Empirical Processes: Notes 1

Moulinath Banerjee

March 9, 2010

Classical empirical process theory deals with the empirical distribution function based on n i.i.d. random variables. If $X_1, X_2, \ldots, X_n$ are i.i.d. real–valued random variables with distribution function F (and corresponding probability measure P on $\mathbb{R}$), then the empirical distribution function is given by:

$$\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i)\,.$$

The corresponding empirical process is:

$$\mathbb{G}_n(x) = \sqrt{n}(\mathbb{F}_n(x) - F(x))\,.$$

The two basic results concerning $\mathbb{F}_n$ and $\mathbb{G}_n$ are the celebrated Glivenko–Cantelli and Donsker theorems.

Glivenko–Cantelli theorem: We have:

$$\|\mathbb{F}_n - F\|_\infty = \sup_{-\infty<x<\infty} \mid \mathbb{F}_n(x) - F(x) \mid \rightarrow_{a.s} 0\,.$$

See, for example, Billingsley (Prob. & Measure) for a proof. Donsker's theorem deals with the convergence, in distribution, of the empirical process. In what follows I will assume the basic concepts of convergence in distribution for stochastic processes assuming values in a metric space. Billingsley's book on weak convergence (especially the 2nd edition) is an excellent reference (in particular, Chapter 1 for the general theory).

Donsker's Theorem Version 1: Suppose that the common distribution of the X_i's is Uniform(0,1). Then (we can restrict the domain of x in the definition of the process $\mathbb{F}_n(x)$ and $\mathbb{G}_n(x)$ to $[0,1]$ and) $\mathbb{G}_n \Rightarrow \mathbb{G}$ as a process in the space $D[0,1]$ where $D[0,1]$ is the space of cadlag functions on $[0,1]$ equipped with the Borel σ-field generated by the Skorohod topology, and $\mathbb{G}$ is a (tight) Brownian bridge process on $[0,1]$ i.e. $\mathbb{G}$ is a mean 0 Gaussian process indexed by $[0,1]$ with covariance structure $K(s,t) = \mathrm{Cov}(\mathbb{G}(s), \mathbb{G}(t)) = s \wedge t - st$.

Note that the process $\mathbb{G}$ has (uniformly) continuous sample paths (almost surely) with respect to the Euclidean metric and therefore lives, with probability 1, in the (complete separable) subspace $C[0,1]$. This is not true of the converging processes $\mathbb{G}_n$.

A natural question that rears its head in this context is the choice of the function space in which we study distributional convergence of the processes $\mathbb{G}_n$. We can certainly think about each $\mathbb{G}_n$ as an element of the space $l^\infty[0,1]$: the class of bounded functions from $[0,1]$ to $\mathbb{R}$. A natural topology to impose on this space is that corresponding to the uniform metric. More precisely, for $f, g \in l^\infty[0,1]$, $d(f,g) = \|f - g\|_\infty$. This is certainly a much nicer metric to deal with (than the Skorohod, at least in my view). The problem arises from the fact that the processes $\mathbb{G}_n$ are not measurable with respect to the σ-field generated by the uniform metric and therefore the standard theory of weak convergence does not apply. In fact, $l^\infty[0,1]$ under the uniform metric is not even separable, so probability measures defined on this space are not automatically tight.

However, it turns out that the lack of measurability of the processes $\mathbb{G}_n$ does not impede the development of a fruitful theory of weak convergence. The extended theory of weak convergence developed by Dudley and Hoffman-Jorgensen comes to our rescue (see Section 1.3 of Van der Vaart and Wellner for the main results). In the setting of this discussion, weak convergence of the sequence of processes $\mathbb{G}_n$ to a limit process $\mathbb{G}$ (where $\mathbb{G}$ is a measurable random element living in $l^\infty[0,1]$ whose induced probability distribution is tight) amounts to $E^\star(h(\mathbb{G}_n)) \to E(h(\mathbb{G}))$ for all bounded continuous real–valued functions defined on $l^\infty[0,1]$. Here $E^\star$ denotes *outer expectation* (see Section 1.2 of VW) and is the same as the ordinary expectation, provided one has measurability. We will say more about the lack of measurability of $\mathbb{G}_n$ later. The continous mapping theorem holds as before and the power of weak convergence is therefore preserved.

Donsker's Theorem Version 2: Suppose now that the X_i's have a continuous distribution F supported on the entire real line. Consider the processes $\{\mathbb{G}_n f_x : x \in \mathbb{R}\}$ where $\mathcal{F} \equiv \{f_x = 1(-\infty, x](\cdot) : x \in \mathbb{R}\}$ as elements of the space $l^\infty(\mathbb{R})$ Consider the process $\mathbb{G} \circ F$: this is a tight Borel–measurable random element from some underlying probability space into $l^\infty(\mathbb{R})$ with continous sample paths with respect to the Euclidean metric. Then $\mathbb{G}_n \Rightarrow \mathbb{G} \circ F$ as a process in $l^\infty(\mathbb{R})$, with the limit process concentrating on a complete separable subspace of $l^\infty(\mathbb{R})$.

Empirical Processes on General Sample Spaces: The modern theory of empirical processes aims to generalize the classical results to empirical measures defined on general sample spaces ($\mathbb{R}^d$, Riemannian manifolds, spaces of functions..). In other words, the goal is to investigate under what conditions uniform consistency results and distributional convergence results hold for empirical processes that are no longer indexed by simple subsets of the real line, but, say, by general classes of functions. Another goal is to develop efficient methods of verifying that conditions for consistency and possibly distributional convergence are met. This leads to the study of the "complexity" or "largeness" of function classes as measured by what are called "entropy numbers" of different types and VC dimensions. But first, let us lay down a concrete framework.

Let $(\mathcal{X}, \mathcal{A}, P)$ denote the sample space of interest and consider a sequence of random variables $X_1, X_2, \ldots$ that are i.i.d. P. While this is not necessary at this stage, it will be eventually required

to assume that the X_i's are defined as co-ordinate projections on a product space (this guarantees that certain techniques can be employed but more on this later). Thus, we can think of the underlying probability space as $(\Omega = \mathcal{X}^\infty, \mathcal{A}^\infty, P^\infty)$, the generic sample point $\omega = (x_1, x_2, x_3, \ldots)$ and $X_i(\omega) = x_i$. Consider now a class of (bounded) functions $\mathcal{F}$ with domain $\mathcal{X}$ and range $\mathbb{R}$ and envelopem function F. We will define the empirical process indexed by the class of functions $\mathcal{F}$. Let

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

denote the empirical measure for the observed data. Note that we have tacitly assumed points to be measurable. For a measure Q and a real-valued (measurable) function f on $\mathcal{X}$ we write Qf for $\int f \, dQ$. Thus by $\mathbb{P}_n f$ we denote $\int f \, d\mathbb{P}_n = n^{-1} \sum_{i=1}^n f(X_i)$. The empirical process $\mathbb{G}_n(\cdot)$ is viewed as a map from Ω to $l^\infty(\mathcal{F})$ and is defined as:

$$\mathbb{G}_n^\omega(f) = \sqrt{n}(\mathbb{P}_n^\omega - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(f(X_i(\omega)) - Pf\right).$$

Thus, for each fixed ω, $\mathbb{G}_n^\omega(\cdot)$ is a bounded function from $\mathcal{F}$ to $\mathbb{R}$. Typically (as in all probability) the dependence on ω is suppressed and the empirical process is just denoted by $\mathbb{G}_n$. Notice that for fixed f, $\mathbb{G}_n(f)$ is a bona–fide random variable. But as we will see in a second, $\mathbb{G}_n$ is not a measurable tranformation from $(\mathcal{X}^\infty, \mathcal{A}^\infty, P^\infty)$ to $l^\infty(\mathcal{F})$. Also note that the classical empirical distribution function for real valued random variables can be viewed as a special case of this more general set–up with $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{1_{(-\infty, x]} : x \in \mathbb{R}\}$.

Lack of measurability: Consider the case where $\mathcal{X} = [0,1]$, $\mathcal{A}$ is the Borel σ-field and P is Lebesgue measure (the uniform distribution). Then $X_1, X_2, \ldots$ are i.i.d. $U(0,1)$ random variables. Consider the empirical process for $n = 1$ given by

$$\mathbb{G}_1(t) \equiv \mathbb{G}_1(1_{(0,t]}) = 1_{(0,t]}(X_1) - t \equiv \tilde{\mathbb{G}}_1(t) - t\,.$$

Take A to be a non–measurable subset of $[0,1]$ (such a set exists). Consider the subset of $l^\infty[0,1]$ given by $\mathcal{F}_A = \{1_{(0,\cdot]}(s) : s \in A\}$. The uniform distance between any two functions in this set is 1; hence this set is closed in $l^\infty[0,1]$ and therefore measurable. We will show that the map $\omega \mapsto 1_{(0,\cdot]}(X_1(\omega)) \equiv \tilde{\mathbb{G}}_1^\omega(\cdot)$ is not measurable. But $1_{(0,\cdot]}(X_1(\omega)) \in \mathcal{F}_A \Leftrightarrow X_1(\omega) \in A$. So the inverse image of $\mathbb{G}_1(\cdot)$ is $A \times [0,1] \times [0,1] \times \ldots$ which is certainly not measurable.

For general empirical processes, the natural questions are the following: (a) For what classes of functions $\mathcal{F}$ do we have a natural generalization of the Glivenko–Cantelli theorem? (b) For what classes of functions $\mathcal{F}$ do we have a natural generalization of Donsker's theorem?

If $\mathcal{F}$ is a class of functions for which

$$\|\mathbb{P}_n - P\|_{\mathcal{F}}^\star = \left(\sup_{f \in \mathcal{F}} \mid \mathbb{P}_n(f) - P(f) \mid \right)^\star \to_{a.s.} 0$$

then we say that $\mathcal{F}$ is a *P-Glivenko-Cantelli* class of functions. On the other hand, if $\mathcal{F}$ is a class of functions for which

$$\mathbb{G}_n = \sqrt{n}\,(\mathbb{P}_n - P) \Rightarrow \mathbb{G} \quad \text{in } l^\infty(\mathcal{F})$$

where $\mathbb{G}$ is a tight Borel measurable random element, then $\mathcal{F}$ is called a *Donsker* class of functions. Tight Borel measures on $l^\infty(\mathcal{F})$ are completely characterized by their finite dimensional distributions. It follows from the CLT that the finite dimensional distributions $(\mathbb{G}(f_1), \mathbb{G}(f_2), \ldots, \mathbb{G}(f_k))$ are multivariate normal with mean 0 and $\mathrm{Cov}(\mathbb{G}(f_i), \mathbb{G}(f_j)) = P(f_i f_j) - Pf_i\, Pf_j$. Thus $\mathbb{G}$ is a tight Gaussian process.

Define the metric ρ_P as follows:

$$\begin{aligned}\rho_P(f,g) &\equiv (E(\mathbb{G}(f) - \mathbb{G}(g))^2)^{1/2} \\ &= (\mathrm{Var}(f(X_1)) + \mathrm{Var}(g(X_1)) - 2\,\mathrm{Cov}(f(X_1), g(X_1)))^{1/2} \\ &= \mathrm{s.d}(f(X_1) - g(X_1))\,.\end{aligned}$$

From the theory of Gaussian processes it follows (see Page 41 of VW) it follows that the pseudometric space $(\mathcal{F}, \rho_P)$ must be totally bounded and that the process $\mathbb{G}$ must have uniformly continuous sample paths (almost surely) with respect to this metric. Since $\mathbb{G}_n$ converges to a tight limit, the sequence $\{\mathbb{G}_n\}$ must itself be *asymptotically tight* (see, for example, Section 1.3 of VW) which, under the circumstances, is equivalent to the *uniform asymptotic equicontinuity* of the sequence $\mathbb{G}_n$ with respect to the metric ρ_P, which is defined as the following property. For any $\epsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \Pr^\star \left(\sup_{\rho_P(f,g) \le \delta} \mid \mathbb{G}_n(f) - \mathbb{G}_n(g) \mid > \epsilon \right) = 0\,.$$

A motivating example from statistics: A commonly recurring theme in statistics is that we seek to establish consistency or asymptotic normality of some statistic which is not a sum of independent random variables but can be related to some natural sum of random functions indexed by a parameter in a suitable (metric) space. The following example illustrates this idea and was one of the key motivating examples used by Pollard (1989) in his review paper.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. P where P is a probability distribution on the real line. Set $\mu = E\,X_1$. Consider the mean absolute sample deviation; i.e. let

$$M_n = \frac{1}{n} \sum_{i=1}^{N} |X_i - \overline{X}_n|\,.$$

What can we say about the asymptotic distribution of M_n? A natural guess for the in-probability limit of M_n is $M \equiv E\,|X - \mu|$. The next question is: Can we show that $\sqrt{n}\,(M_n - M)$ converges to a normal distribution? This may still not be unreasonable to expect. After all if $\overline{X}_n$ were replaced by μ in the definition of M_n this would be an outcome of the CLT (assuming a finite variance for the X_is) and $\overline{X}_n$ is the natural estimate of μ. Write $M_n = \mathbb{P}_n\,|X - \overline{X}_n|$. For this problem, we need to consider the following class of functions: $\mathcal{F} := \{|x - t| \equiv f_t(x) : t \in [\mu - \delta_0, \mu + \delta_0]\}$ for

some $\delta_0 > 0$. Note that M_n as $\mathbb{P}_n f_{\overline{X}_n} \equiv \mathbb{P}_n f_{\overline{X}_n}(X)$. The crux of the following arguments is that $\overline{X}_n$ concentrates with arbitrarily high probability in $[\mu - \delta_0, \mu + \delta_0]$ as n goes to ∞ and therefore *stochastic properties of the empirical process indexed by $\mathcal{F}$ that hold uniformly* can be exploited to get a handle on $\mathbb{P}_n f_{\overline{X}_n}\}$.

First, consistency. Consider,

$$\begin{aligned} \mathbb{P}_n |X - \overline{X}_n| - P|X - \mu| &= \{\mathbb{P}_n |X - \overline{X}_n| - \mathbb{P}_n |X - \mu|\} + \{\mathbb{P}_n |X - \mu| - P|X - \mu|\} \quad (0.1) \\ &\equiv I_n + II_n \,, \end{aligned}$$

where II_n converges to 0 almost surely by SLLN. That I_n converges to 0 almost surely is an outcome of the fact that

$$I_n \leq \mathbb{P}_n | \, |X - \overline{X}_n| - |X - \mu| \, | \leq \mathbb{P}_n |\overline{X}_n - \mu| = |\overline{X}_n - \mu| \to_{a.s} 0 \,.$$

Next consider,

$$\begin{aligned} \sqrt{n}(M_n - M) &= \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} |X_i - \overline{X}_n| - E|X - \mu| \right] \\ &= \sqrt{n} \, [\mathbb{P}_n f_{\overline{X}_n} - P f_\mu] \\ &= \sqrt{n} \, (\mathbb{P}_n - P) f_\mu + \sqrt{n} \, [\mathbb{P}_n f_{\overline{X}_n} - \mathbb{P}_n f_\mu] \\ &= \sqrt{n} \, (\mathbb{P}_n - P) f_\mu + \sqrt{n} \, (\mathbb{P}_n - P)(f_{\overline{X}_n} - f_\mu) + \sqrt{n} \, (\psi(\overline{X}_n) - \psi(\mu)) \\ &\equiv A_n + B_n + C_n \,, \end{aligned}$$

where $\psi(t) = P f_t = E_P |X - t|$. We will argue later that B_n is asymptotically negligible using an equi-continuity argument. For the moment concentrate on $A_n + C_n$. Assume that P has a Lebesgue density. Then,

$$\psi(t) = \mu - 2 \int_{-\infty}^{t} x f(x) dx - t + 2\, t \, F_P(t)$$

with derivative $(2\, \mathrm{F}_P(t) - 1)$. The delta method gives:

$$A_n + C_n = \mathbb{G}_n f_\mu + \sqrt{n}(\overline{X}_n - \mu) \, \psi'(\mu) + o_p(1) = \mathbb{G}_n (f_\mu(x) + x \, \psi'(\mu)) \,.$$

The usual CLT now gives the limit distribution of $\sqrt{n} \, (M_n - M)$ which is left for you to work out. The only step that remains to be justified is that B_n is $o_p(1)$ and it is in this step that the empirical process techniques kick-in.

Proposition A: Let $\mathcal{F}$ be a Donsker class of functions. Let f_0 be a fixed function and let $\hat{f}_n$ be a random function (depending on $X_1, X_2, \ldots, X_n$) such that $\rho_P(\hat{f}_n, f_0) \to_P 0$. Then,

$$|\mathbb{G}_n \hat{f}_n - \mathbb{G}_n f_0| \to_P 0 \,.$$

Proof: Let $\eta, \epsilon > 0$ be given. Since $\mathcal{F}$ is Donsker, we have *uniform asymptotic equicontinuity*: i.e.

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P(\sup_{\rho_P(f,g) < \delta} |\mathbb{G}_n f - \mathbb{G}_n g| > \eta) = 0 \,.$$

Thus, we can find $\delta_0 > 0$ such that,

$$\limsup_{n \to \infty} P(\sup_{\rho_P(f,g) < \delta_0} |\mathbb{G}_n f - \mathbb{G}_n g| > \eta) < \epsilon \,,$$

showing that for all sufficiently large n,

$$P(\sup_{\rho_P(f,g) < \delta_0} |\mathbb{G}_n f - \mathbb{G}_n g| > \eta) < 2\,\epsilon \,.$$

Let $\Omega_n := \{\rho_P(\hat{f}_n, f_0) < \delta_0\}$. By hypothesis $P(\Omega_n) \geq 1 - \epsilon$ eventually. Let $\tilde{\Omega}_n := \{\sup_{\rho_P(f,g)<\delta_0} |\mathbb{G}_n f - \mathbb{G}_n g| \leq \eta\}$. Then, $P(\tilde{\Omega}_n) \geq 1 - 2\,\epsilon$ eventually. Thus $P(\Omega_n \cap \tilde{\Omega}_n) \geq 1 - 3\,\epsilon$ eventually. But $\Omega_n \cap \tilde{\Omega}_n \subset \{|\mathbb{G}_n \hat{f}_n - \mathbb{G}_n f_0| \leq \eta\}$, showing that $P(|\mathbb{G}_n \hat{f}_n - \mathbb{G}_n f_0| > \eta)$ is eventually less than $3\,\epsilon$. □

Lemma: Let ρ be a 'natural' metric on the class of functions $\mathcal{F}$ with the property that if $\rho(f_n, f_0) \to 0$, then $\rho_P(f_n, f_0)$ goes to 0 as well. If $\rho(\hat{f}_n, f_0) \to_P 0$ for a random function $\hat{f}_n$, then $|\mathbb{G}_n \hat{f}_n - \mathbb{G}_n f_0| \to_P 0$ as well.
This lemma follows directly from the proposition above.

Proposition B: Suppose that $\mathcal{F}$ is a Donsker class of functions indexed by a parameter $\theta \in \Theta$. Thus, $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$. Then $l^\infty(\mathcal{F})$ can be identified with $l^\infty(\Theta)$. Let d be the induced metric on Θ defined by $\tilde{d}(\theta_1, \theta_2) = \rho_P(f_{\theta_1}, f_{\theta_2})$. Let d be a "natural metric" on Θ such that if $d(\theta_n, \theta_0) \to 0$, then $\tilde{d}(\theta_n, \theta_0)$ goes to 0. Suppose that $\hat{\theta}_n$ is a random variable (depending on $X_1, X_2, \ldots, X_n$) such that $d(\hat{\theta}_n, \theta_0) \to_P 0$. Then $|\mathbb{G}_n f_{\hat{\theta}_n} - \mathbb{G}_n f_{\theta_0}| \to_P 0$.

This proposition is a direct consequence of Proposition A and the lemma following it. We are now in a position to show that B_n is $o_p(1)$. We take $\Theta = [\mu - \delta_0, \mu + \delta_0]$. Then $\{f_\theta(x) : \theta \in \Theta\}$ is a P–Donsker class of functions provided $E_P(X^2)$ is finite (which we have assumed), as it is then a VC class of functions with square–integrable envelope function. This fact will need characterizations of Donsker classes that we take up later in the course. Let $\hat{\theta}_n = \overline{X}\, 1(\overline{X} \in [\mu - \delta_0, \mu + \delta_0]) + \mu\, 1(\overline{X} \notin [\mu - \delta_0, \mu + \delta_0])$ and $\theta_0 = \mu$. Now,

$$B_n = \sqrt{n}\,(\mathbb{P}_n - P)(f_{\overline{X}_n} - f_\mu) = (\mathbb{G}_n f_{\hat{\theta}_n} - \mathbb{G}_n f_{\theta_0}) + (\mathbb{G}_n f_{\overline{X}_n} - \mathbb{G}_n f_{\hat{\theta}_n}) \,.$$

The second term is $o_p(1)$ by virtue of the fact that $P(\hat{\theta}_n \neq \overline{X}_n)$ goes to 0. Now take $d(\theta_1, \theta_2)$ to be $|\theta_1 - \theta_2|$; clearly $d(\hat{\theta}_n, \theta_0)$ is $o_p(1)$ and it suffices to show, by Proposition B, that $\tilde{d}(\theta_n, \theta_0)$ is $o(1)$ whenever $d(\theta_n, \theta_0)$ is. So, consider:

$$\begin{aligned} \rho_P^2(f_{\theta_n}, f_{\theta_0}) &= \mathrm{Var}_P\{|X - \theta_n| - |X - \theta_0|\} \\ &\leq P\left[(|X - \theta_n| - |X - \theta_0|)^2\right] \\ &\leq (\theta_n - \theta_0)^2 = o(1) \,. \end{aligned}$$

This finishes the proof.

Another Example – The Binary Choice Model: This example is from Van de Geer's book on 'Empirical Processes in M-Estimation'. Available are i.i.d. data $\{Y_i, Z_i\}_{i=1}^n$ where Y_i assumes the values 1 or 0 (depending on whether, say, the individual has a job) and Z_i is a continuous covariate (say, the education level measured on some continuous scale) on $[0, 1]$. The logit model postulates that

$$P(Y = 1|Z = z) = F_0(\alpha_0 + \theta_0 z)$$

where $F_0(x) = e^x/(1 + e^x)$ is the logistic distribution. In what follows, we assume $\alpha_0 = 0$. Let $\hat{\theta}_n$ denote the MLE of θ_0. The $\hat{\theta}_n$ maximizes

$$\sum_{i=1}^{n} \log p_\theta(Y_i, Z_i) = \sum_{i=1}^{n} \left[Y_i \log F_0(\theta Z_i) + (1 - Y_i) \log(1 - F_0(\theta Z_i)) \right].$$

Letting $l_\theta(y, z) = \log p_\theta(y, z)$, we have: $l_\theta(y, z) = y \log F_0(\theta z) + (1 - y) \log(1 - F_0(\theta z))$ and some algebra shows that:

$$\dot{l}_\theta(y, z) = \frac{\partial}{\partial \theta} l_\theta(y, z) = z(y - F_0(\theta z)).$$

Thus $\hat{\theta}_n$ solves $\mathbb{P}_n \dot{l}_\theta(y, z) = 0$. To study the asymptotic normality of $\hat{\theta}_n$, we introduce the following quantities:

$$m(\theta) = E_{\theta_0}(\dot{l}_\theta(Y, Z)) \text{ and } \sigma^2(\theta) = \mathrm{Var}_{\theta_0}(\dot{l}_\theta(Y, Z)).$$

Note that:

$$\sigma^2(\theta_0) = I(\theta_0) = -E_{\theta_0}(\ddot{l}_\theta(Y, Z)).$$

Next,

$$\left\{ \dot{l}_\theta(y, z) \equiv z \left[y - \frac{e^{\theta z}}{1 + e^{\theta z}} \right] : \theta \in [-K, K] \right\}$$

is a P_{θ_0} Donsker class of functions. This can be established quite easily by using *preservation properties of Donsker classes of functions*. By asymptotic equicontinuity, if θ_n converges to θ_0 in the $\rho_{P_{\theta_0}}$ metric, it must be the case that $\mathbb{G}_n \dot{l}_{\theta_n} - \mathbb{G}_n \dot{l}_{\theta_0} \to 0$. Let us now compute the $\rho_{P_{\theta_0}} \equiv \rho_0$ metric for this problem. We have:

$$\begin{aligned}
\rho_0^2(\theta_1, \theta_2) &= \mathrm{Var}_{\theta_0}[\dot{l}_{\theta_1}(Y, Z) - \dot{l}_{\theta_2}(Y, Z)] \\
&\leq E_{\theta_0}\left[Z^2 \left(Y - \frac{e^{\theta_1 Z}}{1 + e^{\theta_1 Z}} - Y + \frac{e^{\theta_2 Z}}{1 + e^{\theta_2 Z}} \right)^2 \right] \\
&= E_{\theta_0}\left[Z^2 \left(\frac{e^{\theta_1 Z}}{1 + e^{\theta_1 Z}} - \frac{e^{\theta_2 Z}}{1 + e^{\theta_2 Z}} \right)^2 \right].
\end{aligned}$$

A trite application of the mean value theorem which involves computing the derivative of $e^{\theta z}/(1 + e^{\theta z})$ with respect to θ and using the facts that θ_1 and θ_2 lie in a compact set and that Z is bounded shows that the random variable within the square brackets on the right side of the above display is

bounded almost surely, up to a constant, by $(\theta_2 - \theta_1)^2$. It follows immediately that if θ_n converges to θ_0 in the usual Euclidean sense then convergence also happens in the ρ_0 metric, whence the Lemma above can be invoked to conclude that $\mathbb{G}_n \dot{l}_{\hat{\theta}_n} - \mathbb{G}_n \dot{l}_{\theta_0} \to_P 0$ if $\hat{\theta}_n \to \theta_0$ in the Euclidean metric. This fact is not difficult to establish and follows via standard arguments. We are now in a position to wrap up the proof of asymptotic normality. Expanding the identity $\sqrt{n}\,\mathbb{P}_n \dot{l}_{\hat{\theta}_n} = 0$ in a telescoping sum we write:

$$\mathbb{G}_n \dot{l}_{\theta_0} + \mathbb{G}_n (\dot{l}_{\hat{\theta}_n} - \dot{l}_{\theta_0}) + \sqrt{n}\,(P_{\theta_0} \dot{l}_{\hat{\theta}_n} - P_{\theta_0} \dot{l}_{\theta_0}) = 0\,.$$

By what we have shown, the above equality can be rewritten as:

$$\sqrt{n}(\mathbb{P}_n - P)\,\dot{l}_{\theta_0} = -\sqrt{n}(\hat{\theta}_n - \theta_0)\,\frac{m(\hat{\theta}_n) - m(\theta_0)}{\hat{\theta}_n - \theta_0} + o_p(1)\,;$$

since $m'(\theta_0) = -I(\theta_0) \neq 0$ (easily checked),

$$\xi_n \equiv \left(\frac{m(\hat{\theta}_n) - m(\theta_0)}{\hat{\theta}_n - \theta_0) } \right)^{-1} \to_P -I(\theta_0)^{-1}$$

and, in particular, is $O_p(1)$. Multiplying both sides of the display preceding the last by ξ_n we get:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \xi_n\,\sqrt{n}(\mathbb{P}_n - P)\,\dot{l}_{\theta_0} + o_p(1)$$

and it follows immediately that the limit distribution is $N(0, I(\theta_0)^{-1})$ as one expects from classical theory.

1 Problems

- 1. (i) Let F be a distribution function as in the second version of Donsker's Theorem. Explain (with explicit arguments) why the limit distribution of $\sup_{x\in\mathbb{R}} \sqrt{n}|\mathbb{F}_n(x) - F(x)|$ does not depend on F. What about the finite sample distribution?
  (ii) In Donsker's Theorem second version, the class of functions $\mathcal{F}$ indexing the empirical process can be identified with $\mathbb{R}$. However, $\mathbb{R}$ is not totally bounded with respect to the Euclidean distance. Is there a paradox here in view of the discussion on Page 4 of these notes? Explain mathematically.

- 2. Let $\{X_n\}$ be a sequence of stochastic processes assuming values in $l^\infty(T)$ and let ρ be a pseudo-metric on $l^\infty(T)$. Show that the following are equivalent:
  (a) For every $\epsilon > 0$,

$$\lim_{\delta\to 0} \limsup_{n\infty} P^\star(\sup_{\rho(s,t)<\delta} |X_n(s) - X_n(t)| > \epsilon) = 0\,.$$

  (b) For every $\epsilon > 0$ and $\{\delta_n\}$ decreasing to 0,

$$\lim P^\star (\sup_{\rho(s,t)<\delta_n} |X_n(s) - X_n(t)| >) = 0\,.$$

- 3. Show that the empirical process based on i.i.d real-valued random variables $X_1, X_2, \ldots, X_n$ is not measurable when considered as a map from the underlying probability space into $l^\infty(\mathbb{R})$ (equipped with the Borel σ-field corresponding to the uniform metric) *for any* n.