Stat210B: Theoretical Statistics **Lecture Date: February 26, 2008**

The Chaining Lemma

Lecturer: Michael I. Jordan *Scribe: Fabian Wauthier*

Recall from last time the definition of a P-Donsker class.

Definition. A class of functions is called *P-Donsker* if $\mathbb{G}_n$ converges weakly to a tight limit process in $l^\infty(\mathcal{F})$, which is a P-Brownian bridge $\mathbb{G}_p$ with zero mean and covariance function $E(\mathbb{G}_p f \mathbb{G}_p g) = Pfg - PfPg$. Here the empirical process $\mathbb{G}_n$ is defined as $\mathbb{G}_n = \sqrt{n}(P_n - P)$. This means in particular, that for any finite collection of functions, the elements $\mathbb{G}_n f$ converge to a zero mean multivariate Gaussian, with aforementioned covariance function.

Furthermore, recall Theorem 19.5 stated last time.

Theorem 19.5 (Donsker). *Every class $\mathcal{F}$ of measurable functions with $J_{[]}(1, \mathcal{F}, L_2(P)) < \infty$ is P-Donsker. Here we defined $J_{[]}(\delta, \mathcal{F}, L_2(P)) = \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2(P))} d\varepsilon$.*

In this lecture we will be concerned mostly with proving the Chaining Lemma, which is instrumental to the proof of this theorem. Before commencing the presentation, we first illustrate some properties of P-Donsker classes.

Combining P-Donsker classes

The definition of P-Donsker classes gives rise to an algebra for combining any two P-Donsker classes. In particular, suppose that $f \in \mathcal{F}$ and $g \in \mathcal{G}$ are both P-Donsker. If $\phi(\cdot, \cdot)$ is a Lipschitz transformation, then $\phi(f, g)$ is P-Donsker. Examples of such Lipschitz transformations include: $f + g$, $f \wedge g$, $f \vee g$, fg if $\mathcal{F}$ and $\mathcal{G}$ are uniformly bounded, and $1/f$ if $\mathcal{F}$ is bounded away from zero.

Chaining Lemma

In this lecture we give a thorough treatment of the core of empirical process theory by proving the Chaining Lemma (lemma 19.34 in van der Vaart). The presentation is based on section 19.6 in van der Vaart (1998). We begin by stating two relevant lemmas. The first one, Bernstein's inequality, represents a tightening of the Hoeffding bound we previously discussed. This strengthening will be required for the following argument.

Lemma 19.32 (Bernstein's inequality). *For one function f and any $x > 0$,*

$$P(|\mathbb{G}_n f| > x) \leq 2 \exp\left\{ -\frac{1}{4} \frac{x^2}{Pf^2 + x||f||_\infty / \sqrt{n}} \right\}. \tag{1}$$

Note that as in Hoeffding, the upper bound is twice the exponential of some function. Here, the Pf^2 term in the exponential accounts for something like the variance, whereas in Hoeffding there was an upper bound on the variance through terms $\sum_i (b_i - a_i)^2$. An additional term has also been introduced to the denominator.

The next lemma will relate Bernstein's inequality to finite collections.

Lemma 19.33. *For any finite class $\mathcal{F}$ of bounded, measurable and square-integrable functions, with $|\mathcal{F}|$ elements, we have*

$$E(||\mathbb{G}_n||_{\mathcal{F}}) \lesssim \max_f \frac{||f||_\infty}{\sqrt{n}} \log(1+|\mathcal{F}|) + \max_f ||f||_{P,2}\sqrt{\log(1+|\mathcal{F}|)}. \tag{2}$$

Here, we have adopted the notation $\lesssim$ to express that the left hand side is less than the right hand side, up to a universal multiplicative constant. The proof idea behind this Lemma lies in breaking the left hand side into two pieces using the triangle inequality, and then applying Bernstein's inequality to both.

We now turn to the Chaining Lemma. The motivation for this lemma lies in the difficulty of carrying out an independent analysis of fluctuations for each element f of an uncountably infinite set of functions $\mathcal{F}$. To get control over the infinite set, we need to tie functions together to a finite number of grid cells. We can introduce suitable structure on $\mathcal{F}$ via a multi-resolution grid. At the coarse top level very few cells partition $\mathcal{F}$; at progressively deeper levels each grid cell is partitioned into a set of smaller cells. By choosing one representative function for each grid cell, the fluctuations between any two functions in $\mathcal{F}$ can be related to fluctuations along edges on the grid tree.

Lemma 19.34 (Chaining Lemma). *Define* $\text{Log } x = 1 \vee \log(x)$ *and* $a(\delta) = \delta/\sqrt{\text{Log } N_{[]}(\delta, \mathcal{F}, L_2(P))}$. *For any class $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$ so that, for some common δ^2, $Pf^2 \leq \delta^2, \forall f \in \mathcal{F}$, and F an envelope function,*

$$E(||\mathbb{G}_n||_{\mathcal{F}}) \lesssim J_{[]}(\delta, \mathcal{F}, L_2(P)) + \sqrt{n}PF\left\{F > \sqrt{n}a(\delta)\right\}. \tag{3}$$

Proof. We begin the proof by focussing on the first term on the right hand side. For $|f| \leq g$ by the triangle inequality

$$\begin{aligned} |\mathbb{G}f| &= \sqrt{n}|P_n f - Pf| && (4)\\ &\leq \sqrt{n}(P_n|f| + P|f|) && (5)\\ &\leq \sqrt{n}(P_n g + Pg). && (6) \end{aligned}$$

This implies that for an envelope function F

$$\begin{aligned} E\left(||\mathbb{G}_n f\left\{F > \sqrt{n}a(\delta)\right\}||_{\mathcal{F}}\right) &\leq \sqrt{n}E\left(P_n F\left\{F > \sqrt{n}a(\delta)\right\} + PF\left\{F > \sqrt{n}a(\delta)\right\}\right) && (7)\\ &= 2\sqrt{n}PF\left\{F > \sqrt{n}a(\delta)\right\}. && (8) \end{aligned}$$

This demonstrates the inequality for the second term on the right hand side. We continue the derivation on $||\mathbb{G}_n f\left\{F \leq \sqrt{n}a(\delta)\right\}||$ and show that it is less than or equal to $J_{[]}(\delta, \mathcal{F}, L_2(P))$. Since the set of remaining functions we work with has shrunk, it has smaller bracketing number than $\mathcal{F}$. For notational convenience, continue by assuming that $f \leq \sqrt{n}a(\delta), \forall f \in \mathcal{F}$. At this point we turn to the multi-resolution structure on $\mathcal{F}$ which we previously noted. Choose an integer q_0 such that $4\delta \leq 2^{-q_0} \leq 8\delta$. Also choose a nested sequence of partitions $\mathcal{F}_{q_i}$ of $\mathcal{F}$ indexed by integers $q \geq q_0$; that is, if at level q there are N_q disjoint sets, then $\mathcal{F} = \cup_{i=1}^{N_q}\mathcal{F}_{q_i}$. Choose this nested sequence of partitions and measurable functions $\Delta_{q_i} \leq 2F$, so that

$$\begin{aligned} \sum_{q \geq q_0} 2^{-q}\sqrt{\text{Log } N_q} &\lesssim \int_0^\delta \sqrt{\text{Log } N_{[]}(\varepsilon, \mathcal{F}, L_2(P))}d\varepsilon && (9)\\ \sup_{f,g \in \mathcal{F}_{q_i}} |f - g| \leq \Delta_{q_i}, &\quad P\Delta_{q_i}^2 < 2^{-2q}. && (10) \end{aligned}$$

The functions Δ_{q_i} are the difference between upper and lower brackets and act as envelopes.

We continue by choosing a representative function within each cell of each level. Fix for each level $q > q_0$ and each partition $\mathcal{F}_{q_i}$ one representative f_{q_i} and define, if $f \in \mathcal{F}_{q_i}$

$$\pi_q f = f_{q_i} \text{ (Nearest neighbor function)} \tag{11}$$
$$\Delta_q f = \Delta_{q_i}. \tag{12}$$

Here is where $\mathcal{F}$ is attributed a finite representation. At scale q, $\pi_q f$ and $\Delta_q f$ run over N_q functions as f runs over $\mathcal{F}$. Define

$$a_q = 2^{-q}/\sqrt{\text{Log } N_{q+1}}, \tag{13}$$
$$A_{q-1} f = \mathbb{I}\left\{\Delta_{q_0} f \le \sqrt{n} a_{q_0}, \ldots, \Delta_{q-1} f \le \sqrt{n} a_{q-1}\right\}, \tag{14}$$
$$B_q f = \mathbb{I}\left\{\Delta_{q_0} f \le \sqrt{n} a_{q_0}, \ldots, \Delta_{q-1} f \le \sqrt{n} a_{q-1}, \Delta_q f > \sqrt{n} a_q\right\}. \tag{15}$$

Now decompose the difference between any f and the representative $\pi_{q_0} f$ using the newly defined sets as a telescoping sum,

$$f - \pi_{q_0} f = \sum_{q_0+1}^{\infty} (f - \pi_q f) B_q f + \sum_{q_0+1}^{\infty} (\pi_q f - \pi_{q-1} f) A_{q-1} f. \tag{16}$$

We observe that either all of the $B_q f$ are zero[1] in which case the $A_{q-1} f$ are 1 (we always have small fluctuations). Alternatively, one $B_{q_1} f = 1$ for some $q_1 > q_0$ (and zero for all other q), in which case $A_q f = 1$ for $q < q_1$ and $A_q f = 0$ for $q \ge q_1$. In that last case we have a sequence of small fluctuations, followed by one large fluctuation

$$f - \pi_{q_0} f = (f - \pi_{q_1} f) + \sum_{q_0+1}^{q_1} (\pi_q f - \pi_{q-1} f) A_{q-1} f. \tag{17}$$

By the construction of partitions and our choice of q_0 we have

$$2a(\delta) = \frac{2\delta}{\sqrt{\text{Log } N_{[]}(\delta, \mathcal{F}, L_2(P))}} \tag{18}$$
$$\le \frac{2^{-q_0}}{\sqrt{\text{Log } N_{q_0+1}}} \tag{19}$$
$$= a_{q_0}. \tag{20}$$

This implies that $\Delta_{q_0} f \le a_{q_0}\sqrt{n}$ and therefore $A_{q_0} f = 1$. Furthermore, nesting implies $\Delta_q f B_q f \le \Delta_{q-1} f B_q f \le \sqrt{n} a_{q-1}$. The last inequality holds if $B_q f = 0$ and also if $B_q f = 1$ by definition. It follows that since $B_q f$ is an indicator where $\Delta_q f > \sqrt{n} a_q$ that $\sqrt{n} a_q P(\Delta_q f B_q f) \le P(\Delta_q f B_q f)^2 = P(\Delta_q f)^2 B_q f \le 2^{-2q}$ by the choice of $\Delta_q f$. We now apply the empirical process $\mathbb{G}_n$ to both series on the right of the equation 16 and use the triangular inequality on the supremum over absolute values. Because $|\mathbb{G}_n f| \le \mathbb{G}_n g + 2\sqrt{n} P g$ for $|f| < g$ we get, by applying Lemma 19.33

$$E\left\|\sum_{q_0+1}^{\infty} \mathbb{G}_n (f - \pi_q f) B_q f\right\|_{\mathcal{F}} \le \sum_{q_0+1}^{\infty} E||\mathbb{G}_n \Delta_q f B_q f||_{\mathcal{F}} + \sum_{q_0+1}^{\infty} 2\sqrt{n}||P\Delta_q f B_q f||_{\mathcal{F}} \tag{21}$$
$$\overset{19.33}{\lesssim} \sum_{q_0+1}^{\infty} \left[a_{q-1}\text{Log } N_q + 2^{-q}\sqrt{\text{Log } N_q} + \frac{4}{a_q} 2^{-2q}\right]. \tag{22}$$

We note that the third term arises in part from our earlier observation that $P(\Delta_q f B_q f) \le 2^{-2q}/\sqrt{n} a_q$. However, it was unclear in class where the additional factor of 2 stems from. All three terms in the infinite

[1] There is a typo in van der Vaart (1998) page 287, where the author states that "either all $B_q f$ are *1*".

sum will become essentially like the middle one, which we know from inequality 9 can be bounded by a multiple of $J_{[]}(\delta, \mathcal{F}, L_2(P))$. Thus we have bounded one more term.

To establish a similar bound for the second part of equation 16, note that there are at most N_q functions $\pi_q f - \pi_{q-1} f$ and at most N_{q-1} indicators $A_{q-1} f$. Nesting implies $|\pi_q f - \pi_{q-1} f| A_{q-1} f \leq \Delta_{q-1} f A_{q-1} f \leq \sqrt{n} a_{q-1}$. The $L_2(P)$ norm of $|\pi_q f - \pi_{q-1} f|$ is upper bounded by $2^{-(q+1)}$. Now using Lemma 19.33 we find that

$$E \left|\left| \sum_{q_0+1}^{\infty} \mathbb{G}_n (\pi_q f - \pi_{q-1} f) A_{q-1} f \right|\right|_{\mathcal{F}} \lesssim \sum_{q_0+1}^{\infty} \left[a_{q-1} \text{Log } N_q + 2^{-q} \sqrt{\text{Log } N_q} \right]. \tag{23}$$

As before, note that the first and second terms on the right are identical and that each can be bounded by a multiple of $J_{[]}(\delta, \mathcal{F}, L_2(P))$.

As the final step in this proof we need to establish a bound for terms $\pi_{q_0} f$. Note that for the envelope function F, we have $|\pi_{q_0} f| \leq F$. Also, recall that since early in the derivation we are only considering the class of functions $f \{F \leq \sqrt{n} a(\delta)\}$ where f ranges over $\mathcal{F}$, so that $F \leq \sqrt{n} a(\delta)$. Moreover, $\sqrt{n} a(\delta) \leq \sqrt{n} a_{q_0}$ by a similar argument as in derivation 18-20. Recall also that one of the preconditions of this lemma is that $Pf^2 < \delta^2, \forall f \in \mathcal{F}$, so that in particular $P(\pi_{q_0} f)^2 \leq \delta^2$. Applying Lemma 19.33 again, we find that

$$E ||\mathbb{G}_n \pi_{q_0} f||_{\mathcal{F}} \lesssim a_{q_0} \text{Log } N_{q_0} + \delta \sqrt{\text{Log } N_{q_0}}. \tag{24}$$

By the choice of q_0 at the onset and inequality 9, both terms can be bounded by a multiple of $J_{[]}(\delta, \mathcal{F}, L_2(P))$.

This concludes the proof of Lemma 19.34. We summarise briefly. The proof was carried out by using an envelope function F to split the function space $\mathcal{F}$ into two sets. In inequality 8 we quickly saw that one set gives rise to one of the terms in the final result. We then defined a multi-resolution tree on the remaining subset of $\mathcal{F}$ so that we could consider fluctuations via suitably defined events $A_{q-1} f$ and $B_q f$. In the following we repeatedly applied Lemma 19.33 to yield inequalities 22, 23, and 24, each of which can be upper bounded by a multiple of $J_{[]}(\delta, \mathcal{F}, L_2(P))$. In the final result, these three parts are represented by one copy of $J_{[]}(\delta, \mathcal{F}, L_2(P))$. □

References

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.