Empirical Processes: Glivenko–Cantelli Theorems

Moulinath Banerjee

June 6, 2010

1 Glivenko–Cantelli classes of functions

The reader is referred to Chapter 1.6 of Wellner's Torgnon notes, Chapter ??? of VDVW and Chapter 8.3 of Kosorok. First, a theorem using bracketing entropy. Let $(\mathcal{F}, \|\ \|)$ be a subset of a normed space of real functions $f : \mathcal{X} \to \mathbb{R}$. Given real functions l and u on $\mathcal{X}$ (but not necessarily in $\mathcal{F}$), the bracket $[l, u]$ is defined as the set of all functions $f \in \mathcal{F}$ satisfying $l \leq f \leq u$. The functions l, u are assumed to have finite norms. An ϵ-bracket is a bracket with $\|u - l\| \leq \epsilon$. The bracketing number $N_{[]}(\epsilon, \mathcal{F} \,\|\, \|)$ is the minimum number of ϵ-brackets with which $\mathcal{F}$ can be covered and the bracketing entropy is the log of this number.

Theorem 1.1 *Let $\mathcal{F}$ be a class of measurable functions with $N_{[]}(\epsilon, \mathcal{F} \,\|\, \|) < \infty$ for all $\epsilon > 0$. Then $\mathcal{F}$ is P-Glivenko-Cantelli, i.e.*

$$\|\mathbb{P}_n - P\|^{\star}_{\mathcal{F}} \to_{a.s} \ 0\,.$$

Brief sketch: For any $\epsilon > 0$ choose finitely many ϵ-brackets $\{l_i, u_i\}_{i=1}^m$ (which can be arranged, by assumption) and argue, by finding a bound on $|(\mathbb{P}_n - P)f|$ (for each f) in terms of the $[l_i, u_i]$ that contains it, that:

$$\sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f| \leq \left\{ \max_{1 \leq i \leq m} (\mathbb{P}_n - P)\, u_i \ \ \vee \ \ \max_{1 \leq i \leq m} (P - \mathbb{P}_n)\, l_i \right\} + \epsilon\,,$$

and conclude, using the strong law for random variables, that the right side of the above display is almost surely less than $2\,\epsilon$ eventually. □

GC theorem for a continuous distribution function on the line: Let F be a continuous cdf and P the corresponding measure. By uniform continuity of F on the line, for every $\epsilon > 0$,

we can find $-\infty = t_0 < t_1 < t_2 < \ldots < t_k < t_{k+1} = \infty$, with k a positive integer, such that the union of the brackets $[1(x \leq t_i), 1(x \leq t_{i+1})]$ for $i = 0, 1, \ldots, k$ contains $\{1(x \leq t : t \in \mathbb{R}\}$ and satisfy $F(t_{i+1}) - F(t_i) \leq \epsilon$. The above theorem now applies directly. Note that the continuity of the distribution function F was used crucially. The GC theorem on the line holds for arbitrary distribution functions though. This more general result will be seen to be a corollary of a subsequent GC theorem.

The next lemma provides a setting which guarantees a finite bracketing number for appropriate classes of functions and finds a ready application in inference in parametric statistical models.

Lemma 1.1 *Suppose that $\mathcal{F} = \{f(\cdot, t) : t \in T\}$, where T is a compact subset of a metric space (D, d) and the functions $f : \mathcal{X} \times T \to \mathbb{R}$ are continuous in t for P–almost $x \in \mathcal{X}$. Assume that the envelope function F defined by $F(x) = \sup_{t \in T} |f(x, t)|$ satisfies $P^\star F < \infty$. Then $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$, for each $\epsilon > 0$.*

The proof is given in Chapter 1.6 of Wellner's Torgnon notes. We skip it but show next how the above result is helpful for deducing consistency in parametric statistical models.

Consistency in parametric models: Let $\{p(x, \theta) : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$ be a class of parametric densities and consider $X_1, X_2, \ldots,$ generated from some $p(x, \theta_0)$. Also assume that Θ is compact and that $p(x, \theta)$ is continuous in θ for P_{θ_0}-almost x. Define $M(\theta) = E_{\theta_0}\, l(X_1, \theta)$ where $l(x, \theta) = \log p(x, \theta)$. Finally assume that $\sup_{\theta \in \Theta} |l(x, \theta)| \leq B(x)$ for some B with $E_{\theta_0}\, B(X_1) < \infty$. Then, note that $M(\theta)$ if finite for all θ and moreover, continuous on Θ. If P_0 denotes the measure corresponding to θ_0, $M(\theta) = P_{\theta_0}\, l(\cdot, \theta)$. The MLE of θ is given by $\hat{\theta}_n = \text{argmax}_\theta\, \mathbb{M}_n(\theta)$ where $\mathrm{M}_n(\theta) = \mathbb{P}_n\, l(\cdot, \theta)$. Under the assumption that the model is identifiable (i.e. the probability distributions corresponding to different θ's are different), it is easily seen that $M(\theta)$ is uniquely minimized at θ_0. Finally, note that θ_0 is a *well-separated* maximizer in the sense that for any $\eta > 0$, $\sup_{\theta \in \Theta \cap B_\eta(\theta_0)^c} M(\theta) < M(\theta_0)$, with $B_\eta(\theta_0)$ being the open ball of radius η centered at θ_0. Let $\psi(\eta) = M(\theta_0) - \sup_{\theta \in \Theta \cap B_\eta(\theta_0)^c} M(\theta)$. Then $\psi(\eta) > 0$.

Our goal is to show that $\hat{\theta}_n \to_{P\ \theta_0^\star} \theta_0$. So, given $\epsilon > 0$, consider $P^\star(\hat{\theta}_n \in B_\epsilon(\theta_0)^c$. Now,

$$\begin{aligned}
\hat{\theta}_n \in B_\epsilon(\theta^0)^c \quad &\Rightarrow \quad M(\hat{\theta}_n) \leq \sup_{\theta \in \Theta \cap B_\eta(\theta_0)^c} M(\theta) \\
&\Leftrightarrow \quad M(\hat{\theta}_n) - M(\theta_0) \leq -\psi(\epsilon) \\
&\Rightarrow \quad M(\hat{\theta}_n) - M(\theta_0) + \mathbb{M}_n(\theta_0) - \mathbb{M}_n(\hat{\theta}_n) \leq -\psi(\epsilon)
\end{aligned}$$

$$\Rightarrow \quad 2 \sup_{\theta \in \Theta} |\mathbb{M}_n(\theta) - M(\theta)| \geq \psi(\epsilon)\,.$$

Thus,

$$P^\star(\hat{\theta}_n \in B_\epsilon(\theta_0)^c) \leq P^\star(\sup_{\theta \in \Theta} |\mathbb{M}_n(\theta) - M(\theta)| \geq \psi(\epsilon)/2) \equiv P^\star(\sup_{\theta \in \Theta} |(\mathbb{P}_n - P_{\theta_0})\, l(\cdot, \theta)| \geq \psi(\epsilon)/2)\,,$$

and this goes to 0, owing to the fact that $(\sup_{\theta \in \Theta} |(\mathbb{P}_n - P_{\theta_0})\, l(\cdot, \theta)|)^\star \;\rightarrow_{a.s.}\; 0$ (since under our assumptions on the parametric model, we can conclude from Lemma 1.1 that $N_{[]}(\eta, \{\mathrm{l}(\cdot, \theta) : \theta \in \Theta\}, L_1(P_{\theta_0})) < \infty$ for every $\eta > 0$ and then invoke Theorem 1.1).

We next state (and partly prove) a result that provides necessary and sufficient conditions for a class of functions $\mathcal{F}$ to be Glivenko-Cantelli in terms of covering numbers.

Theorem 1.2 *Let $\mathcal{F}$ be a P-measurable class of measurable functions bounded in $L_1(P)$. Then $\mathcal{F}$ is P-Glivenko Cantelli if and only if:*
(a) $P^\star F < \infty$,
(b)

$$\lim_{n \to \infty} \frac{E^\star \log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))}{n} = 0\,,$$

for all $M < \infty$ and $\epsilon > 0$. Here $\mathcal{F}_M = \{f\, 1(F \leq M) : f \in \mathcal{F}\}$.

Discussion: We will only consider the 'if' part of the proof. This will be provided later. First, we note that L_2 can be replaced by any $L_r, r \geq 1$. At least for the if part, this will be obvious from the proof. Secondly, for the 'if' part, the second condition can be replaced by the weaker condition that $\log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))/n \rightarrow_{P^\star} 0$. Thirdly, since $N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n)) \leq N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n))$ for all $M > 0$, condition (b) in the theorem can be replaced by the alternative condition that $E^\star(\log N(\epsilon, \mathcal{F}, L_2(\mathbb{P}_n)/n) \rightarrow 0$ (or a condition involving convergence in probability for the 'if' part). Finally, if $\mathcal{F}$ has a measurable and integrable envelope, F, then $\mathbb{P}_n F$ is finite almost surely (simple strong law) and it is readily argued that:

$$\forall \epsilon > 0, (\log N(\epsilon, \mathcal{F}, L_1(\mathbb{P}_n)))^\star = o_p(n) \Leftrightarrow \forall \epsilon > 0, (\log N(\epsilon \|F\|_{\mathbb{P}_n, 1}, \mathcal{F}, L_1(\mathbb{P}_n)))^\star = o_p(n)\,.$$

To see this quickly, use the characterization of in-probability convergence in terms of almost sure convergence along subsequences. It turns out that there is a large class of functions, called VC classes of functions, for which the quantity $\log N(\epsilon \|F\|_{\mathbb{P}_n, 1}, \mathcal{F}, L_1(\mathbb{P}_n))$ is bounded, uniformly in n and ω; in fact, for such a class $\mathcal{F}$ of functions, for:

$$\sup_Q N(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq K_1 \left(\frac{1}{\epsilon^r}\right)^M\,,$$

for an integer $M \geq 1$ that depends solely on $\mathcal{F}$, a constant K_1 that depends only on $\mathcal{F}$, and the supremum is taken over all probability measures for which $\|F\|_{Q,r} > 0$. Thus, a VC class of functions with integrable envelope F is easily Glivenko-Cantelli for any probability measure on the corresponding sample space. The fortunate thing is that functions formed by combining VC classes of functions via various mathematical operations often satisfy similar entropy bounds as in the above display, so that such (more) complex classes continue to remain Glivenko-Cantelli under integrability hypotheses.

As a special case, consider $\mathcal{F} = \{f_t(x) = 1_{-\infty,t]}(x) : t \in \mathbb{R}^d\}$. Thus $f_t(x)$ is simply the indicator of the infinite rectangle to the 'south-west' of the point t. For all probability measures Q on d-dimensional Euclidean space:

$$N(\epsilon, F, L_1(Q)) \leq M_d \left(\frac{K}{\epsilon}\right)^d ,$$

which immediately implies the classical Glivenko-Cantelli theorem in $\mathbb{R}^d$.

Proof of Theorem 1.2: We prove the 'if' part. By P-measurability of the class $\mathcal{F}$ and Corollary 1.1 of the symmetrization notes applied with Φ being the identity,

$$\begin{aligned} E^\star \|\mathbb{P}_n - P\|_{\mathcal{F}} &\leq 2\,E \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}} \\ &= 2\,E_X\,E_\epsilon \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}} \\ &\leq 2\,E_X\,E_\epsilon \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}_M} + 2\,P^\star(F\,1(F > M))\,. \end{aligned}$$

Given any $\epsilon > 0$, an appropriate choice of M ensures that the second term is no larger than ϵ. It suffices to show that for this choice of M, the first term is eventually smaller than ϵ. To this end, first fix $X_1, X_2, \ldots, X_n$. An ϵ-net $\mathcal{G}$ (assumed to be of minimal size) over $\mathcal{F}_M$ in $L_2(\mathbb{P}_n)$ is also an ϵ-net in $L_1(\mathbb{P}_n)$. It follows that:

$$E_\epsilon \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}_M} \leq E_\epsilon \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{G}} + \epsilon\,.$$

Before going further, note that each $g \in \mathcal{G}$ can be assumed to be uniformly bounded (in absolute value) by M. This can be achieved since each f in $\mathcal{F}_M$ is bounded (in absolute value) by M. So,

given an arbitrary ϵ-net $\mathcal{G}$, perturb each g to a $\tilde{g}$ which coincides with g whenever $|g| \leq M$ and on the complement of this set equals $(g) \times M$. These perturbed functions continue to constitute an ϵ-net over $\mathcal{F}_M$.

Consider the first term on the right of the above display. Since the L_1 norm is bounded (up to a constant) by the ψ_1 Orlicz norm, which is bounded upto a constant by the ψ_2 Orlicz norm, we can use Lemma 1.1 in the chaining notes to bound the first term, up to a constant, by:

$$B_n = \sqrt{1 + \log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))} \max_{f \in \mathcal{G}} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\psi_2 | X} .$$

As a consequence of Hoeffding's inequality (see the first page of the symmetrization notes):

$$\left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\psi_2 | X} \leq \sqrt{6} \frac{1}{\sqrt{n}} (\mathbb{P}_n f^2)^{1/2} \leq \sqrt{6} \frac{1}{\sqrt{n}} M ,$$

and thus

$$B_n \leq \sqrt{6}\, M \sqrt{\frac{1 + \log N(\epsilon, \mathcal{F}_M, L_2(\mathbb{P}_n))}{n}} \to 0 ,$$

by Condition (b) of the theorem. Conclude that:

$$E_\epsilon \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}_M} \to_P 0 .$$

Since the above random variable is bounded, conclude that:

$$E_X\, E_\epsilon \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right\|_{\mathcal{F}_M} \to 0 .$$

It follows that $E^\star (\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$. Our goal is however to show almost sure convergence. This is deduced by a submartingale argument, a simplified version of which is presented at the end of these notes. The idea here is to show that $\|\mathbb{P}_n - P\|^\star_{\mathcal{F}}$ is a reverse submartingale with respect to a (decreasing) filtration that converges to the symmetric sigma-field and therefore has an almost sure limit. This almost sure limit, being measurable with respect to the symmetric sigma field, must be a non-negative constant almost surely. The fact that the expectation converges to 0 then forces this constant to be 0. The full undiluted version of the argument is presented in Lemma 2.4.5 of VDVW. □

Uniform and universal GC classes: If $\mathcal{F}$ is P-Glivenko-Cantelli for all probability measures

P on $(\mathcal{X}, \mathcal{A})$, it is called a universal Glivenko-Cantell class. For example, VC classes of functions (that appear in the discussion preceding the proof of Theorem 1.2) are universal GC-classes provided they are uniformly bounded (so that there is an integrable envelope for every probability measure P).

A stronger GC property can be formulated in terms of the uniformity of the convergence of the empirical measure to the true measure over all probability measures on $(\mathcal{X}, \mathcal{A})$. Say that $\mathcal{F}$ is a *strong uniform GC class* if, for all $\epsilon > 0$,

$$\sup_{P \in \mathcal{P}(\mathcal{X}, \mathcal{A})} Pr_P^\star \left(\sup_{m \geq n} \|\mathbb{P}_m - P\|_{\mathcal{F}} > \epsilon \right) \to 0 \,.$$

Note that the almost sure convergence of $\|\mathbb{P}_n - P\|$

* to 0 for a fixed P is equivalent to the condition: For every $\epsilon > 0$,

$$Pr_P^\star \left(\sup_{m \geq n} \|\mathbb{P}_m - P\|_{\mathcal{F}} > \epsilon \right) \to 0 \,.$$

Uniform Glivenko Cantelli classes are sometimes useful in statistical applications, for example in situations where the parent distribution from which a statistical model is generated is allowed to vary with the sample size n, or situations where there are two indices m, n that go to infinity, with n being the sample size, and m an index that labels the statistical model. Consistency arguments for such situations can be constructed via the notion of uniform GC classes of functions. A compelling application is presented in the paper by Sen, Banerjee and Michailidis (2010) (available on Banerjee's webpage) where the problem is one of estimating the minimum effective dose in a dose–response setting (the largest dose beyond which the response is positive) and n is the number of distinct doses with each dose administered to a distinct set of m individuals. Consistency of a least squares estimate of the minimum effective dose is established as $m, n \to \infty$ and the notion of uniform GC classes is heavily used. Section 2.8.1 of VDVW deals with these ideas; see Theorem 2.8.1 which can be used to deduce that VC classes of functions are uniformly Glivenko-Cantelli under appropriate integrability restrictions.

GC preservation: Preservation of GC properties are important from the perspective of applications. Often, in a statistical application, it becomes necessary to show the GC property for a class of functions with complex functional forms to which tailor-made GC theorems are difficult to apply. However, if such classes can be built up from simple GC classes of functions via simple

mathematical operations, the GC property often translates to the complex classes of interest. Section 1.6 of Wellner's notes has a discussion of preservation properties as does Section 9.3 of Kosorok.

Some discussion from Kosorok:

An example: Suppose that $\mathcal{X} = \mathbb{R}$ and that $X \sim P$.

(i) For $0 < M < \infty$ and $a \in \mathbb{R}$, let $f(x,t) = |x-t|$ and $\mathcal{F} = \mathcal{F}_{a,M} = \{f(x,t) : |t-a| \leq M\}$. Show that if $E(|X|) < \infty$, $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$.

Derivation: Chop the interval $[a-M, a+M]$ into an evenly spaced (finite) grid of points $\{s_i\}$ including the end-points such that successive points on the grid are separated by a distance no larger than $\tilde{\epsilon} < \epsilon$. Construct a set of brackets $\{l_j, u_j\}$ where $l_j(x) = |x-s_j|1(x \leq s_j) + |x-s_{j+1}|\,1(x \geq s_{j+1})$ and $u_j(x) = |x-s_j| \vee |x-s_{j+1}|$. Each l_j, u_j has finite norm since $E_P(|X|) < \infty$. A simple picture should now convince you that $u_j - l_j$ is non-negative and no larger than $\tilde{\epsilon}$ pointwise and hence in the $L_1(P)$ norm. Every point t in $[a-M, a+M]$ lies in some $[s_j, s_{j+1}]$ and the function $f(x,t)$ then belongs to the bracket $[l_j, u_j]$, showing that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$.

(ii) Same as before but let $f(x,t) = |x-t| - |x-a|$. Show that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ but without the assumption that $E_P(|X|) < \infty$. □

Derivation: Take the l_j, u_j's constructed above and define $\tilde{u}_j = u_j - |x-a|$ and $\tilde{u}_j = l_j - |x-a|$. Consider $\{\tilde{l}_j, \tilde{u}_j\}$. It is easy to show, using the fact that $|\,|x-t| - |x-t'|\,| \leq |t-t'|$ that each $\tilde{u}_j$ and each $\tilde{l}_j$ is bounded and therefore integrable, irrespective of whether $E(|X|) < \infty$. If $t \in [s_j, s_{j+1}]$, $f(x,t)$ lies in the bracket $[\tilde{l}_j, \tilde{u}_j]$. □