GC and Donsker classes

Lab 8, BIO251

04/10/2014

1 Glivenko-Cantelli Theorems

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 Orlicz 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:

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

In the above we used the obvious inequality $\log(1 + x) \leq 1 + \log(x)$, and the fact that $||\cdot||_1 \leq ||\cdot||_2 \leq ||\cdot||_{\psi_2}$. 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:

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

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.

Remark As a corollary it can be shown that a sufficient condition so that the random entropy condition holds is:

$$\sup_Q \log N(\varepsilon ||F||_{Q,1}, \mathcal{F}_M, L_1(Q)) < \infty$$

Where the supremum is taken over all finitely discrete measures (so that it coveres all $\mathbb{P}_n$ for any data realization).

2 A Donsker Theorem

We now provide a sufficient condition for a class being Donsker. This condition will be defined in terms of a "uniform entropy" of the covering numbers, but there are other sufficient conditions using the entropy of the bracketing numbers which we won't consider. We have the following result:

Theorem. Let $\mathcal{F}$ be a class of measurable functions, with envelope F, that satisfies the following uniform entropy bound:

$$\int_0^\infty \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon < \infty$$

The supremum is taken over all finitely discrete probability measures Q on $(\mathcal{X}, \mathcal{A})$, such that $||F||_{Q,2}^2 = \int F^2 dQ > 0$. Let the classes $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, ||f - g||_{P,2} < \delta\}$ and $\mathcal{F}_\infty^2 = \{f^2 : f \in \mathcal{F}_\infty\}$ be P-measurable for any $\delta > 0$. If $P^* F^2 < \infty$, then F is P-Donsker.

Proof. Let $\delta_n \downarrow 0$ be arbitrary. By Markov's inequality we have:

$$P^*(||\mathbb{G}_n||_{\mathcal{F}_{\delta_n}} \geq x) \leq \frac{\mathrm{E}^* ||\mathbb{G}_n||_{\mathcal{F}_{\delta_n}}}{x}$$

We then use the symmetrization lemma we proved last time:

$$\frac{\mathrm{E}^* ||\mathbb{G}_n||_{\mathcal{F}_{\delta_n}}}{x} \leq \frac{2}{x} \mathrm{E}^* \left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right|\right|_{\mathcal{F}_{\delta_n}}$$

Now since we are assuming the class $\mathcal{F}_{\delta_n}$ is P-measurable the E^* can be replaced by iterative expectation $\mathrm{E}^* = \mathrm{E}_X \mathrm{E}_\varepsilon$. We now fix the values of $X_1, \ldots, X_n$. We next apply Hoeffding's inequality to the Rademacher process $f \mapsto \{n^{-1/2} \sum_{i=1}^n \varepsilon_i f(X_i)\}$ to conclude that this process is sub-Gaussian for the $L_2(\mathbb{P}_n)$-seminorm:

$$||f||_n = \sqrt{\frac{1}{n} \sum_{i=1}^n f^2(X_i)}$$

Now we use the second part of the corollary we derived in the end of Lab 6 (with $f_0 = 0$), to conclude that:

$$\mathrm{E}_\varepsilon \left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right|\right|_{\mathcal{F}_{\delta_n}} \lesssim \int_0^\infty \sqrt{\log D(\varepsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n))} d\varepsilon \lesssim \int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n))} d\varepsilon$$

Where the last inequality follows upon noting that we can change the variable $\varepsilon \to 1/2\varepsilon$, and increase the inequality constant a bit.

Note here that when ε is large enough the space $\mathcal{F}_{\delta_n}$ can be contained in only 1 ball. This certainly happens when $\varepsilon > \theta_n$, where:

$$\theta_n^2 = \sup_{f \in \mathcal{F}_{\delta_n}} ||f||_n^2 = \left|\left| \frac{1}{n} \sum_{i=1}^n f^2(X_i) \right|\right|_{\mathcal{F}_{\delta_n}}$$

This is true because then we can center a ball at 0 and radius ε to cover the whole class $\mathcal{F}_{\delta_n}$. Thus we have:

$$\mathrm{E}_\varepsilon \left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right|\right|_{\mathcal{F}_{\delta_n}} \lesssim \int_0^{\theta_n} \sqrt{\log N(\varepsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n))} d\varepsilon$$

Furthermore obviously the covering numbers of the class $\mathcal{F}_\delta \subset \mathcal{F}_\infty$ are bounded by the covering numbers of $\mathcal{F}_\infty$. The latter numbers satisfy $N(\varepsilon, \mathcal{F}_\infty, L_2(Q)) \leq N^2(\varepsilon/2, \mathcal{F}, L_2(Q))$ for all measures Q (see why in the appendix). Therefore upon a change of variables, and bounding the integrand we consequently get:

$$\begin{aligned}
\mathrm{E}_\varepsilon \left|\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right|\right|_{\mathcal{F}_{\delta_n}} &\lesssim \int_0^{\theta_n/||F||_n} \sqrt{\log N(\varepsilon ||F||_n, \mathcal{F}, L_2(\mathbb{P}_n))} d\varepsilon ||F||_n \\
&\lesssim \int_0^{\theta_n/||F||_n} \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon ||F||_n \\
&\leq \int_0^\infty \mathbb{1}(\varepsilon \leq \theta_n^*/||F_*||_n) ||F^*||_n \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon
\end{aligned}$$

Where the supremum is taken over all finitely discrete measures Q (this set trivially incudes $\mathbb{P}_n$ so the bound is trivial in that sense). Note that by doing this trivial bounding we got rid of the dependence of the integrand on the dataset. The only thing that depend on the dataset still is $\mathbb{1}(\varepsilon \leq \theta_n^*/||F_*||_n)||F^*||_n$. Everything so far was conditional on the data $X_1, \ldots, X_n$.

Note that we can always add the constant 1 to the envelope function F without changing the second moment condition. We still need to get the expectation E_X.This will ensure that $F_* \geq 1$. Taking it results in:

$$\begin{aligned}
&\int_0^\infty \mathrm{E}_X \, \mathbb{1}(\varepsilon \leq \theta_n^*/||F_*||_n) ||F^*||_n \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon \\
&\overset{\text{CS}}{\leq} \underbrace{\sqrt{\mathrm{E}_X \, ||F^*||_n^2}}_{O(1)} \int_0^\infty \sqrt{P(\varepsilon \leq \theta_n^*)} \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon
\end{aligned}$$

Now since the integrand $\sqrt{P(\varepsilon \le \theta_n^*)}\sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} \le \sup_Q \sqrt{\log N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))}$ which is integrable by our assumption, the Dominated Convergence Theorem would ensure that the above integral converges to 0 provided that $\theta_n \overset{P_*}{\to} 0$, which will finish the proof of the asymptotic equicontinuity because then $P(||\mathbb{G}_n||_{\mathcal{F}_{\delta_n}} \ge x)$ will converge to 0, as $\delta_n \to 0$.

Note that $\theta_n = \left|\left|\mathbb{P}_n f^2\right|\right|_{\mathcal{F}_{\delta_n}}$. Note then that since $\sup\{Pf^2 : f \in \mathcal{F}_{\delta_n}\} \to 0$, and $\mathcal{F}_{\delta_n} \subset \mathcal{F}_\infty$, it is enough to show that:

$$\left|\left|\mathbb{P}_n f^2 - Pf^2\right|\right|_{\mathcal{F}_\infty} \to 0$$

This is of course a ULLN for the class $\mathcal{F}^2_\infty$. This class has an integrable envelope $(2F)^2$, and is P-measurable by assumption. For any pair of functions $f, g \in \mathcal{F}_\infty$ we have:

$$\mathbb{P}_n|f^2 - g^2| \le \mathbb{P}_n|f-g|4F \le ||f-g||_n \, ||4F||_n$$

Therefore if $||f-g||_n \le \varepsilon ||F||_n$ it follows that $\mathbb{P}_n|f^2 - g^2| \le \varepsilon(2||F||_n)^2$. This statement translated to covering numbers is $N(\varepsilon||2F||_n^2, \mathcal{F}^2_\infty, L_1(\mathbb{P}_n)) \le N(\varepsilon||F||_n, \mathcal{F}_\infty, L_2(\mathbb{P}_n))$. As we argued earlier we have:

$$N(\varepsilon||F||_n, \mathcal{F}_\infty, L_2(\mathbb{P}_n)) \le N^2(\varepsilon||F||_n/2, \mathcal{F}, L_2(\mathbb{P}_n))$$

and the latter must be a finite number in order for us to have the uniform entropy bounded. It is shown in the appendix that the condition $N(\varepsilon||2F||_n^2, \mathcal{F}^2_\infty, L_1(\mathbb{P}_n))$ is a finite number implies the GC condition in the GC theorem above (or you can argue from the remark in the end of section 1). This concludes the asymptotic equicontinuity part of the proof.

The final step of the proof is to show that, the space $\mathcal{F}$ is totally bounded in $L_2(P)$. From the result that we just showed we know that there exists a sequence of finitely discrete measures P_n such that $||(P_n - P)f^2||_{\mathcal{F}_\infty}$ converges to 0. Take n sufficiently large so that the supremum is bounded by ε^2. We know that $N(\varepsilon, \mathcal{F}, L_2(P_n))$ is finite (this can be shown along the lines of the fact shown in the appendix). Any ε-net for $\mathcal{F}$ in $L_2(P_n)$ is a $\sqrt{2}\varepsilon$-net in $L_2(P)$, since $P(f-g)^2 \le \varepsilon^2 + P_n(f-g)^2 \le 2\varepsilon^2$. This concludes the proof.

Example. The set $\mathcal{F}$ of all indicator functions $\mathbb{1}_{(-\infty,t]}$ of cells in $\mathbb{R}$ satisfies:

$$N(\varepsilon, \mathcal{F}, L_2(Q)) \le N_{[]}(\varepsilon^2, \mathcal{F}, L_1(Q)) \le \frac{2}{\varepsilon^2}$$

for any probability measure and $\varepsilon \le 1$, for any probability measure Q. The first inequality follows from the fact that, $\sqrt{Q(f-g)^2} = \sqrt{Q|f-g|}$ so that if $Q|f-g| \le \varepsilon^2$ we would have $\sqrt{Q(f-g)^2} \le \varepsilon$. Therefore $N(\varepsilon, \mathcal{F}, L_2(Q)) \le N(\varepsilon^2, \mathcal{F}, L_1(Q))$. Furthermore if we have a bracket on the set $\mathcal{F}$ – $[l, u]$ we can put a ball with a center at the function $\frac{l+u}{2}$ and radius ε^2, this ball will obviously cover all the functions within the bracket. And thus $N(\varepsilon^2, \mathcal{F}, L_1(Q)) \le N_{[]}(\varepsilon^2, \mathcal{F}, L_1(Q))$. The right inequality follows because the total probability mass is 1, and we are splitting it in intervals of ε^2 we have at most $1/\varepsilon^2 + 1 \le 2/\varepsilon^2$ brackets. Therefore for the uniform entropy in $[0,1]$ we have $\lesssim \int_0^1 \log(1/\varepsilon) d\varepsilon < \infty$. Of course when $\varepsilon > 1$ the number of brackets required is only 1 so that it's 0 in the integral. Thus the class $\mathcal{F}$ would be Donsker if we can show that $\mathcal{F}_\delta$ and $\mathcal{F}^2_\infty$ are P-measurable. This is not hard however, since $\mathbb{Q}$ is dense in $\mathbb{R}$ so we can get to the supremums by countable number of operations.

Compare this result to Slide 55, noteset 2!

We consider much more general classes that satisfy the uniform entropy condition next time.

A Some Details

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

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

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 $\sqrt{6}\sqrt{\frac{1}{n^2}\sum_{i=1}^n f^2(X_i)} = \sqrt{\frac{6}{n}}(\mathbb{P}_n f^2)^{1/2}$.

We now show why $N(\varepsilon, \mathcal{F}_\infty, L_2(Q)) \leq N^2(\varepsilon/2, \mathcal{F}, L_2(Q))$ for all measures Q. Take an $\epsilon/2$ covering of $\mathcal{F}$ consisting of $N = N(\varepsilon/2, \mathcal{F}, L_2(Q))$. Denote with $S = \{f_1, f_2, \ldots, f_N\}$ the covering set. We show that the set $\{f - g : f, g \in S\}$, which are N^2 points, is an ε-cover of $\mathcal{F}_\infty$. Take any point $h \in \mathcal{F}_\infty$. We know that $h = s - t$ for some functions $s, t \in \mathcal{F}$. Take f and g such that $\sqrt{Q(s-f)^2} < \varepsilon/2$ and $\sqrt{Q(t-g)^2} < \varepsilon/2$. Use triangle inequality to conclude that $\sqrt{Q((s-t)-(f-g))^2} < \varepsilon$, which concludes the proof.

Here we show that if $N(\varepsilon||2F||_n^2, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$ is finite for all ε, it must be the case that $\log N(\varepsilon, \mathcal{F}_{\infty,M}^2, L_1(\mathbb{P}_n)) = o_P^*(n)$, in fact this number turns out to also be finite. First note that $N(\varepsilon, \mathcal{F}_{\infty,M}^2, L_1(\mathbb{P}_n)) \leq N(\varepsilon, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$. Fix $\varepsilon > 0$. Since $P^*F^2 < \infty$, then there exists S such that $||2F||_n^2 \leq S$ with probability 1. Thus $N(\varepsilon S, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n)) \leq N(\varepsilon||2F||_n^2, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$ with probability 1, and thus since $\varepsilon > 0$ was arbitrary $N(\varepsilon, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$ is $O_P^*(1)$, and therefore $\log N(\varepsilon, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n)) = o_P^*(n)$, which implies that $\log N(\varepsilon, \mathcal{F}_{\infty,M}^2, L_1(\mathbb{P}_n)) = o_P^*(n)$.