PROBABILITY AND STATISTICS DAY @MIT
in honor of
RICHARD DUDLEY

4 OCTOBER 2003

Some of my favorite Dudley papers

David Pollard
Yale University
http://www.stat.yale.edu/~pollard/

Too much to choose from

- CLTs for empirical measures (Ann Prob 1978)
- St. Flour notes (1984)
- Aarhus notes (1976, 1999)
- RAP (1989)
- UCLT (1999)
- Weak convergence of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces. *Illinois J. Math.* (1966)
- Convergence of Baire measures. *Studia Math.* (1966)
- Distances of probability measures and random variables. *Ann. Math. Statist.* (1968)
- An extended Wichura theorem, definitions of Donsker class, and weighted empirical distributions. *Prob. in Banach Spaces* (1985)
- …

Some characteristics

- weak convergence/convergence in distribution
- gaussian processes $\leftrightarrow$ empirical processes
- looks for "the best and shortest available proofs"
- serious respect for history
- serious concern for measurability difficulties

Without using (7.4)–(7.7), but directly from the definition (7.3) and the recurrence relation

(7.8) $${}_{N}C_{\leqslant k} = {}_{N-1}C_{\leqslant k} + {}_{N-1}C_{\leqslant k-1},$$

Vapnik and Červonenkis ((1971), Lemma 1) prove:

(7.9). THEOREM (Vapnik-Červonenkis). *If X is any set, $\mathcal{C}$ any collection of subsets of X, and $V(\mathcal{C}) \leqslant v$, then $m^{\mathcal{C}}(n) <{}_{N}C_{\leqslant v}$ for all $n \geqslant v$.*

They note that ${}_{n}C_{\leqslant k} \leqslant n^{k} + 1$. (Their 1974 book, pages 214–219, shows that $m^{\mathcal{C}}(n) \leqslant {}_{n}C_{\leqslant V(\mathcal{C})-1}$. Note: in the 1971 paper and the 1974 book, pages 97 and 214, are three disagreeing definitions of "$\Phi(k, n)$.") They prove that for $n > k \geqslant 1$, ${}_{n}C_{\leqslant k} \leqslant 1.5n^{k}/k!$. Hence

(7.10) $$\text{for} \quad n > v := V(\mathcal{C}) \geqslant 1,$$ $$m^{\mathcal{C}}(n) \leqslant 1.5n^{v-1}/(v-1)! < n^{v}.$$

For $n < v$, $m^{\mathcal{C}}(n) = 2^{n} \leqslant 2^{v} \leqslant n^{v}$. If $v = 0$, $\mathcal{C}$ is empty. Thus (without using (7.10)) we have:

(7.11) For any collection $\mathcal{C}$ of sets, $m^{\mathcal{C}}(n) \leqslant n^{V(\mathcal{C})}$ for all $n \geqslant 2$, and $m^{\mathcal{C}}(n) \leqslant n^{V(\mathcal{C})} + 1$ for all $n \geqslant 0$.

Now for any sets $A_1, \cdots, A_m$, let $\mathcal{A}(A_1, \cdots, A_m)$ denote the algebra of subsets of X generated by $A_1, \cdots, A_m$.

(7.12). PROPOSITION. *For any* VCC $\mathcal{C}$ *and any* $k < +\infty$,

$$\mathcal{A}_k(\mathcal{C}) := \bigcup \{\mathcal{A}(A_1, \cdots, A_k) : A_1, \cdots, A_k \in \mathcal{C}\} \quad \textit{is a } \text{VCC}.$$

PROOF. By induction, we may assume $k = 2$. Let $\mathcal{D} := \{A \cap B : A, B \in \mathcal{C}\}$. Then $m^{\mathcal{D}}(n) \leqslant m^{\mathcal{C}}(n)^2 \leqslant (n^{V(\mathcal{C})} + 1)^2 < 2^n$ for n large, so $\mathcal{D}$ is a VCC.

We may assume $\phi \in \mathcal{C}$ and $X \in \mathcal{C}$. If $\mathcal{S} := \{A \setminus B : A, B \in \mathcal{C}\}$ then $\mathcal{S}$ is a VCC as above. A finite union of VCC's is likewise a VCC. Now every set in $\mathcal{A}(A, B)$ is a union of some of the four atoms $A \cap B$, $A \setminus B$, $B \setminus A$, and $(X \setminus A) \setminus B$. Unions of at most four sets can be treated also as above, completing the proof.

(7.13). LEMMA. *If $(X, \mathcal{A}, P)$ is a probability space, $\mathcal{C} \subset \mathcal{A}$, $\mathcal{C}$ is a* VCC *and $v := V(\mathcal{C})$, there is a constant $K = K(v)$ (not depending on P) such that for $0 < \varepsilon \leqslant \frac{1}{2}$,*

$$N(\varepsilon, \mathcal{C}, P) \leqslant K\varepsilon^{-v}|\ln \varepsilon|^{v}.$$

PROOF. Suppose $A_1, \cdots, A_m \in \mathcal{C}$, and $P(A_i \,\Delta A_j) \geqslant \varepsilon$ for $i \neq j$. We may assume $m \geqslant 2$. If $n \geqslant 2$ is so large that $m(m-1)(1-\varepsilon)^n < 2$, then $\Pr\{P_n(A_i \,\Delta A_j) > 0$ for all $i \neq j\} > 0$. In that case, $m \leqslant m^{\mathcal{C}}(n) \leqslant n^{v}$ by (7.11). If we take the smallest n for which $m^2(1-\varepsilon)^n < 2$, then $m^2(1-\varepsilon)^{n-1} \geqslant 2$ so $n - 1 \leqslant (2 \ln m - \ln 2)/|\ln(1-\varepsilon)|$, $n \leqslant (2 \ln m)/\varepsilon$, and $m \leqslant (2 \ln m)^{v}\varepsilon^{-v}$.

The Annals of Mathematical Statistics
1968, Vol. 39, No. 5, 1563–1572

DISTANCES OF PROBABILITY MEASURES AND RANDOM VARIABLES

BY R. M. DUDLEY[1]

Massachusetts Institute of Technology

1. Introduction. Let (S, d) be a separable metric space. Let $\mathcal{P}(S)$ be the set of Borel probability measures on S. $\mathcal{C}(S)$ denotes the Banach space of bounded continuous real-valued functions on S, with norm

$$\|f\|_\infty = \sup \{|f(x)| : x \in S\}.$$

On $\mathcal{P}(S)$ we put the usual weak-star topology TW^*, the weakest such that

$$P \to \int f \, dP, \quad P \in \mathcal{P}(S)$$

is continuous for each $f \in \mathcal{C}(S)$.

It is known ([8], [11], [1]) that TW^* on $\mathcal{P}(S)$ is metrizable. The main purpose of this paper is to discuss and compare various metrics and uniformities on $\mathcal{P}(S)$ which yield the topology TW^*.

For S complete, V. Strassen [10] proved the striking and important result that if $\mu, \nu \in \mathcal{P}(S)$, the Prokhorov distance $\rho(\mu, \nu)$ is exactly the minimum distance "in probability" between random variables distributed according to μ and ν. Theorems 1 and 2 of this paper extend Strassen's result to the case where S is measurable in its completion, and, with "minimum" replaced by "infimum", to an arbitrary separable metric space S. We use the finite combinatorial "marriage lemma" at the crucial step in the proof rather than the separation of convex sets (Hahn-Banach theorem) as in [10]. This offers the possibility of a constructive method of finding random variables as close as possible with the given distributions.

For S complete, V. Skorokhod ([9], Theorem 3.1.1, p. 281) proved the related result that if $\mu_n \to \mu_0$ for TW^* there exist random variables X_n with distributions μ_n such that $X_n \to X_0$ almost surely. This is proved in Section 3 below for a general separable S. Note that it is not sufficient to establish consistent finite-dimensional joint distributions for the X_n ; the Kolmogorov existence theorem for stochastic processes is not available in this generality. Instead we construct the joint distribution of $\{X_n\}_{n=0}^{\infty}$ out of suitable infinite Cartesian product measures.

When S is the real line R, various special constructions involving distribution and characteristic functions are known. In Section 4, we compare some of these uniformities on $\mathcal{P}(R)$.

2. Strassen's theorem. The metric of Prokhorov [8] is defined as follows. For any $x \in S$ and $T \subset S$ let

$$d(x, T) = \inf (d(x, y) : y \in T),$$

Received 11 January 1968.

[1] Fellow of the A. P. Sloan Foundation.

Convergence in "distribution"

- $\{X_n : n \in \mathbb{N}\}$ measurable(?) maps into metric space $(\mathcal{X}, d)$
- P a probability measure on Borel sigma-field of $\mathcal{X}$
- define $X_n \rightsquigarrow P$ to mean

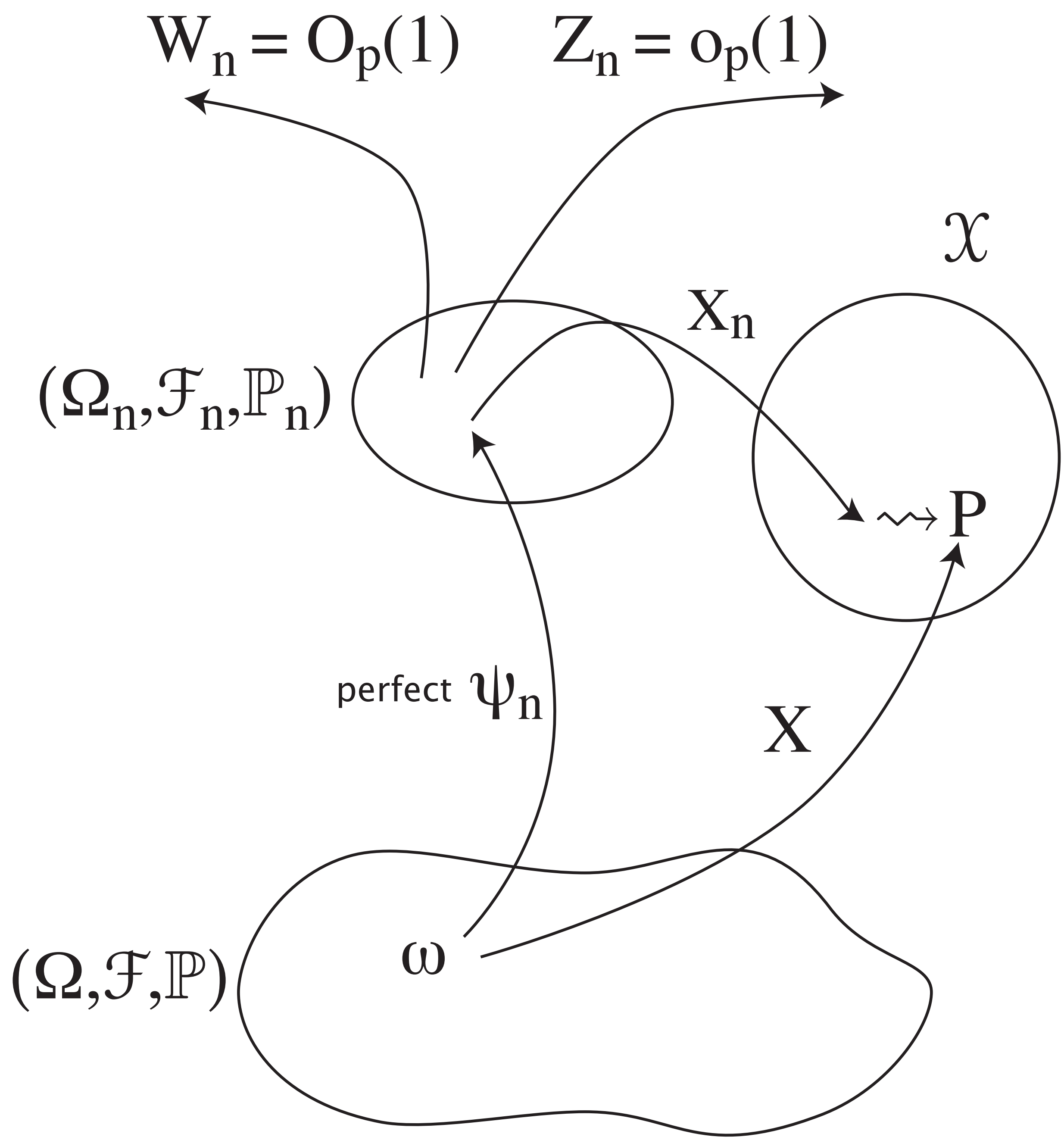
$$\int^* f(X_n)\, d\mathbb{P} \to \int f\, dP$$

$$\int_* f(X_n)\, d\mathbb{P} \to \int f\, dP$$

  for all bounded, (Lipschitz)-continuous, real functions f on $\mathcal{X}$
- Problem: Construct new measurable(?) maps $\{\widetilde{X}_n : n \in \mathbb{N}\}$ with $\widetilde{X}_n$ having "same distribution" as X_n, and $\widetilde{X}$ with distribution P, such that

$$\widetilde{X}_n \to \widetilde{X} \qquad \text{almost surely}$$

  (Or: almost uniformly?)
- Dudley (1985): Build new probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}_n =$ image of $\mathbb{P}$ under a perfect map $\psi_n : \Omega \to \Omega_n$.



Can we make $Z_n \circ \psi_n = o(1)$ almost surely?

Can we make $W_n \circ \psi_n = O(1)$ almost surely?

- No problem with $o_p(1)$:

$$(X_n, Z_n) \rightsquigarrow P \otimes \delta_0$$

- Problem with $O_p(1)$, but $W_n \circ \psi_n$ is $O_p(1)$ under $\mathbb{P}$.

Example

- Suppose:

  (i) stochastic processes $\{X_n(\theta) : \theta \in \mathbb{R}\}$

  (ii) estimators $\widehat{\theta}_n = \operatorname{argmax}_{\theta \in \mathbb{R}} X_n(\theta)$

  (iii) $X_n \rightsquigarrow X$ in the sense of metric for uniform convergence on compacta

  (iv) $\widehat{\theta}_n = O_p(1)$

- Can we deduce that

$$\widehat{\theta}_n \rightsquigarrow \operatorname*{argmax}_{\theta} X(\theta) \qquad ?$$

Comparison of experiments

- Probability measure $\mathbb{P}, \mathbb{P}_1, \ldots, \mathbb{P}_k$ on $(\mathcal{X}, \mathcal{B})$ with $d\mathbb{P}_i/d\mathbb{P} = X_i$
  Probability measure $\mathbb{Q}, \mathbb{Q}_1, \ldots, \mathbb{Q}_k$ on $(\mathcal{Y}, \mathcal{C})$ with $d\mathbb{Q}_i/d\mathbb{Q} = Y_i$

- Distribution of $X := (X_1, \ldots, X_k)$ under $\mathbb{P}$ close to distribution of $Y := (Y_1, \ldots, Y_k)$ under $\mathbb{Q}$.
- Find a randomization (Markov kernel) $K_x(dy)$ such that, for all i and all $|g| \leq 1$,

$$|\int g(y)\mathbb{Q}_i(dy) - \int g(y)K_x(dy)\mathbb{P}_i(dx)| < \text{tiny}$$

- WLOG(???) $\mathcal{X} = \mathcal{Y} = \mathbb{R}^k$ with the X_i and Y_i as coordinate maps (work with image of $\mathbb{P}$ under X and image of $\mathbb{Q}$ under Y)

- $$\Delta := \sup_{\|\ell\|_{Lip} \le 1} | \int \ell(x)\mathbb{P}(dx) - \int \ell(y)\mathbb{Q}(dy)|$$
- Construct probability $\mathbb{K}$ on $\mathbb{R}^k \times \mathbb{R}^k$ with marginals $\mathbb{P}$ and $\mathbb{Q}$ and
$$\iint |x - y|\mathbb{K}(dx, dy) = \Delta$$
- Take K_x as conditional distribution (under $\mathbb{K}$) of y given x
- Then, for $|g| \le 1$,

$$\begin{aligned}
| \int & g(y)\mathbb{Q}_i(dy) - \int g(y)K_x(dy)\mathbb{P}_i(dx)| \\
&= | \int g(y)y_i\mathbb{Q}(dy) - \int g(y)K_x(dy)x_i\mathbb{P}(dx)| \\
&\le \iint |g(y)(y_i - x_i)|\mathbb{K}(dx, dy) \\
&\le \Delta
\end{aligned}$$