Empirical Processes: Symmetrization

Moulinath Banerjee

October 12, 2010

1 Symmetrization Lemmas

These are comments on the Notes in Chapter 1.5 of Wellner's Torgnon Notes, Section 8.2 of Kosorok and Chapter 2.3 of VDVW. The main results are on inequalities connecting sums of independent processes to symmetrized versions using independent Rademacher variables. Symmetrized versions of empirical properties are sub-Gaussian (in an appropriate sense) whence bounds on the modulus of continuity can be obtained in terms of covering numbers/packing numbers. Sub-Gaussianity of symmetrized empirical processes falls out of Hoeffding's inequality (Lemma 8.7 of Kosorok (2008)).
Hoeffding's inequality: Let $a = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ and let $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ be independent Rademacher variables. Then:

$$P\left(\left|\sum_{i=1}^{n} \epsilon_i a_i\right|\right) > x \leq 2\, e^{-\|x\|^2/2a^2},$$

where $\|\|$ is the Euclidean norm. Hence $\|\sum \epsilon a\|_{\psi_2} \leq \sqrt{6}\,\|a\|$.

Next, the key symmetrization lemma for expectations. We consider sums of independent stochastic processes $\{Z_i(f) : f \in \mathcal{F}\}$. The processes Z_i don't need to have any measurability properties beyond the measurability of all marginals $Z_i(f)$. For computing outer expectations, however, we need to ensure that the underlying probability space is a product space of the form $\prod_{i=1}^{n} (\mathcal{X}_i, \mathcal{A}_i, P_i) \times (\mathcal{Z}, \mathcal{C}, Q)$ and each Z_i is a function of the i'th co-ordinate of $(x, z) = (z_1, z_2, \ldots, z_n)$ only. The Rademacher variables to be used in the symmetrization depend on z only. The empirical process case is obtained by taking $Z_i(f) = f(X_i) - Pf$ where P is the common distribution and will be most relevant to the subsequent development.

Lemma 1.1 *Let $Z_1, Z_2, \ldots, Z_n$ be independent stochastic processes with 0 mean. Then, for any nondecreasing convex function $\Phi : \mathbf{R} \mapsto \mathbf{R}$ and arbitrary functions $\mu_i : \mathcal{F} \mapsto \mathbf{R}$,*

$$E^\star \Phi \left(\frac{1}{2} \left\| \sum_{i=1}^n \epsilon_i Z_i \right\|_{\mathcal{F}} \right) \leq E^\star \Phi \left(\left\| \sum_{i=1}^n Z_i \right\|_{\mathcal{F}} \right) \leq E^\star \Phi \left(2 \left\| \sum_{i=1}^n \epsilon_i (Z_i - \mu_i) \right\|_{\mathcal{F}} \right) .$$

Proof: Let $Y_1, Y_2, \ldots, Y_n$ be independent copies of $X_1, X_2, \ldots, X_n$ defined formally as the co-ordinate projections on the last n co-ordinates in the product space $\prod_{i=1}^n (\mathcal{X}_i, \mathcal{A}_i, P_i) \times (\mathcal{Z}, \mathcal{C}, Q) \times \prod_{i=1}^n (\mathcal{X}_i, \mathcal{A}_i, P_i)$. The outer expectations in the statement of the lemma are unaffected by this enlargement of the original probability space, because co-ordinate projections are perfect maps. First, the inequality on the right. Note that for fixed Z_i's we have:

$$\left\| \sum_{i=1}^n Z_i(f) \right\|_{\mathcal{F}} = \left\| \sum_{i=1}^n Z_i(f) - E\, Y_i(f) \right\|_{\mathcal{F}} \leq E_Y^\star \left\| \sum_{i=1}^n (Z_i(f) - Y_i(f)) \right\|_{\mathcal{F}} .$$

Next,

$$\begin{aligned}
\Phi \left(\left\| \sum_{i=1}^n Z_i(f) \right\|_{\mathcal{F}} \right) &\leq \Phi \left(E_Y^\star \left\| \sum_{i=1}^n (Z_i(f) - Y_i(f)) \right\|_{\mathcal{F}} \right) \\
&= \Phi \left(E_Y \left\| \sum_{i=1}^n (Z_i(f) - Y_i(f)) \right\|_{\mathcal{F}}^{\star, Y} \right) \\
&\leq E_Y \Phi \left(\left\| \sum_{i=1}^n (Z_i(f) - Y_i(f)) \right\|_{\mathcal{F}}^{\star, Y} \right) \\
&= E_Y^\star \Phi \left(\left\| \sum_{i=1}^n (Z_i(f) - Y_i(f)) \right\|_{\mathcal{F}} \right)
\end{aligned}$$

where the $\star, Y$ – the symbol for the smallest measurable majorant with respect to the Y_i's – can be moved out of the argument to Φ since Φ is continuous and non-decreasing (Lemma 6.8 of Kosorok). Taking expectations, now, with respect to the Z_i's, we have:

$$E^\star \Phi \left(\left\| \sum_{i=1}^n Z_i(f) \right\|_{\mathcal{F}} \right) \leq E_Z^\star E_Y^\star \Phi \left(\left\| \sum_{i=1}^n (Z_i(f) - Y_i(f)) \right\|_{\mathcal{F}} \right) \leq E^\star \Phi \left(\left\| \sum_{i=1}^n (Z_i(f) - Y_i(f)) \right\|_{\mathcal{F}} \right) ,$$

where $E^\star$ in the rightmost expression is the joint expectation wrt both Z_i's and Y_i's on the product probability space, by the outer expectation version of Fubini's theorem. Observe now that

$$E^\star \Phi \left(\left\| \sum_{i=1}^n (Z_i(f) - Y_i(f)) \right\|_{\mathcal{F}} \right) = E^\star \Phi \left(\left\| \sum_{i=1}^n e_i (Z_i(f) - Y_i(f)) \right\|_{\mathcal{F}} \right) ,$$

for every n–tuple $(e_1, e_2, \ldots, e_n)$ in $\{-1, 1\}^n$. Conclude that:

$$E^\star \, \Phi \left(\left\| \sum_{i=1}^n Z_i(f) \right\|_{\mathcal{F}} \right) \leq E_\epsilon \, E^\star_{Z,Y} \Phi \left(\left\| \sum_{i=1}^n \epsilon_i \left(Z_i(f) - Y_i(f) \right) \right\|_{\mathcal{F}} \right) .$$

Add and subtract μ_i insider the right side and use the triangle inequality and convexity of Φ to dominate the right side of the above display by:

$$\frac{1}{2} \, E_\epsilon \, E^\star_{Z,Y} \Phi \left(2 \left\| \sum_{i=1}^n \epsilon_i \left(Z_i(f) - \mu_i(f) \right) \right\|_{\mathcal{F}} \right) + \frac{1}{2} \, E_\epsilon \, E^\star_{Z,Y} \Phi \left(2 \left\| \sum_{i=1}^n \epsilon_i \left(Z_i(f) - \mu_i(f) \right) \right\|_{\mathcal{F}} \right) .$$

By perfectness of co-ordinate projections, $E^\star_{Z,Y}$ above is the same as $E^\star_Z$ and $E^\star_Y$ in the two terms respectively. Finally replace the iterated expectation by the joint outer expectation for each of the terms and note that the two resulting terms are equal to complete the proof.

Next we prove the inequality on the left. The joint expectation on the left-most side of the string of inequalities in the statement of the lemma can be written as an iterated expectation:

$$E^\star \, \Phi \left(\frac{1}{2} \left\| \sum_{i=1}^n \epsilon_i \, Z_i(f) \right\|_{\mathcal{F}} \right) = E_\epsilon \, E^\star_Z \, \Phi \left(\frac{1}{2} \left\| \sum_{i=1}^n \epsilon_i \, Z_i(f) - E \, Y_i(f) \right\|_{\mathcal{F}} \right) .$$

For a fixed $(e_1, e_2, \ldots, e_n) \in \{-1, 1\}^n$, consider

$$E^\star_Z \, \Phi \left(\frac{1}{2} \left\| \sum_{i=1}^n e_i \, Z_i(f) - E \, Y_i(f) \right\|_{\mathcal{F}} \right) .$$

Dominate the argument to Φ on the right side of the above display by

$$E^\star_Y \left\| \sum_{i=1}^n \frac{1}{2} e_i \left(Z_i(f) - Y_i(f) \right) \right\|_{\mathcal{F}}$$

and use the same steps as at the beginning of the proof of this lemma to conclude that

$$E^\star_Z \, \Phi \left(\frac{1}{2} \left\| \sum_{i=1}^n e_i \, Z_i(f) \right\|_{\mathcal{F}} \right) \leq E^\star_{Z,Y} \, \Phi \left\| \sum_{i=1}^n \frac{1}{2} \, e_i \left(Z_i(f) - Y_i(f) \right) \right\|_{\mathcal{F}} .$$

But this last expression is invariant to the choice of $(e_1, e_2, \ldots, e_n)$ and therefore

$$E^\star_Z \, \Phi \left(\frac{1}{2} \left\| \sum_{i=1}^n e_i \, Z_i(f) \right\|_{\mathcal{F}} \right) \leq E^\star_{Z,Y} \, \Phi \left\| \sum_{i=1}^n \frac{1}{2} \left(Z_i(f) - Y_i(f) \right) \right\|_{\mathcal{F}} .$$

It follows that:

$$E^\star\,\Phi\left(\frac{1}{2}\left\|\sum_{i=1}^n \epsilon_i\, Z_i(f)\right\|_{\mathcal{F}}\right) \leq E^\star_{Z,Y}\,\Phi\left\|\sum_{i=1}^n \frac{1}{2}\left(Z_i(f) - Y_i(f)\right)\right\|_{\mathcal{F}}.$$

As before, we now use the triangle inequality inside the argument to Φ followed by Jensen's inequality to complete the proof. $\square$

We return to the setting of empirical processes. Define

$$\mathbb{P}^0_n(f) = \frac{1}{n}\sum_{i=1}^n \epsilon_i\, f(X_i) \;\text{ and }\; \mathbb{P}^\dagger_n(f) = \frac{1}{n}\sum_{i=1}^n \epsilon_i\left(f(X_i) - Pf\right).$$

An important special corollary obtains by choosing $Z_i(f) = f(X_i) - Pf$ and $\mu_i(f) = 0$ as well as $\mu_i(f) = -Pf$.

Corollary 1.1

$$E^\star\,\Phi(\|\mathbb{P}^\dagger_n\|_{\mathcal{F}}/2) \leq E^\star\,\Phi(\|\mathbb{P}_n - P\|_{\mathcal{F}}) \leq E^\star\,\Phi(2\|\mathbb{P}^0_n\|_{\mathcal{F}}) \;\wedge\; E^\star\,\Phi(2\|\mathbb{P}^\dagger_n\|_{\mathcal{F}}).$$

The above lemma and its corollary appear in Section 1.5 of Wellner's Torgnon notes. The lemma appears also in Section 2.3.2 of VDVW and a special case of the lemma as Lemma 2.3.1 of VDVW. This last case is, perhaps, of most interest and is generally used for $\Phi(x) = x$. A slight variant of these lemmas, in the setting of empirical processes, appears as Theorem 8.8 of Kosorok where the rightmost inequality is somewhat different.

Some measurability considerations: These symmetrization results will be most useful to us when the supremum $\|\mathbb{P}^0_n\|_{\mathcal{F}}$ is measurable and Fubini's theorem allows us to write the joint expectation as an iterated expectation; first, with respect to ϵ holding X fixed and then with respect to X. The conditional expectation given X will be controlled using the entropy bound on sub-Gaussian processes. This technique is at the heart of a Donsker theorem involving entropy integrals where the class $\mathcal{F}$ will be a set of functions formed by differencing pairs of functions belonging to the original class and satisfying other appropriate constraints. We will therefore require some measurability hypothesis on the classes of functions over which we seek to prove uniform central limit theorems. One way to formulate such hypotheses is through the notion of a P-measurable class discussed in VDVW (Section 2.3) and Kosorok (Pages 141-144). A somewhat stronger hypothesis that guarantees P-measurability is that of *pointwise measurable*

classes of functions, denoted PM in Kosorok. PM is readily satisfied by many function classes, like indicators of cells in Euclidean space, indicators of balls etc. Kosorok provides a good discussion of the preservation of the PM property under different mathematical operations. In statistical applications, the measurability of the supremums will generally not be an issue.

Symmetrization for probabilities: We will not discuss this in any great detail. Results from symmetrization for probabilities don't follow from the results established above since the function $\Phi(x) = 1(x > a)$ is non-convex for any choice of a. Nevertheless, important results which relate the tail behavior of the supremum of sums of independent processes as considered above to their symmetrized (and centered) versions exist. See Lemma 2.3.7 of VDVW in this context. The subsequent discussion in that book is also of interest. In particular, it is established that the stochastic equicontinuity condition needed for a class of functions to be Donsker has an equivalent version in terms of expectations. If you go back to the corollary at the end of the notes on general weak convergence theory, the asymptotic equicontinuity condition is:

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P^\star(\|\mathbb{G}_n\|_{\mathcal{F}_\delta} > \epsilon) = 0\,,$$

for any pre-assigned $\epsilon > 0$, with $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F} \text{ and } \rho_2(f, g) < \delta\}$. This can be equivalently stated as: For any given $\epsilon > 0$, for every $\delta_n \to 0$,

$$\lim_{\delta_n \to 0} P^\star(\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} > \epsilon) = 0\,.$$

Lemma 2.3.11 in VDVW (via its corollary 2.3.12) shows that the above display is equivalent to:

$$\lim_{\delta_n \to 0} E^\star(\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}}) = 0\,.$$

An extensive discussion of symmetrization elaborating on some aspects of that in VDVW is available in the handwritten supplementary notes on symmetrization posted on the webpage.

Exercise: Show that the asymptotic equicontinuity condition in the corollary referred to above is equivalent to

$$\lim_{\delta_n \to 0} P^\star(\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} > \epsilon) = 0\,,$$

for any pre-assigned $\epsilon > 0$.

Proof: Assume the asymptotic equicontinuity condition. Let $\epsilon > 0$ be preassigned and consider $\delta_n \to 0$. Fix an $\eta > 0$. By our assumption, there exists a $\delta_0 > 0$ such that $\limsup_{n\to\infty} P^\star(\|G_n\|_{\mathcal{F}_{\delta_0}} > \epsilon) < \eta/2$. But this implies that for all sufficiently large n, $P^\star(\|G_n\|_{\mathcal{F}_{\delta_0}} > \epsilon) < \eta$. Since δ_n is

eventually less than δ_0 and $\|G_n\|_{\mathcal{F}_\delta}$ is decreasing in δ, for all sufficiently large n, $P^\star(\|G_n\|_{\mathcal{F}_{\delta_n}} > \epsilon) < \eta$.

Now for the converse. Need to show that given $\tau > 0$, there exists δ_0 such that $\limsup_{n\to\infty} P^\star(\|\mathbb{G}_n\|_{F_{\delta_0}} > \epsilon) < \tau$. Suppose not. Then, there is a sequence $\xi_m \to 0$ and a $\tau_0 > 0$, for which $\limsup_{n\to\infty} P^\star(\|\mathbb{G}_n\|_{\mathcal{F}_{\xi_m}} > \epsilon) \geq \tau_0$. Now, we can choose an $m_1 \geq 1$ for which $P^\star(\|\mathbb{G}_{m_1}\|_{\mathcal{F}_{\xi_1}} > \epsilon) \geq \tau_0/2$. Next, choose an $m_2 > m_1$ such that $P^\star(\|\mathbb{G}_{m_2}\|_{\mathcal{F}_{\xi_2}} > \epsilon) \geq \tau_0/2$ and proceed in this fashion, to create a sequence $m_1, m_2, m_3, \ldots$ and so on. Choose your sequence δ_n to be ξ_1 for the first m_1 entries, ξ_2 for the next $m_2 - m_1$ entries and so on and note that $P^\star(\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} > \epsilon)$ does not converge to 0 for this sequence. $\square$