SECTION 1

Introduction

As it has developed over the last decade, abstract empirical process theory has largely been concerned with uniform analogues of the classical limit theorems for sums of independent random variables, such as the law of large numbers, the central limit theorem, and the law of the iterated logarithm. In particular, the Glivenko-Cantelli Theorem and Donsker's Theorem, for empirical distribution functions on the real line, have been generalized and extended in several directions. Progress has depended upon the development of new techniques for establishing maximal inequalities for sums of independent stochastic processes. These inequalities can also be put to other uses in the asymptotic theory of mathematical statistics and econometrics. With these lecture notes I hope to explain some of the theoretical developments and illustrate their application by means of four nontrivial and challenging examples.

The notes will emphasize a single method that has evolved from the concept of a Vapnik-Červonenkis class of sets. The results attained will not be the best possible of their kind. Instead I have chosen to strive for just enough generality to handle the illustrative examples without having to impose unnatural extra conditions needed to squeeze them into the framework of existing theory.

Usually the theory in the literature has concerned independent (often, also identically distributed) random elements $\xi_1, \xi_2, \dots$ of an abstract set Ξ. That is, for some σ-field on Ξ, each ξ_i is a measurable map from a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ into Ξ. For each n, the $\{\xi_i\}$ define a random probability measure on the set Ξ: the *empirical measure* P_n puts mass $1/n$ at each of the points $\xi_1(\omega), \dots, \xi_n(\omega)$. Each real-valued function f on Ξ determines a random variable,

$$P_n f = \frac{1}{n} \sum_{i \leq n} f(\xi_i(\omega)).$$

For fixed f, this is an average of independent random variables, which, under appropriate regularity conditions and with the proper standardizations, will satisfy a law of large numbers or a central limit theorem. The theory seeks to generalize these classical results so that they hold uniformly (in some sense) for f ranging over various classes $\mathcal{F}$ of functions on Ξ.

In asymptotic problems, $\mathcal{F}$ is often a parametric family, $\{f(\cdot,t) : t \in T\}$, with T not necessarily finite dimensional. One can then simplify the notation by writing $f_i(\omega,t)$ instead of $f(\xi_i(\omega),t)$. In my opinion, this is the most natural notation for the methods that will be developed in these notes. It accommodates gracefully applications where the function f is allowed to change with i (or n). For example, in Section 11 we will encounter a triangular array of processes,

$$f_{ni}(\omega,t) = |y_i(\omega)^+ - (x_i'\theta_0 + z_{ni}'t)^+| - |y_i(\omega)^+ - (x_i'\theta_0)^+| \qquad \text{for } i = 1,\ldots,n,$$

generated by a reparametrization of a censored regression. The $\{z_{ni}\}$ will be constructed from the deterministic vectors $\{x_i\}$ by means of a transformation that depends on n. Such processes do not fit comfortably into the traditional notation, but their analysis depends on the same symmetrization and conditioning arguments as developed in the literature for the empirical measure P_n.

The notation also allows for transformations that depend on i, as with the $f_i(\omega,t)/i$ that will appear in Section 8. It also eliminates an unnecessary notational distinction between empirical processes and partial-sum processes, bringing both closer to the theory for sums of independent random elements in Banach space. In these notes, however, I will concentrate on problems and methods that are usually identified as belonging to empirical process theory.

The general problem to be attacked in the next six sections will be that of finding probabilistic bounds for the maximal deviation of a sum of independent stochastic processes,

$$S_n(\omega,t) = \sum_{i\le n} f_i(\omega,t),$$

from its expectation,

$$M_n(t) = \mathbb{P}S_n(\cdot,t) = \sum_{i\le n} \mathbb{P}f_i(\cdot,t).$$

That is, we will seek to bound $\Delta_n(\omega) = \sup_{t\in T} |S_n(\omega,t) - M_n(t)|$. In applications the f_i will often acquire a second subscript to become a triangular array. But, since most of the argument is carried out for fixed n, there is no need to complicate the notation prematurely.

For a general convex, increasing function Φ on $\mathbb{R}^+$, Section 2 will derive a bound for $\mathbb{P}\,\Phi(\Delta_n)$. The strategy will be to introduce a more variable process,

$$L_n(\boldsymbol{\sigma},\omega) = \sup_t \Big|\sum_{i\le n} \sigma_i f_i(\omega,t)\Big|,$$

defined by means of a new sequence of independent random variables $\{\sigma_i\}$, each σ_i taking only the values $+1$ and -1, both with probability $1/2$. We will find that $\mathbb{P}\,\Phi(\Delta_n)$ is less than $\mathbb{P}\,\Phi(2L_n)$.

With ω held fixed, L_n is a very simple process indexed by a subset of $\mathbb{R}^n$,

$$\mathcal{F}_\omega = \{(f_1(\omega,t),\ldots,f_n(\omega,t)) : t \in T\}.$$

The indexing of the points of $\mathcal{F}_\omega$ by T will become irrelevant; the geometry of $\mathcal{F}_\omega$

will be all that matters. In terms of the usual inner product on $\mathbb{R}^n$,

$$L_n(\boldsymbol{\sigma}, \omega) = \sup_{\mathbf{f} \in \mathcal{F}_\omega} |\boldsymbol{\sigma} \cdot \mathbf{f}|.$$

Section 3 will establish a general inequality for processes like this, but indexed by fixed subsets of $\mathbb{R}^n$; it will be applied conditionally to L_n. The inequality will take the form of a bound on an *Orlicz norm.*

If Φ is a convex, increasing function on $\mathbb{R}^+$ with $0 \le \Phi(0) < 1$, the Orlicz norm $\|Z\|_\Phi$ of a random variable Z is defined by

$$\|Z\|_\Phi = \inf\{C > 0 : \mathbb{P}\Phi(|Z|/C) \le 1\},$$

with $+\infty$ as a possible value for the infimum. If $\mathbb{P}\Phi(|Z|/C_0) < \infty$ for some finite C_0, a dominated convergence argument shows that $\mathbb{P}\Phi(|Z|/C) \to \Phi(0) < 1$ as $C \to \infty$, which ensures that $\|Z\|_\Phi$ is finite. If one identifies random variables that are equal almost everywhere, $\|\cdot\|_\Phi$ defines a norm on the space $\mathcal{L}^\Phi$ of all random variables Z for which $\|Z\|_\Phi < \infty$. (The space $\mathcal{L}^\Phi$ is even complete under this norm, a property we will not need.) In the special case where $\Phi(x) = x^p$ for some $p \ge 1$, the norm $\|\cdot\|_\Phi$ coincides with the usual $\|\cdot\|_p$, and $\mathcal{L}^\Phi$ is the usual space of random variables with finite p^{th} absolute moments. Finiteness of $\|Z\|_\Phi$ places a constraint on the rate of decrease for the tail probabilities via the inequality

$$\begin{aligned} \mathbb{P}\{|Z| \ge t\} &\le \mathbb{P}\Phi(|Z|/C)/\Phi(t/C) \\ &\le 1/\Phi(t/C) \qquad \text{if } C = \|Z\|_\Phi. \end{aligned}$$

The particular convex function

$$\Psi(x) = \tfrac{1}{5}\exp(x^2)$$

would give tails decreasing like $\exp(-Ct^2)$ for some constant C. Such a rate of decrease will be referred to as subgaussian tail behavior.

The inequality in Section 3 will be for processes indexed by a subset $\mathcal{F}$ of $\mathbb{R}^n$. It will take the form of a bound on the particular Orlicz norm,

$$\Big\| \sup_{\mathbf{f} \in \mathcal{F}} |\boldsymbol{\sigma} \cdot \mathbf{f}| \Big\|_\Psi,$$

involving the *packing numbers* for the set $\mathcal{F}$. [The packing number $D(\epsilon, \mathcal{F})$ is the largest number of points that can be packed into $\mathcal{F}$ with each pair at least ϵ apart.] In this way we transform the study of maximal inequalities for Δ_n into a study of the geometry of the set $\mathcal{F}_\omega$.

Section 4 will make the connection between packing numbers and the combinatorial methods that have evolved from the approach of Vapnik and Červonenkis. It will develop the idea that a bounded set $\mathcal{F}$ in $\mathbb{R}^n$ that has a weak property shared by V-dimensional subspaces should have packing numbers like those of a bounded subset of $\mathbb{R}^V$. The three sections after that will elaborate upon the idea, with Section 7 summarizing the results in the form of several simple maximal inequalities for Δ_n.

Section 8 will transform the maximal inequalities into simple conditions for uniform analogues of the law of large numbers. Sections 9 and 10 will transform them into uniform analogues of the central limit theorem—functional limit theorems that are descendents of Donsker's Theorem for the empirical distribution function on the real line. The approach there will depend heavily on the method of almost sure representation.

Section 9 will be the only part of these notes where particular care is taken with questions of measurability. Up to that point any measurability difficulties could be handled by an assumption that T is a Borel (or analytic) subset of a compact metric space and that each of the functions $f_i(\omega, t)$ is jointly measurable in its arguments ω and t. Such niceties are left to the reader.

The challenging applications will occupy the last four sections.

The key to the whole approach taken in these notes is an important combinatorial lemma, a refinement of the so-called *Vapnik-Červonenkis Lemma.* It deserves an immediate proof so that the reader might appreciate the simplicity of the foundation upon which all else rests.

In what follows, $\mathcal{S}$ will denote the set of all 2^n possible n-tuples of $+1$'s and -1's. The pointwise minimum of two vectors $\boldsymbol{\sigma}$ and $\boldsymbol{\eta}$ in $\mathcal{S}$ will be denoted by $\boldsymbol{\sigma} \wedge \boldsymbol{\eta}$. The symbol # will denote cardinality of a set. Inequalities between vectors in $\mathcal{S}$ should be interpreted coordinatewise.

(1.1) Basic Combinatorial Lemma. *For each map η from $\mathcal{S}$ into itself there exists a one-to-one map θ from $\mathcal{S}$ onto itself such that $\theta(\boldsymbol{\sigma}) \wedge \boldsymbol{\sigma} = \eta(\boldsymbol{\sigma}) \wedge \boldsymbol{\sigma}$ for every $\boldsymbol{\sigma}$.*

Proof. Replacing $\eta(\boldsymbol{\sigma})$ by $\eta(\boldsymbol{\sigma}) \wedge \boldsymbol{\sigma}$ if necessary, we may simplify the notation by assuming that $\eta(\boldsymbol{\sigma}) \leq \boldsymbol{\sigma}$ for every $\boldsymbol{\sigma}$. Then for each $\boldsymbol{\sigma}$ in $\mathcal{S}$ we need to choose $\theta(\boldsymbol{\sigma})$ from the set $K(\boldsymbol{\sigma}) = \{\boldsymbol{\alpha} \in \mathcal{S} : \boldsymbol{\alpha} \wedge \boldsymbol{\sigma} = \eta(\boldsymbol{\sigma})\}$. For each subset $\mathcal{A}$ of $\mathcal{S}$ define

$$K(\mathcal{A}) = \bigcup_{\boldsymbol{\sigma} \in \mathcal{A}} K(\boldsymbol{\sigma}).$$

The idea is to prove that $\#K(\mathcal{A}) \geq \#\mathcal{A}$, for every choice of $\mathcal{A}$. The combinatorial result sometimes known as the Marriage Lemma (Dudley 1989, Section 11.6) will then imply existence of a one-to-one map θ from $\mathcal{S}$ onto itself such that $\theta(\boldsymbol{\sigma}) \in K(\boldsymbol{\sigma})$ for every $\boldsymbol{\sigma}$, as required.

For the special case where $\eta(\boldsymbol{\sigma}) = \boldsymbol{\sigma}$ for every $\boldsymbol{\sigma}$, the inequality holds trivially, because then $\boldsymbol{\sigma} \in K(\boldsymbol{\sigma})$ for every $\boldsymbol{\sigma}$, and $K(\mathcal{A}) \supseteq \mathcal{A}$ for every $\mathcal{A}$. The general case will be reduced to the trivial case by a sequence of n modifications that transform a general η to this special η.

The first modification changes the first coordinate of each $\eta(\boldsymbol{\sigma})$. Define a new map η^* by putting $\eta^*(\boldsymbol{\sigma})_i = \eta(\boldsymbol{\sigma})_i$ for $2 \leq i \leq n$, and $\eta^*(\boldsymbol{\sigma})_1 = \sigma_1$. Let $K^*(\boldsymbol{\sigma})$ be the subset of $\mathcal{S}$ defined using η^*. We need to show that

$$\#K(\mathcal{A}) \geq \#K^*(\mathcal{A}).$$

To do this, partition $\mathcal{S}$ into 2^{n-1} sets of pairs $\{\boldsymbol{\beta}^-, \boldsymbol{\beta}^+\}$, where each $\boldsymbol{\beta}^-$ differs from its $\boldsymbol{\beta}^+$ only in the first coordinate, with $\beta_1^- = -1$ and $\beta_1^+ = +1$. It is good enough to show that

$$\#[K(\mathcal{A}) \cap \{\boldsymbol{\beta}^-, \boldsymbol{\beta}^+\}] \geq \#[K^*(\mathcal{A}) \cap \{\boldsymbol{\beta}^-, \boldsymbol{\beta}^+\}]$$

for every such pair. This will follow from: (i) if $\boldsymbol{\beta}^- \in K^*(\mathcal{A})$ then $K(\mathcal{A})$ contains both $\boldsymbol{\beta}^-$ and $\boldsymbol{\beta}^+$; and (ii) if $\boldsymbol{\beta}^+ \in K^*(\mathcal{A})$ but $\boldsymbol{\beta}^- \notin K^*(\mathcal{A})$ then at least one of $\boldsymbol{\beta}^-$ and $\boldsymbol{\beta}^+$ must belong to $K(\mathcal{A})$.

Let us establish (i). Suppose $\boldsymbol{\beta}^- \in K^*(\mathcal{A})$. Then, for some $\boldsymbol{\sigma}$ in $\mathcal{A}$, we have $\boldsymbol{\beta}^- \in K^*(\boldsymbol{\sigma})$, that is $\boldsymbol{\beta}^- \wedge \boldsymbol{\sigma} = \eta^*(\boldsymbol{\sigma})$. For this $\boldsymbol{\sigma}$ we must have

$$-1 = \min[-1, \sigma_1] = \eta^*(\boldsymbol{\sigma})_1 = \sigma_1.$$

Since $\eta(\boldsymbol{\sigma}) \leq \boldsymbol{\sigma}$, it follows that $\eta(\boldsymbol{\sigma})_1 = -1$ and $\eta(\boldsymbol{\sigma}) = \eta^*(\boldsymbol{\sigma})$. Thus $\boldsymbol{\beta}^+ \wedge \boldsymbol{\sigma} = \boldsymbol{\beta}^- \wedge \boldsymbol{\sigma} = \eta(\boldsymbol{\sigma})$, as required for (i).

For (ii), suppose $\boldsymbol{\beta}^+$ belongs to $K^*(\mathcal{A})$ but $\boldsymbol{\beta}^-$ does not. Then, for some $\boldsymbol{\sigma}$ in $\mathcal{A}$, we have $\boldsymbol{\beta}^+ \wedge \boldsymbol{\sigma} = \eta^*(\boldsymbol{\sigma}) \neq \boldsymbol{\beta}^- \wedge \boldsymbol{\sigma}$. Both vectors $\boldsymbol{\beta}^+ \wedge \boldsymbol{\sigma}$ and $\boldsymbol{\beta}^- \wedge \boldsymbol{\sigma}$ agree with $\eta^*(\boldsymbol{\sigma})$, and hence with $\eta(\boldsymbol{\sigma})$, in coordinates 2 to n. Either $\eta(\boldsymbol{\sigma})_1 = -1 = (\boldsymbol{\beta}^- \wedge \boldsymbol{\sigma})_1$ or $\eta(\boldsymbol{\sigma})_1 = \sigma_1 = +1 = (\boldsymbol{\beta}^+ \wedge \boldsymbol{\sigma})_1$. Thus either $\eta(\boldsymbol{\sigma}) = \boldsymbol{\beta}^+ \wedge \boldsymbol{\sigma}$ or $\eta(\boldsymbol{\sigma}) = \boldsymbol{\beta}^- \wedge \boldsymbol{\sigma}$, as required for (ii).

We have now shown that the modification in the first coordinate of the η map reduces the cardinality of the corresponding $K(\mathcal{A})$. A similar modification of η^* in the second coordinate will give a similar reduction in cardinality. After n such modifications we will have changed η so that $\eta(\boldsymbol{\sigma}) = \boldsymbol{\sigma}$ for all $\boldsymbol{\sigma}$. The corresponding $K(\mathcal{A})$ has cardinality bigger than the cardinality of $\mathcal{A}$, because it contains $\mathcal{A}$, but smaller than the cardinality of the $K(\mathcal{A})$ for the original η. □

REMARKS. Several authors have realized the advantages of recasting abstract empirical processes as sums of independent stochastic processes. For example, Alexander (1987b) has developed general central limit theorems that apply to both empirical processes and partial-sum processes; Gaenssler and Schlumprecht (1988) have established moment inequalities similar to one of the inequalities that will appear in Section 7.

The proof of the Basic Combinatorial Lemma is based on Lemmas 2 and 3 of Ledoux and Talagrand (1989). It is very similar to the method used by Steele (1975) to prove the Vapnik-Červonenkis Lemma (see Theorem II.16 of Pollard 1984).