d}$, let $L_n(f) = \sum_{i=1}^n \mathbb{I}(f(X_i) \neq f_*(X_i))$ and $L(f) = \mathbb{E}[\mathbb{I}(f(X) \neq f_*(X))]$ for $X \sim P$. The empirical risk minimizer is $f_n = \arg \min_{f \in C_d} L_n(f)$. We used the multiplicative Chernoff bound to obtain the following proposition:

Proposition 6.1. For $\delta \in (0, 1)$, $f \in \mathbb{R}^{2^d}$ and $n, N \in \mathbb{N}$, let $\beta_\delta^n(f, N) = \sqrt{\frac{2L(f) \log(\frac{N}{\delta})}{n}}$. For all $f_0 \in C_d$ and $\delta \in (0, 1)$, let $U(\delta, f_0, C_d)$ be the event that:

$$U(\delta, f_0, C_d) := \left\{ \forall f \in C_d : L(f) \leq L_n(f) + \beta_\delta^n(f, N+1) \right\} \cap \left\{ L_n(f_0) \leq L(f_0) + \beta_\delta^n(f_0, N+1) + \frac{\log(\frac{N+1}{\delta})}{3n} \right\}.$$

It follows that $\mathbb{P}(U(\delta, f_0, C_d)) \geq 1 - \delta$.

For all $f_0 \in C_d$, on the event $U(\delta, f_0, C_d)$, we have that:

$$\begin{aligned} L(f_n) &\leq L_n(f_n) + \beta_\delta^n(f_n, N+1) \\ &\leq L_n(f_0) + \beta_\delta^n(f_n, N+1) && (f_n \text{ is the sol. to ERM}) \\ &\leq L(f_0) + \beta_\delta^n(f_0, N+1) + \beta_\delta^n(f_n, N+1) + \frac{\log(\frac{N+1}{\delta})}{3n}, \end{aligned}$$

which gives us the following theorem:

Theorem 6.2. For all $f_0 \in C_d$, w.p. $1 - \delta$,

$$L(f_n) \leq L(f_0) + \beta_\delta^n(f_0, N+1) + \beta_\delta^n(f_n, N+1) + \frac{\log(\frac{N+1}{\delta})}{3n}.$$

Since the above theorem holds for all $f_0 \in C_d$, we can take the infimum:

Corollary 6.3. w.p. $1 - \delta$,

$$L(f_n) \leq \beta_\delta^n(f_n, N+1) + \frac{\log(\frac{N+1}{\delta})}{3n} + \inf_{f \in C_d(\delta)} (L(f) + \beta_\delta^n(f, N+1))$$

Remark 6.4. In our current setting, $\inf_{f \in C_d(\delta)} (L(f) + \beta_\delta^n(f, N+1)) = 0$ because $L(f_*) + \beta_\delta^n(f_*, N+1) = 0$. Corollary 6.3 cannot buy us anything more than the bound we got in the last class because there is still a factor of $\sqrt{1/n}$ in $\beta_\delta^n(f_n, N+1)$. However, in more general settings where $L(f_*) \neq 0$, i.e., noises are injected to $f_*(X_i)$, we may get some benefit from Corollary 6.3.

6.2 Empirical Process

Now consider an arbitrary function class $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ which is potentially infinite and an arbitrary (measurable) loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ (instead of the 0-1 loss we considered in the previous section). Let $f_n = \arg \max_{f \in \mathcal{F}} L_n(f)$ be the empirical risk minimizer on \mathcal{F} . If we were to apply the technique in Proposition 6.1, the term $L_n(f) - L(f)$ for some $f \in \mathcal{F}$, would be the quantity that we would like to bound. To do that, one of the options is to bound:

$$\sup_{f \in \mathcal{F}} |L_n(f) - L(f)| = \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \ell(f(X_i), Y_i) - \int \ell(f(x), y) P(dx, dy) \right| \quad (6.1)$$

To reduce clutter, we define $D_i : \mathcal{F} \rightarrow \mathbb{R}$ for $i \in \mathbb{N}$ such that

$$D_i(f) = \ell(f(X_i), Y_i) - \int \ell(f(x), y) P(dx, dy),$$

and $\bar{D}_n : \mathcal{F} \rightarrow \mathbb{R}$ such that

$$\bar{D}_n(f) = \frac{1}{n} \sum_{i=1}^n D_i(f), \quad \forall f \in \mathcal{F}.$$

Note that $D_1(f), D_2(f), \dots$ are i.i.d. random variables. Then Eq. (6.1) can be written as:

$$\sup_{f \in \mathcal{F}} \bar{D}_n(f).$$

We call $\{\bar{D}_n(f)\}_{n=1}^{\infty}$ an empirical process. Empirical process theory is a subarea of probability theory that studies the question of convergence of the process to 0 in different ways, e.g., convergence in probability or almost sure convergence. If $\bar{D}_n(f) \rightarrow 0$ in probability, it is called the *Weak Law of Large Number* and when $\sup_{f \in \mathcal{F}} \bar{D}_n(f) \rightarrow 0$ happens, we say that *uniform convergence* happens.

6.3 Lower Bracketing Number

Now we further reduce the clutter by introducing new notations. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

$$G = \{(x, y) \rightarrow \ell(f(x), y) : f \in \mathcal{F}\} \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}} = \mathbb{R}^{\mathcal{Z}}.$$

Let $Z_1, Z_2, \dots, Z_n \sim P \in \mathcal{M}_1(\mathcal{Z})$ and let $P_n(dz) = \frac{1}{n} \sum_{i=1}^n \delta_{Z_i}(dz)$ be the *empirical distribution* where $\delta_{Z_i}(\{z\}) = 1$ if $z = Z_i$ and 0 otherwise. Note that δ_{Z_i} is a random measure. For $P \in \mathcal{M}_1(\mathcal{Z})$, let $Pg := \int g dP$ for $g \in \mathcal{G}$. Then Eq. (6.1) can be written as:

$$\sup_{g \in \mathcal{G}} |P_n g - P g|$$

Definition 6.5. Let $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ and fix $P \in \mathcal{M}_1(\mathcal{Z})$. For a fixed $\varepsilon, g_1, \dots, g_m \in \mathbb{R}^{\mathcal{Z}}$ is called a lower bracketing cover of $\mathcal{G} @ P @ \varepsilon$ if for all $g \in \mathcal{G}$, there exists $j \in [m]$ such that:

1. $g_j \leq g$,
2. $Pg \leq P g_j + \varepsilon$.

Note that g_1, \dots, g_m is not necessarily in \mathcal{G} .

Theorem 6.6. Let $\mathcal{G} \subset [0, 1]^{\mathcal{Z}}$, $P \in \mathcal{M}_1(\mathcal{Z})$ and $Z_1, \dots, Z_n \sim P$ for $n \in \mathbb{N}$. For all $\varepsilon > 0, \delta \in (0, 1)$ and $g \in \mathcal{G}$, it follows that w.p. $1 - \delta$,

$$Pg - P_n g \leq \inf_{\varepsilon > 0} \left[\varepsilon + \sqrt{\frac{\log(N_\varepsilon / \delta)}{2n}} \right],$$

where for all $\varepsilon > 0$,

$$N_\varepsilon = \min\{n \in \mathbb{N} : \text{there exists } g_1, \dots, g_n \text{ such that } (g_1, \dots, g_n) \text{ is a lower bracketing cover of } \mathcal{G} @ P @ \varepsilon\}$$

Proof. Fix an $\varepsilon > 0$. Let $m = N_\varepsilon$ and g_1, \dots, g_m be a lower bracketing cover of $\mathcal{G}@P@_\varepsilon$. Using additive Chernoff bound, we have that w.p. at least $1 - \delta$, it follows that

$$Pg_j \leq P_n g_j + \sqrt{\frac{\log(N_\varepsilon/\delta)}{2n}}. \quad (6.2)$$

Pick $g \in \mathcal{G}$ and by definition of lower bracketing cover, there exists $j \in [m]$ such that

$$Pg \leq Pg_j + \varepsilon \leq P_n g_j + \varepsilon + \sqrt{\frac{\log(N_\varepsilon/\delta)}{2n}} \quad (\text{Definition 6.5(1) and Eq. (6.2)})$$

$$\leq P_n g + \varepsilon + \sqrt{\frac{\log(N_\varepsilon/\delta)}{2n}}. \quad (\text{Definition 6.5(2)})$$

Since ε was arbitrary, we then take the infimum over ε :

$$Pg \leq P_n g + \inf_{\varepsilon > 0} \left[\varepsilon + \sqrt{\frac{\log(N_\varepsilon/\delta)}{2n}} \right].$$

□

Corollary 6.7. Let $\hat{g}_n = \arg \min_{g \in \mathcal{G}} P_n g$ be the empirical risk minimizer, then it follows that w.p. at least $1 - \delta$:

$$P\hat{g}_n \leq \inf_{g \in \mathcal{G}} Pg + 2 \inf_{\varepsilon} \left[\varepsilon + \sqrt{\frac{\log((N_\varepsilon + 1)/\delta)}{2n}} \right]$$

Proof. Fix an $\varepsilon > 0$, by definition of infimum, there exists a g_ε such that

$$Pg_\varepsilon \leq \inf_{g \in \mathcal{G}} Pg + \varepsilon \quad (6.3)$$

Denote the lower bracketing cover of $\mathcal{G}@P@_\varepsilon =: C_{LB}(G, P, \varepsilon)$. Let $U(\delta, g_\varepsilon, C_{LB}(G, P, \varepsilon))$ be:

$$U(\delta, g_\varepsilon, C_{LB}(G, P, \varepsilon)) := \left\{ \forall g \in C_{LB}(G, P, \varepsilon) : Pg \leq P_n g + \sqrt{\frac{\log((N_\varepsilon + 1)/\delta)}{2n}} \right\} \cup \left\{ P_n g_\varepsilon \leq Pg_\varepsilon + \sqrt{\frac{\log((N_\varepsilon + 1)/\delta)}{2n}} \right\}.$$

Then $U(\delta, g_\varepsilon, C_{LB}(G, P, \varepsilon))$ holds w.p. $1 - \delta$. On $U(\delta, g_\varepsilon, C_{LB}(G, P, \varepsilon))$, we have that there exists a $j \in [m]$ such that:

$$P\hat{g}_n \leq Pg_j + \varepsilon \quad (\text{Defn. of lower bracketing cover})$$

$$\leq P_n g_j + \varepsilon + \sqrt{\frac{\log((N_\varepsilon + 1)/\delta)}{2n}} \quad (\text{Chernoff's bound})$$

$$\leq P_n \hat{g}_n + \varepsilon + \sqrt{\frac{\log((N_\varepsilon + 1)/\delta)}{2n}} \quad (\text{Defn. of lower bracketing cover})$$

$$\leq P_n g_\varepsilon + \varepsilon + \sqrt{\frac{\log((N_\varepsilon + 1)/\delta)}{2n}} \quad (\text{Defn. of } \hat{g}_n)$$

$$\leq Pg_\varepsilon + \varepsilon + 2\sqrt{\frac{\log((N_\varepsilon + 1)/\delta)}{2n}} \quad (\text{Chernoff's bound})$$

$$\leq \inf_{g \in \mathcal{G}} Pg_\varepsilon + 2\varepsilon + 2\sqrt{\frac{\log((N_\varepsilon + 1)/\delta)}{2n}} \quad (\text{Eq. (6.3)})$$

□

Since ε was arbitrary, we then take the infimum over ε :

$$P\hat{g}_n \leq \inf_{g \in \mathcal{G}} Pg + 2 \inf_{\varepsilon} \left[\varepsilon + \sqrt{\frac{\log((N_{\varepsilon} + 1)/\delta)}{2n}} \right]$$

Similarly, using the multiplicative Chernoff bound, we can get the following corollary:

Corollary 6.8. *Let $\hat{g}_n = \arg \min_{g \in \mathcal{G}} P_n g$ be the empirical risk minimizer, then it follows that w.p. at least $1 - \delta$:*

$$P\hat{g}_n \leq \inf_{g \in \mathcal{G}, \varepsilon > 0} \left[Pg + 2\varepsilon + \sqrt{\frac{2Pg \log((N_{\varepsilon} + 1)/\delta)}{2n}} + \sqrt{\frac{P\hat{g}_n \log((N_{\varepsilon} + 1)/\delta)}{2n}} + \frac{\log((N_{\varepsilon} + 1)/\delta)}{3n} \right]$$