Symmetrization and GC classes

Lab 7, BIO251

03/31/2014

1 Symmetrization

In this section we consider a "symmetrized" process, which is easier to analyze compared to the empirical process, and we show that the two cannot be too far from each other.

Let $\varepsilon_1, \ldots, \varepsilon_n$ be iid Rademacher random variables. The empirical process is:

$$f \mapsto (\mathbb{P}_n - P)f = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - PF)$$

Instead consider the process:

$$f \mapsto \mathbb{P}_n^0 f = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)$$

where here we are considering that the $\varepsilon_1, \ldots, \varepsilon_n$ are independent of $X_1, \ldots, X_n$. Note that both processes have mean 0. The symmetrized process, conditional on the data is a Rademacher process, and we can use the corollary we derived last time. Therefore in what follows, we focus on bounding $||\mathbb{P}_n - P||_{\mathcal{F}}$ in terms of the symmetrized process. Since the supremum $||\mathbb{P}_n - P||_{\mathcal{F}}$ is involved, we need to use outer expectations, because as we discussed the supremum might not be measurable.

Due to technical issues partly related to the the outer expectations, for this problem we would understand independent as variables defined on a product probability space. In other words we will assume that $X_1, X_2, \ldots, X_n$ are the coordinate projections on the product space $(\mathcal{X}^n, \mathcal{A}^n, P^n)$. Thus outer expectations of functions $h(X_1, \ldots, X_n)$ are calculated wrt to the measure P^n. If we have more random variables the space will be considered to be $(\mathcal{X}^n, \mathcal{A}^n, P^n) \times (\mathcal{Z}, \mathcal{C}, Q)$ and the other variables are going to be only in the $(n+1)$ coordinate of the distribution.

We have the following lemma, which is mostly used with $\Phi(x) = x$.

Lemma (Symmetrization). For every convex, nondecreasing $\Phi : \mathbb{R} \mapsto \mathbb{R}$ and a class of measurable functions $\mathcal{F}$,

$$\mathrm{E}^* \, \Phi(||\mathbb{P}_n - P||_{\mathcal{F}}) \leq \mathrm{E}^* \, \Phi(2||\mathbb{P}_n^0||_{\mathcal{F}})$$

The outer expectations are computed as indicating in the preceding paragraph.

Proof. Let $Y_1, \ldots, Y_n$ be independent copies of $X_1, \ldots, X_n$, defined as the last n coordinate projections on the probability space $(\mathcal{X}^n, \mathcal{A}^n, P^n) \times (\mathcal{Z}, \mathcal{C}, Q) \times (\mathcal{X}^n, \mathcal{A}^n, P^n)$. A technical remark here is that the outer expectations wouldn't be affected by adding the variables in such a way because the

coordinate projections are *perfect* maps. For fixed values of $X_1, \ldots, X_n$ we have:

$$||\mathbb{P}_n - P||_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} [f(X_i) - \mathrm{E}\, f(Y_i)] \right| \leq \mathrm{E}_Y^* \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} [f(X_i) - f(Y_i)] \right|$$

Here by E_Y^* we mean the expectation wrt to the Y distribution for given fixed values of $X_1, \ldots, X_n$. Now we apply Jensen's inequality:

$$\Phi(||\mathbb{P}_n - P||_{\mathcal{F}}) \leq \mathrm{E}_Y\, \Phi \left(\left\| \frac{1}{n} \sum_{i=1}^{n} [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}^{*Y} \right)$$

Here by $*Y$ we understand the least measurable majorant of the supremum with respect to $Y_1, \ldots, Y_n$ with holding the $X_1, \ldots, X_n$ fixed. Because Φ is increasing and continuous we can take the $*Y$ from inside of Φ and move it to E_Y^* (see proof in the appendix). And we have:

$$\Phi(||\mathbb{P}_n - P||_{\mathcal{F}}) \leq \mathrm{E}_Y^*\, \Phi \left(\left\| \frac{1}{n} \sum_{i=1}^{n} [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}} \right)$$

Now take an expectation wrt $X_1, \ldots, X_n$ to get:

$$\mathrm{E}^*\, \Phi(||\mathbb{P}_n - P||_{\mathcal{F}}) \leq \mathrm{E}_X^*\, \mathrm{E}_Y^*\, \Phi \left(\left\| \frac{1}{n} \sum_{i=1}^{n} [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}} \right)$$

It follows from Fubini's theorem for outer expectations, which we considered in Lab 1, that the double expectation above is bounded above by the outer expectation E^*. Note here that changing the sign of $[f(X_i) - f(Y_i)]$ is the same as changing the roles of $X_i \leftrightarrow Y_i$. Since these variables are defined on a product probability space the outer expectation of any function $f(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ would be the same under permutation of the functions arguments.

Thus the expression:

$$\mathrm{E}^*\, \Phi \left(\left\| \frac{1}{n} \sum_{i=1}^{n} e_i [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}} \right)$$

Is the same for any n-tuple $(e_1, \ldots, e_n) \in \{-1, 1\}^n$. Therefore:

$$\mathrm{E}^*\, \Phi(||\mathbb{P}_n - P||_{\mathcal{F}}) \leq \mathrm{E}_\varepsilon\, \mathrm{E}_{X,Y}^*\, \Phi \left(\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}} \right)$$

Using triangle's inequality we can split X and Y and then apply Jensen's inequality:

$$\begin{aligned}
&\mathrm{E}_\varepsilon\, \mathrm{E}_{X,Y}^*\, \Phi \left(\frac{2 \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i [f(X_i)] \right\|_{\mathcal{F}} + 2 \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i [f(Y_i)] \right\|_{\mathcal{F}}}{2} \right) \\
&\leq \frac{1}{2}\, \mathrm{E}_\varepsilon\, \mathrm{E}_{X,Y}^*\, \Phi \left(2 \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i [f(X_i)] \right\|_{\mathcal{F}} \right) + \frac{1}{2}\, \mathrm{E}_\varepsilon\, \mathrm{E}_{X,Y}^*\, \Phi \left(2 \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i [f(Y_i)] \right\|_{\mathcal{F}} \right) \\
&= \frac{1}{2}\, \mathrm{E}_\varepsilon\, \mathrm{E}_X^*\, \Phi \left(2 \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i [f(X_i)] \right\|_{\mathcal{F}} \right) + \frac{1}{2}\, \mathrm{E}_\varepsilon\, \mathrm{E}_Y^*\, \Phi \left(2 \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i [f(Y_i)] \right\|_{\mathcal{F}} \right)
\end{aligned}$$

by perfectness of coordinate projections (which simply means that the outer expectations $\mathrm{E}^*_{X,Y} = \mathrm{E}^*_X$ and $\mathrm{E}^*_{X,Y} = \mathrm{E}^*_Y$, which is not generally true for other maps). This finally concludes:

$$\mathrm{E}^* \, \Phi(||\mathbb{P}_n - P||_{\mathcal{F}}) \leq \underbrace{\mathrm{E}_\varepsilon \, \mathrm{E}^*_X}_{\leq \mathrm{E}^*} \Phi \left(2 \frac{1}{n} \left|\left| \sum_{i=1}^n \varepsilon_i [f(X_i)] \right|\right|_{\mathcal{F}} \right)$$

We conclude with a definition. This definition is required because at some point we will deal with an expectation of the sort $\mathrm{E}^*_X \, \mathrm{E}_\varepsilon$ and we would like to claim that this is equal to the joint outer expectation. However, since Fubini's theorem is not true for outer expectations we would need some sort of measurability. Since in this case the ε's are discrete the measurability is true if and only if the maps:

$$(X_1, X_2, \ldots, X_n) \mapsto \left|\left| \sum_{i=1}^n e_i f(X_i) \right|\right|_{\mathcal{F}}$$

are measurable for all $\{e_1, \ldots, e_n\} \in \{-1, 1\}^n$. For the use of Fubini's theorem, it turns out that something a little weaker suffices, i.e. that the maps above are measurable in the completion of $(\mathcal{X}^n, \mathcal{A}^n, P^n)$.

Definition (Measurable class). A class $\mathcal{F}$ of measurable functions $f : \mathcal{F} \mapsto \mathbb{R}$ on a probability space $(\mathcal{X}, \mathcal{A}, P)$ is called a P-measurable class if the map defined above is measurable for all $\{e_1, \ldots, e_n\} \in \mathbb{R}^n$ on the completion of the space $(\mathcal{X}^n, \mathcal{A}^n, P^n)$ (see appendix for definition of complete).

2 Glivenko-Cantelli Theorems

Finally we are here.

Definition (Bracketing numbers). Given two functions l and u, the *bracket* $[l, u]$ is the set of all functions f with $l \leq f \leq u$ pointwise. An *ε-bracket* is a bracket $[l, u]$ with $||u - l|| < \varepsilon$. The *bracketing number* $N_{[]}(\varepsilon, \mathcal{F}, ||\cdot||)$ ($||\cdot||$ here is the norm on $\mathcal{F}$) is the minimum number of ε-brackets needed to cover $\mathcal{F}$. Note that u and l need not belong to $\mathcal{F}$, but are assumed to have finite norms.

Theorem. Let $\mathcal{F}$ be a class of measurable functions such that $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\varepsilon > 0$. Then $\mathcal{F}$ is Glivenko-Cantelli.

Proof. Fix $\varepsilon > 0$. Choose finitely many ε-brackets $[l_i, u_i]$ to cover the space $\mathcal{F}$ and $P(u_i - l_i) < \varepsilon$ for every i. Thus for every $f \in \mathcal{F}$ we have a bracket such that:

$$(\mathbb{P}_n - P)f \leq (\mathbb{P}_n - P)u_i + P(u_i - f) \leq (\mathbb{P}_n - P)u_i + \varepsilon$$

The last inequality gives us:

$$\sup_{f \in \mathcal{F}} (\mathbb{P}_n - P)f \leq \max_i (\mathbb{P}_n - P)u_i + \varepsilon$$

By SLLN the RHS converges a.s. to ε. We have a similar argument for $\inf_{f \in \mathcal{F}} (\mathbb{P}_n - P)f$, and get that $\limsup ||\mathbb{P}_n - P||^*_{\mathcal{F}} \leq \varepsilon$ a.s. for every $\varepsilon \geq 0$. Take a sequence $\varepsilon = \frac{1}{m}$ to conclude that the $\limsup$ is precisely 0.

Example. Take $\mathcal{F}$ to be the set of all indicator functions $\mathbb{1}_{(-\infty,c]}$. This class posses a finite bracketing number, for any underlying distribution and for any $\varepsilon > 0$ and thus is GC. To see this consider the brackets $[\mathbb{1}_{(-\infty,t_i]}, \mathbb{1}_{(-\infty,t_{i+1})}]$ for $i = 1, \ldots, m$ where $-\infty = t_0 < t_1 < \ldots < t_m = \infty$ are selected such that $P(t_i < x < t_{i+1}) < \varepsilon$.

Next we show a more involved theorem. It's sufficiency condition can be verified for many classes of functions by combinatorial arguments using the so called VC dimension, which we will talk about later on.

Theorem. Let $\mathcal{F}$ be a P-measurable class of measurable functions with envelope F such that $P^*F < \infty$. Let $\mathcal{F}_M$ be the class of functions $f\mathbb{1}_{F \leq M}$ when f ranges over $\mathcal{F}$. If $\log N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) = o_P^*(n)$ for every ε and $M > 0$, then $||\mathbb{P}_n - P||_{\mathcal{F}}^* \to 0$ both almost surely and in mean. In particular $\mathcal{F}$ is a GC class.

Proof. By symmetrization and measurability of the class $\mathcal{F}$, and Fubini's theorem we have:

$$\begin{aligned}
\mathrm{E}^* ||\mathbb{P}_n - P||_{\mathcal{F}} &\leq 2\,\mathrm{E}_X\,\mathrm{E}_\varepsilon \left|\left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i [f(X_i)] \right|\right|_{\mathcal{F}} \\
&\leq 2\,\mathrm{E}_X\,\mathrm{E}_\varepsilon \left|\left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i [f(X_i)] \right|\right|_{\mathcal{F}_M} + 2P^*F\{F > M\}
\end{aligned}$$

Where we used the triangle inequality for the last inequality. We can make the right term arbitrary small by picking a large enough M. To show convergence in mean, it suffices to show that the first term goes to 0 for a fixed M. Fix $X_1, X_2, \ldots, X_n$. Define $\mathcal{G}$ to be an ε-net in $L_1(\mathbb{P}_n)$ over $\mathcal{F}_M$. We then have:

$$\mathrm{E}_\varepsilon \left|\left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i [f(X_i)] \right|\right|_{\mathcal{F}_M} \leq \mathrm{E}_\varepsilon \left|\left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i [f(X_i)] \right|\right|_{\mathcal{G}} + \varepsilon$$

Details on this inequality can be found in the appendix. Note here that the size of $\mathcal{G}$ can be selected to be $N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n))$. Now as we know from last time the Rademacher process is sub-Gaussian, and then we can bound the Orlizc norm $\psi_2(x) = e^{x^2} - 1$. Before that we make use of the maximal inequality we derived last time to get that the expression above is further bounded by a multiple of :

$$\sqrt{1 + \log N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n))} \sup_{f \in \mathcal{G}} \left|\left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i [f(X_i)] \right|\right|_{\psi_2 | X} + \varepsilon$$

Here the Orlizc norm is taken wrt to $\varepsilon_1, \ldots, \varepsilon_n$ holding $X_1, \ldots, X_n$ fixed. Apply Hoeffding's inequality to get a bound $\sqrt{\frac{6}{n}}(\mathbb{P}_n f^2)^{1/2} \leq \sqrt{\frac{6}{n}} M$ (more detail in appendix). Putting everything together we get:

$$\sqrt{1 + \log N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n))} \sqrt{\frac{6}{n}} M + \varepsilon \overset{P^*}{\to} \varepsilon$$

We have shown that holding $X_1, \ldots, X_n$ fixed the RHS converges to 0. Taking expectation wrt to X (and noting that everything is bounded by M) we can use the DCT to show it converges to 0.

Thus $||\mathbb{P}_n - P||_{\mathcal{F}}^* \to 0$ in mean. The a.s. convergence follows because $||\mathbb{P}_n - P||_{\mathcal{F}}^*$ is a reverse submartingale wrt to a suitable filtration. We don't show this.

A Some Details

We used the fact that: If $g : \mathbb{R} \mapsto \mathbb{R}$ is nondecreasing and continuous $T : \Omega \mapsto \mathbb{R}$ is an arbitrary map, then $g(T)^* = g(T^*)$. To see this note that because g is nondecreasing $g(T^*) \geq g(T)$. Suppose that there is a measurable functions U such that $g(T^*) \geq U \geq g(T)$. Define $g^{-1}(u) = \sup\{x : g(x) \leq u\}$. Then $g(x) \leq u$ is equivalent to $x \leq g^{-1}(u)$. Now $T^* \geq g^{-1} \circ U \geq T$, but by continuity of g, it follows that $g^{-1} \circ U$ is measurable and therefore has to coincide with T^*.

Complete probability space means that for all subsets $A \subset B, B \in \mathcal{A}$ with $P(B) = 0$ we have $A \in \mathcal{A}$. Turns out that the class of sets $A \cup N$ with $A \in \mathcal{A}$ and $N \subset M$ for some $M \in \mathcal{A}$ with $P(M) = 0$ forms a σ-field – $\bar{\mathcal{A}}$ which obviously contains the original σ-field $\mathcal{A}$. Now, we can extend the measure P to $\bar{P}$ on $\bar{\mathcal{A}}$by $\bar{P}(A \cup N) = \bar{P}(A)$. The space $(\mathcal{X}, \bar{\mathcal{A}}, \bar{P})$ is called a completion of $(\mathcal{X}, \mathcal{A}, P)$.

Here we give a little more detail for the two inequalities. Note that:

$$2\,\mathrm{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i [f(X_i)] \right\|_{\mathcal{F}_M} \leq 2\,\mathrm{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i [f(X_i)] \right\|_{\mathcal{G}} + \varepsilon$$

The ε comes in because for each $f \in \mathcal{F}_M$ we can find a $g \in \mathcal{G}$ with $\mathbb{P}_n |f - g| \leq \varepsilon$. The difference:

$$\mathrm{E}_\varepsilon \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i [f(X_i) - g(X_i)] \right| \leq \mathrm{E}_\varepsilon \frac{1}{n} \sum_{i=1}^n \underbrace{|\varepsilon_i|}_{1} |f(X_i) - g(X_i)| = \mathbb{P}_n |f - g| \leq \varepsilon$$

For the second inequality, note that by Hoeffding's we have that the ψ_2 norm is bounded by $\sqrt{6}\sqrt{\frac{1}{n^2} \sum_{i=1}^n f^2(X_i)} = \sqrt{\frac{6}{n}} (\mathbb{P}_n f^2)^{1/2}$.