Chapter 10

Representations and couplings

SECTION 1 illustrates the usefulness of coupling, by means of three simple examples.

SECTION 2 describes how sequences of random elements of separable metric spaces that converge in distribution can be represented by sequences that converge almost surely.

*SECTION *3 establishes Strassen's Theorem, which translates the Prohorov distance between two probability measures into a coupling.*

*SECTION *4 establishes Yurinskii's coupling for sums of independent random vectors to normally distributed random vectors.*

SECTION 5 describes a deceptively simple example (Tusnády's Lemma) of a quantile coupling, between a symmetric Binomial distribution and its corresponding normal approximation.

SECTION 6 uses the Tusnády Lemma to couple the Haar coefficients for the expansions of an empirical process and a generalized Brownian Bridge.

SECTION 7 derives one of most striking results of modern probability theory, the KMT coupling of the uniform empirial process with the Brownian Bridge process.

1. What is coupling?

A coupling of two probability measures, P and Q, consists of a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting two random elements X and Y, such that X has distribution P and Y has distribution Q. Sometimes interesting relationships between P and Q can be coded in some simple way into the joint distribution for X and Y. Three examples should make the concept clearer.

<1> **Example.** Let P_α denote the $\text{Bin}(n, \alpha)$ distribution. As α gets larger, the distribution should "concentrate on bigger values." More precisely, for each fixed x, the tail probability $P_\alpha[x, n]$ should be an increasing function of α. A coupling argument will give an easy proof.

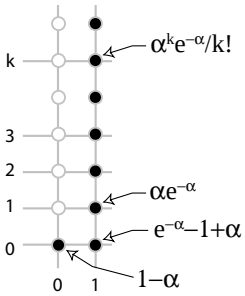
Consider a β larger than α. Suppose we construct a pair of random variables, X_α with distribution P_α and X_β with distribution P_β, such that $X_\alpha \le X_\beta$ almost surely. Then we will have $\{X_\alpha \ge x\} \le \{X_\beta \ge x\}$ almost surely, from which we would recover the desired inequality, $P_\alpha[x, n] \le P_\beta[x, n]$, by taking expectations with respect to $\mathbb{P}$.

How might we construct the coupling? Binomials count successes in independent trials. Couple the trials and we couple the counts. Build the trials from

independent random variables U_i, each uniformly distributed on $(0,1)$. That is, define $X_\alpha := \sum_{i\le n}\{U_i \le \alpha\}$ and $X_\beta := \sum_{i\le n}\{U_i \le \beta\}$. In fact, the construction couples all P_γ, for $0 \le \gamma \le 1$, simultaneously. □

<2> **Example.** Let P denote the Bin(n,α) distribution and Q denote the approximating Poisson$(n\alpha)$ distribution. A coupling argument will establish a total variation bound, $\sup_A |PA - QA| \le n\alpha^2$, an elegant means for expressing the Poisson approximation to the Binomial.

Start with the simplest case, where n equals 1. Find a probability measure $\mathbb{P}$ concentrated on $\{0,1\}\times\mathbb{N}_0$ with marginal distributions $P := \text{Bin}(1,\alpha)$ and $Q := \text{Poisson}(\alpha)$. The strategy is simple: put as much mass as we can on the diagonal, $(0,0)\cup(1,1)$, then spread the remaining mass as needed to get the desired marginals. The atoms on the diagonal are constrained by the inequalities



$$\mathbb{P}(0,0) \le \min\left(P\{0\}, Q\{0\}\right)) = \min\left(1-\alpha, e^{-\alpha}\right),$$
$$\mathbb{P}(1,1) \le \min\left(P\{1\}, Q\{1\}\right)) = \min\left(\alpha, \alpha e^{-\alpha}\right).$$

To maximize, choose $\mathbb{P}(0,0) := 1-\alpha$ and $\mathbb{P}(1,1) := \alpha e^{-\alpha}$. The rest is arithmetic. We need $\mathbb{P}(1,0) := e^{-\alpha} - 1 + \alpha$ to attain the marginal probability $Q\{0\}$, and $\mathbb{P}(0,k) := 0$, for $k = 1,2,\ldots$, to attain the marginal $P\{0\} = 1-\alpha$. The choices $\mathbb{P}(1,k) := Q\{k\}$, for $k = 2,3,\ldots$, are then forced. The total off-diagonal mass equals $\alpha - \alpha e^{-\alpha} \le \alpha^2$.

For the general case, take $\mathbb{P}$ to be the n-fold product of measures of the type constructed for $n = 1$. That is, construct n independent random vectors $(X_1,Y_1),\ldots,(X_n,Y_n)$ with each X_i distributed Bin$(1,\alpha)$, each Y_i distributed Poisson(α), and $\mathbb{P}\{X_i \ne Y_i\} \le \alpha^2$. The sums $X := \sum_i X_i$ and $Y := \sum_i Y_i$ then have the desired Binomial and Poisson distributions, and $\mathbb{P}\{X \ne Y\} \le \sum_i \mathbb{P}\{X_i \ne Y_i\} \le n\alpha^2$. The total variation bound follows from the inequality

$$|\mathbb{P}\{X\in A\} - \mathbb{P}\{Y\in A\}| = |\mathbb{P}\{X\in A, X\ne Y\} - \mathbb{P}\{Y\in A, X\ne Y\}| \le \mathbb{P}\{X\ne Y\},$$

for every subset A of integers. □

The first Example is an instance of a general method for coupling probability measures on the real line by means of quantile functions. Suppose P has distribution function F and Q has distribution function G, with corresponding quantile functions q_F and q_G. Remember from Section 2.9 that, for each $0 < u < 1$,

$$u \le F(x) \qquad \text{if and only if} \qquad q_F(u) \le x.$$

In particular, if U is uniformly distributed on $(0,1)$ then

$$\mathbb{P}\{q_F(U) \le x\} = \mathbb{P}\{U \le F(x)\} = F(x),$$

so that $X := q_F(U)$ must have distribution P. We couple P with Q by using the same U to define the random variable $Y := q_G(U)$ with distribution Q.

A slight variation on the quantile coupling is available when G is one-to-one with range covering the whole of $(0,1)$. In that case, q_G is a true inverse function for G, and $U = G(Y)$. The random variable $X := q_F G(Y)$ is then an increasing function of Y, a useful property. Section 5 will describe a spectacularly successful example of a quantile coupling expressed in this form.

<3> **Example.** Suppose $\{P_n\}$ is a sequence of probability measures on the real line, for which $P_n \rightsquigarrow P$. Write F_n and F for the corresponding distribution functions, and q_n and q for the quantile functions. From Section 7.1 we know that $F_n(x) \to F(x)$ at each x for which $P\{x\} = 0$, which implies (Problem [1]) that $q_n(u) \to q(u)$ at Lebesgue almost all u in $(0, 1)$. If we use a single U, distributed uniformly on $(0, 1)$, to construct the variables $X_n := q_n(U)$ and $X := q(U)$, then we have $X_n \to X$ almost surely. That is we have represented the weakly convergent sequence of measures by an almost surely convergent sequence of random variables.

REMARK. It might happen that the measures $\{P_n\}$ are the distributions of some other sequence of random variables, $\{Y_n\}$. Then, necessarily, $Y_n \rightsquigarrow P$; but the construction does *not* assert that Y_n converges almost surely. Indeed, we might even have the Y_n defined on different probability spaces, which would completely rule out any possible thought of almost sure convergence. The construction ensures that each X_n has marginal distribution P_n, the same as Y_n, but the joint distribution of the X_n's has nothing to do with the joint distribution of the Y_n's (which is only well defined if the Y_n all live on the same probability space). Indeed, that is the whole point of the construction: we have artificially manufactured the joint distribution for the X_n's in order that they converge, not just in the distributional sense, but also in the almost sure sense.

The representation lets us prove facts about weak convergence by means of the tools for almost sure convergence. For example, in the problems to Chapter 7, you were asked to show that $\Delta(P, Q) := \sup\{|P\ell - Q\ell| : \|\ell\|_{BL} \le 1\}$ defines a metric for weak convergence on the set of all Borel probability measures on a separable metric space. (Refer to Section 7.1 for the definition of the bounded Lipschitz norm.) If $\Delta(P_n, P) \to 0$ then $P_n f \to Pf$ for each f with $\|f\|_{BL} < \infty$, that is, $P_n \rightsquigarrow P$. Conversely, if $P_n \rightsquigarrow P$ and can we find X_n with distribution P_n and X with distribution P for which $X_n \to X$ almost surely (see Section 2 for the general case), then

$$\Delta(P_n, P) \le \sup_{\|\ell\|_{BL}\le 1} \mathbb{P}|\ell(X_n) - \ell(X)| \le \mathbb{P}\left(1 \wedge |X_n - X|\right) \to 0.$$

In effect, the general constructions of the representing variables subsume the specific calculations used in Chapter 7 to approximate $\{\ell : \|\ell\|_{BL} \le 1\}$ by a finite collection of functions. □

2. Almost sure representations

The representation from Example <3> has extensions to more general spaces. The result for separable metric spaces gives the flavor of the result without getting us caught up in too many measure theoretic details.

<4> **Theorem.** *For probability measures on the Borel sigma field of a separable metric space $\mathcal{X}$, if $P_n \rightsquigarrow P$ then there exist random elements X_n, with distributions P_n, and X, with distribution P, for which $X_n \to X$ almost surely.*

The main step in the proof involves construction of a joint distribution for X_n and X. To avoid a profusion of subscripts, it is best to isolate this part of the

construction into a separate lemma. Once again, a single uniformly distributed U (that is, with distribution equal to Lebesgue measure $\mathfrak{m}$ on $\mathcal{B}(0,1)$) will eventually provide the thread that ties together the various couplings into a single sequence converging almost surely. The construction builds the joint distribution via a probability kernel, K from $(0,1) \times \mathcal{X}$ into $\mathcal{X}$.

Recall, from Section 4.3, that such a kernel consists of a family of probability measures $\{K_{u,x}(\cdot) : u \in (0,1),\ x \in \mathcal{X}\}$ with $(u,x) \mapsto K_{u,x}B$ measurable for each fixed B in $\mathcal{B}(\mathcal{X})$. We define a measure on the product sigma-field of $(0,1) \times \mathcal{X} \times \mathcal{X}$ by

$$(\mathfrak{m} \otimes P \otimes K)^{u,x,y} f(u,x,y) := \mathfrak{m}^u \left(P^x K_{u,x}^y f(u,x,y) \right).$$

Less formally: we independently generate an observation u from the uniform distribution $\mathfrak{m}$ and an observation x from P, then we generate a y from the corresponding $K_{u,x}$. The expression in parentheses on the right-hand side also defines a probability distribution, $(P \otimes K)_u$, on $\mathcal{X} \times \mathcal{X}$,

$$(P \otimes K)_u^{x,y} f(x,y) := P^x K_{u,x}^y f(x,y) \qquad \text{for each fixed } u.$$

In fact, $\{(P \otimes K)_u : u \in (0,1)\}$ is a probability kernel from $(0,1)$ to $\mathcal{X} \times \mathcal{X}$. Notice also that the marginal distribution $\mathfrak{m}^u P^x K_{u,x}$ for y is a $\mathfrak{m} \otimes P$ average of the $K_{u,x}$ probability measures on $\mathcal{B}(\mathcal{X})$. As an exercise in generating class methods, you might check all the measurability properties needed to make these assertions precise.

<5> **Lemma.** *Let P and Q be probability measures on the Borel sigma-field $\mathcal{B}(\mathcal{X})$. Suppose there is a partition of $\mathcal{X}$ into disjoint Borel sets $B_0, B_1, \ldots, B_m$, and a positive constant ϵ, for which $QB_\alpha \geq (1-\epsilon)PB_\alpha$ for each α. Then there exists a probability kernel K from $(0,1) \times \mathcal{X}$ to $\mathcal{X}$ for which $Q = \mathfrak{m}^u P^x K_{u,x}$ and for which $(P \otimes K)_u$ concentrates on $\cup_\alpha (B_\alpha \times B_\alpha)$ whenever $u \leq 1-\epsilon$.*

Proof. Rewrite the assumption as $QB_\alpha = \delta_\alpha + (1-\epsilon)PB_\alpha$, where the nonnegative numbers δ_α must sum to ϵ because $\sum_\alpha QB_\alpha = \sum_\alpha PB_\alpha = 1$. Write $Q(\cdot \mid B_\alpha)$ for the conditional distribution, which can be taken as an arbitrary probability measure on B_α if $QB_\alpha = 0$. Partition the interval $(1-\epsilon, 1)$ into disjoint subintervals J_α with $\mathfrak{m}J_\alpha = \delta_\alpha$. Define

$$K_{u,x}(\cdot) = \sum\nolimits_\alpha \left(\{u \in J_\alpha\} + \{u \leq 1-\epsilon,\ x \in B_\alpha\} \right) Q(\cdot \mid B_\alpha).$$

When $u \leq 1-\epsilon$ the recipe is: generate y from $Q(\cdot \mid B_\alpha)$ when $x \in B_\alpha$, which ensures that x and y then belong to the same B_α. Integrate over u and x to find the marginal probability that y lands in a Borel set A:

$$\mathfrak{m}^u P^x K_{u,x} A = \sum\nolimits_\alpha \left(\delta_\alpha + (1-\epsilon)PB_\alpha \right) Q(A \mid B_\alpha) = \sum\nolimits_\alpha (QB_\alpha) Q(A \mid B_\alpha) = QA,$$

as asserted. □

REMARK. Notice that the kernel K does nothing clever when $u \in J_\alpha$. If we were hoping for a result closer to the quantile coupling of Example <3>, we might instead try to select y from a B_β that is close to x, in some sense. Such refined behavior would require a more detailed knowledge of the partition.

Proof of Theorem <4>. The idea is simple. For each n we will construct an appropriate probability kernel $K_{u,x}^{(n)}$ from $(0,1) \times \mathcal{X}$ to $\mathcal{X}$, via an appeal to the

Lemma, with Q equal to the corresponding P_n and ϵ depending on n. We then independently generate $X_n(\omega)$ from $K_{u,x}^{(n)}$, for each n, with u an observation from $\mathfrak{m}$ independent of an observation $X(\omega) := x$ from P.

The inequality required by the Lemma would follow from convergence in distribution if each B_α were a P-continuity set (that is, if each boundary ∂B_α had zero P measure—see Section 7.1), for then we would have $P_n B_\alpha \to P B_\alpha$ as $n \to \infty$. Problem [4] shows how to construct such a partition $\pi := \{B_0, B_1, \ldots, B_m\}$ for an arbitrarily small $\epsilon > 0$, with two additional properties,

(i) $P B_0 \le \epsilon$

(ii) diameter$(B_\alpha) \le \epsilon$ for each $\alpha \ge 1$.

We shall need a a whole family of such partitions, $\pi_k := \{B_{\alpha,k} : \alpha = 0, 1, \ldots, m_k\}$, corresponding to values $\epsilon_k := 2^{-k}$ for each $k \in \mathbb{N}$.

To each k there exists an n_k for which $P_n B \ge (1 - \epsilon_k) P B$ for all B in π_k, when $n \ge n_k$. With no loss of generality we may assume that $1 < n_1 < n_2 < \ldots$, which ensures that for each n greater than n_1 there exists a unique $k := k(n)$ for which $n_k \le n < n_{k+1}$. Write $K_{u,x}^{(n)}$ for the probability kernel defined by Lemma <5> for $Q := P_n$ with $\epsilon := \epsilon_{k(n)}$, and $\pi_{k(n)}$ as the partition. Define $\mathbb{P}$ as the probability measure $\mathfrak{m} \otimes P \otimes \left(\otimes_{n\in\mathbb{N}} K_{u,x}^{(n)}\right)$ on the product sigma-field of $\Omega := (0,1) \times \mathcal{X} \times \mathcal{X}^{\mathbb{N}}$. The generic point of Ω is a sequence $\omega := (u, x, y_1, y_2, \ldots)$. Define $X(\omega) := x$ and $X_n(\omega) := y_n$.

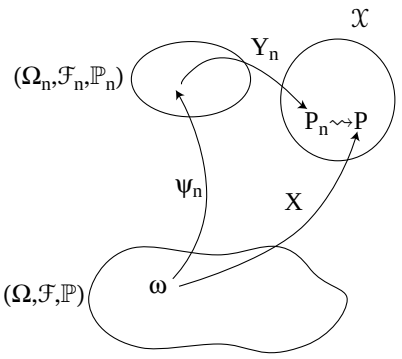
Why does X_n converge $\mathbb{P}$-almost surely to X? First note that $\sum_k P B_{0,k} < \infty$. Borel-Cantelli therefore ensures that, for almost all x and every u in $(0,1)$, there exists a $k_0 = k_0(u,x)$ for which $u \le 1 - \epsilon_k$ and $x \notin B_{0,k}$ for all $k \ge k_0$. For such (u,x) and $k \ge k_0$ we have $(x, y_n) \in \cup_{\alpha\ge1} B_{\alpha,k} \times B_{\alpha,k}$ for $n_k \le n < n_{k+1}$, by the concentration property of the kernels. That is, both $X(\omega)$ and $X_n(\omega)$ fall within the same $B_{\alpha,k}$ with $\alpha \ge 1$, a set with diameter less than ϵ_k. Think your way through that convoluted assertion and you will realize we have shown something even stronger than almost sure convergence. □

<6> **Example.** Suppose $P_n \rightsquigarrow P$ as probability measures on the Borel sigma-field of a separable metric space, and suppose that $\{T_n\}$ is a sequence of measurable maps into another metric space $\mathcal{Y}$. If P-almost all x have the property that $T_n(x_n) \to T(x)$ for every sequence $\{x_n\}$ converging to x, then the sequence of image measures also converges in distribution, $T_n P_n \rightsquigarrow T P$, as probability measures on the Borel sigma-field of $\mathcal{Y}$. The proof is easy is we represent $\{P_n\}$ by the sequence $\{X_n\}$, as in the Theorem. For each ℓ in $BL(\mathcal{Y})$, we have $\ell(T_n(X_n(\omega))) \to \ell(T(X(\omega)))$ for $\mathbb{P}$-almost all ω. Thus

$$(T_n P_n)\ell = \mathbb{P}\ell(T_n(X_n)) \to \mathbb{P}\ell(T(X)) = (T P)\ell,$$

by Dominated Convergence. □

I noted in Example <3> that if Y_n has distribution P_n, and if each Y_n is defined on a different probability space $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$, then the convergence in distribution $Y_n \rightsquigarrow P$ cannot possibly imply almost sure convergence for Y_n. Nevertheless, using an argument similar to the proof of Theorem <4>, Dudley (1985) obtained something almost as good as almost sure convergence.



He built a single probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting measurable maps ψ_n, into Ω_n, and X, into $\mathcal{X}$, with distributions $\mathbb{P}_n = \psi_n(\mathbb{P})$ and $P = X(\mathbb{P})$, for which $Y_n(\psi_n(\omega)) \to X(\omega)$ for $\mathbb{P}$ almost all ω. In effect, the ψ_n maps pull Y_n back to Ω, where the notions of pointwise and almost sure convergence make sense.

Actually, Dudley established a more delicate result, for Y_n that need not be measurable as maps into $\mathcal{X}$, a generalization needed to accommodate an application in the theory of abstract empirical processes. See Pollard (1990, Section 9) for a discussion of some of the conceptual and technical difficulties—such as the meaning of convergence in distribution for maps that don't have distributions in the usual sense—that are resolved by Dudley's construction. See Kim & Pollard (1990, Section 2) for an example of the subtle advantages of Dudley's form of the representation theorem.

*3. Strassen's Theorem

Once again let $(\mathcal{X}, d)$ be a separable metric space equipped with its Borel sigma-field $\mathcal{B}(\mathcal{X})$. For each subset A of $\mathcal{X}$, and each $\epsilon \geq 0$, define A^ϵ to be the closed set $\{x \in \mathcal{X} : d(x, A) \leq \epsilon\}$. The ***Prohorov distance*** between any P and Q from the set $\mathcal{P}$ of all probability measures on $\mathcal{B}(\mathcal{X})$ is defined as

$$\rho(P, Q) := \inf\{\epsilon > 0 : PB \leq QB^\epsilon + \epsilon \text{ for all } B \text{ in } \mathcal{B}(\mathcal{X})\}.$$

Despite the apparent lack of symmetry in the definition, ρ is a metric (Problem [3]) on $\mathcal{P}$.

REMARK. Separability of $\mathcal{X}$ is convenient, but not essential when dealing with the Prohorov metric. For example, it implies that $\mathcal{B}(\mathcal{X} \times \mathcal{X}) = \mathcal{B}(\mathcal{X}) \otimes \mathcal{B}(\mathcal{X})$, which ensures that $d(X, X')$ is measurable for each pair of random elements X and X'; and if $X_n \to X$ almost surely then $\mathbb{P}\{d(X_n, X) > \epsilon\} \to 0$ for each $\epsilon > 0$.

If $\rho(P_n, P) \to 0$ then, for each closed F we have $P_n F \leq PF^\epsilon + \epsilon$ eventually, and hence $\limsup_n P_n F \leq PF$, implying that $P_n \rightsquigarrow P$. Theorem <4> makes it easy to prove the converse. If X_n has distribution P_n and X has distribution P, and if $X_n \to X$ almost surely, then for each $\epsilon > 0$ there is an n_ϵ such that

$$\mathbb{P}\{d(X_n, X) > \epsilon\} < \epsilon \qquad \text{for } n \geq n_\epsilon.$$

For every Borel set B, when $n \geq n_\epsilon$ we have

<7> $$P_n B \leq \mathbb{P}\{X_n \in B,\ d(X_n, X) \leq \epsilon\} + \mathbb{P}\{d(X_n, X) > \epsilon\} \leq \mathbb{P}\{X \in B^\epsilon\} + \epsilon = PB^\epsilon + \epsilon.$$

Thus ρ is actually a metric for weak convergence of probability measures.

The Prohorov metric also has an elegant (and useful, as will be shown by Section 4) coupling interpretation, due to Strassen (1965). I will present a slightly restricted version of the result, by placing a tightness assumption on the probabilities, in order to simplify the statement of the Theorem. (Actually, the proof will establish

a stronger result; the tightness will be used only at the very end, to tidy up.) Also, the role of ϵ is slightly easier to understand if we replace it by two separate constants.

<8> **Theorem.** *Let P and Q be tight probability measures on the Borel sigma field $\mathcal{B}$ of a separable metric space $\mathcal{X}$. Let ϵ and ϵ' be positive constants. There exists random elements X and Y of $\mathcal{X}$ with distributions P and Q such that $\mathbb{P}\{d(X,Y) > \epsilon\} \le \epsilon'$ if and only if $PB \le QB^{\epsilon} + \epsilon'$ for all Borel sets B.*

The argument for deducing the family of inequalities from existence of the coupling is virtually the same as <7>. For the other, more interesting direction, I follow an elegant idea of Dudley (1976, Lecture 18). By approximation arguments he reduced to the case where both P and Q concentrate on a finite set of atoms, and then existence of the coupling followed by an appeal to the classical Marriage Lemma (Problem [5]). I modify his argument to eliminate a few steps, by making an appeal to the following generalization (proved in Problem [6]) of that Lemma.

<9> **Lemma.** *Let ν be a finite measure on a finite set S and μ be a finite measure on a sigma-field $\mathcal{B}$ on a set T. Suppose $\{R_\alpha : \alpha \in S\}$ is a collection of measurable sets with the domination property that $\nu(A) \le \mu(\cup_{\alpha\in A} R_\alpha)$ for all $A \subseteq S$. Then there exists a probability kernel K from S to T with K_α concentrated on R_α for each α and $\sum_{\alpha\in S} \nu\{a\}K_\alpha \le \mu$.*

Proof of Theorem <8>. The measure $\mathbb{P}$ will live on $\mathcal{X}\times\mathcal{X}$, with X and Y as the coordinate maps. It will be the limit of a weakly convergent subsequence of a uniformly tight family $\{\mathbb{P}_\delta : \delta > 0\}$, obtained by an appeal to the Prohorov/Le Cam theorem from Section 7.5.

Construct $\mathbb{P}_\delta$ via a "discretization" of P, which brings the problem within the ambit of Lemma <9>. For a small, positive δ, which will eventually be sent to zero, partition $\mathcal{X}$ into finitely many disjoint, Borel sets $B_0, B_1, \ldots, B_m$ with $PB_0 < \delta$ and diameter$(B_\alpha) < \delta$ for $\alpha \ge 1$. (Compare with the construction in Problem [4].) Define a probability measure ν, concentrated on the finite set $S := \{0, 1, \ldots, m\}$, by $\nu\{\alpha\} := PB_\alpha$ for $\alpha = 0, \ldots, m$. Augment $\mathcal{X}$ by a point ∞. Extend Q to a measure μ on $T := \mathcal{X}\cup\{\infty\}$ by placing mass ϵ' at ∞. Define R_α as $B_\alpha^{\epsilon}\cup\{\infty\}$. With these definitions, the measures ν and μ satisfy the requirements of Lemma <9>: for each subset A of S,

$$\nu(A) = P(\cup_{\alpha\in A} B_\alpha) \le Q(\cup_{\alpha\in A} B_\alpha)^{\epsilon} + \epsilon' = Q\left(\cup_{\alpha\in A} B_\alpha^{\epsilon}\right) + \mu\{\infty\} = \mu(\cup_{\alpha\in A} R_\alpha).$$

The Lemma ensures existence of a probability kernel K, from S to T, with $K_\alpha B_\alpha^{\epsilon} + K_\alpha\{\infty\} = K_\alpha R_\alpha = 1$ for each α and $\sum_\alpha \nu\{\alpha\}K_\alpha A \le \mu A$ for every Borel subset A of T. In particular, $\sum_\alpha \nu\{\alpha\}K_\alpha B \le QB$ for all $B \in \mathcal{B}$. The nonnegative measure $Q - \sum_\alpha \nu\{\alpha\}K_\alpha\big|_{\mathcal{X}}$ on $\mathcal{B}$ has total mass

$$\tau := 1 - \textstyle\sum_\alpha \nu\{\alpha\}K_\alpha\mathcal{X} = \sum_\alpha \nu\{\alpha\}K_\alpha\{\infty\} \le \mu\{\infty\} = \epsilon'.$$

Write this measure as τQ_0, with Q_0 a probability measure on $\mathcal{B}$. (If $\tau = 0$, choose Q_0 arbitrarily.) We then have $Qh = \tau Q_0 h + \sum_\alpha \nu\{\alpha\}K_\alpha h$ for all $h \in \mathcal{M}^+(\mathcal{X})$.

Define a probability measure $\mathbb{P}_\delta$ on $\mathcal{B}\otimes\mathcal{B}$ by

$$\mathbb{P}_\delta f := P^x\left(\textstyle\sum_{\alpha=0}^{m}\{x\in B_\alpha\}(K_\alpha + K_\alpha\{\infty\}Q_0)^y f(x,y)\right) \qquad \text{for } f \in \mathcal{M}^+(\mathcal{X}\times\mathcal{X}).$$

REMARK. In effect, I have converted K to a probability kernel L from $\mathcal{X}$ to $\mathcal{X}$, by setting L_x equal to $K_\alpha|_{\mathcal{X}} + K_\alpha\{\infty\}Q_0$ when $x \in B_\alpha$. The definition of $\mathbb{P}_\delta$ is equivalent to $\mathbb{P}_\delta := P \otimes L$, in the sense of Section 4.4.

The measure $\mathbb{P}_\delta$ has marginals P and Q because, for g and h in $\mathcal{M}^+(\mathcal{X})$,

$$\mathbb{P}_\delta^{x,y} g(x) = P^x \left(\sum\nolimits_\alpha \{x \in B_\alpha\} \left(K_\alpha \mathcal{X} + K_\alpha\{\infty\} \right) g(x) \right) = Pg,$$

$$\mathbb{P}_\delta^{x,y} h(y) = \sum\nolimits_\alpha P\{x \in B_\alpha\} \left(K_\alpha h + K_\alpha\{\infty\} Q_0 h \right) = \sum\nolimits_\alpha \nu\{\alpha\} K_\alpha h + \tau Q_0 h.$$

It concentrates most of its mass on the set $D := \cup_{\alpha=1}^m \left(B_\alpha \times B_\alpha^\epsilon \right)$,

$$\begin{aligned} \mathbb{P}_\delta D &\geq \sum\nolimits_{\alpha=1}^m P^x \left(\{x \in B_\alpha\} K_\alpha^y \{(x,y) \in D\} \right) \\ &= \sum\nolimits_{\alpha=1}^m P^x \left(\{x \in B_\alpha\} K_\alpha^y \{y \in B_\alpha^\epsilon\} \right) \\ &= \sum\nolimits_{\alpha=1}^m \nu\{\alpha\} K_\alpha \mathcal{X} = 1 - \tau - (PB_0)(K_0\mathcal{X}). \end{aligned}$$

When (x, y) belongs to D, we have $x \in B_\alpha$ and $d(y, B_\alpha) \leq \epsilon$ for some B_α with diameter$(B_\alpha) < \delta$, and hence $d(x, y) \leq \delta + \epsilon$. Thus $\mathbb{P}_\delta$ assigns measure at least $1 - \epsilon' - \delta$ to the closed set $F_{\delta+\epsilon} := \{(x, y) \in \mathcal{X} \times \mathcal{X} : d(x, y) \leq \delta + \epsilon\}$.

The tightness of both P and Q will let us eliminate δ, by passing to the limit along a subsequence. For each $\eta > 0$ there exists a compact set C_η for which $PC_\eta^c < \eta$ and $QC_\eta^c < \eta$. The probability measure $\mathbb{P}_\delta$, which has marginals P and Q, puts mass at most 2η outside the compact set $C_\eta \times C_\eta$. The family $\{\mathbb{P}_\delta : \delta > 0\}$ is uniformly tight, in the sense explained in Section 7.5. As shown in that Section, there is a sequence $\{\delta_i\}$ tending to zero for which $\mathbb{P}_{\delta_i} \rightsquigarrow \mathbb{P}$, with $\mathbb{P}$ a probability measure on $\mathcal{B} \otimes \mathcal{B}$. It is a very easy exercise to check that $\mathbb{P}$ has marginals P and Q. For each fixed $t > \epsilon$, the weak convergence implies

$$\mathbb{P}F_t \geq \limsup_i \mathbb{P}_{\delta_i} F_t \geq \limsup_i \mathbb{P}_{\delta_i} F_{\epsilon+\delta_i} \geq 1 - \epsilon'.$$

□ Let t decrease to ϵ to complete the proof.

*4. The Yurinskii coupling

The multivariate central limit theorem gives conditions under which a sum S of independent random vectors $\xi_1, \ldots, \xi_n$ has an approximate normal distribution. Theorem <4> would translate the corresponding distributional convergence into a coupling between the standardized sum and a random vector with the appropriate normal distribution. When the random vectors have finite third moments, Theorem <8> improves the result by giving a rate of convergence (albeit in probability).

<10> **Theorem.** *Let $\xi_1, \ldots, \xi_n$ be independent random k-vectors with $\mathbb{P}\xi_i = 0$ for each i and $\beta := \sum_i \mathbb{P}|\xi_i|^3$ finite. Let $S := \xi_1 + \ldots + \xi_n$. For each $\delta > 0$ there exists a random vector T with a $N(0, \mathrm{var}(S))$ distribution such that*

$$\mathbb{P}\{|S - T| > 3\delta\} \leq C_0 B \left(1 + \frac{|\log(1/B)|}{k} \right) \qquad \text{where } B := \beta k \delta^{-3},$$

for some universal constant C_0.

REMARK. The result stated by Yurinskii (1977) took a slightly different form. I have followed Le Cam (1988, Theorem 1) in reworking the Yurinskii's methods. Both those authors developed bounds on the Prohorov distance, by making an explicit choice for δ. The Le Cam preprint is particularly helpful in its discussion of heuristics behind how one balances the effect of various parameters to get a good bound.

Proof. The existence of the asserted coupling (for a suitably rich probability space) will follow via Theorem <8> if we can show for each Borel subset A of $\mathbb{R}^k$ that

<11> $$\mathbb{P}\{S \in A\} \le \mathbb{P}\{T \in A^{3\delta}\} + \text{ERROR},$$

with the ERROR equal to the upper bound stated in the Theorem. By choosing a smooth (bounded derivatives up to third order) function f that approximates the indicator function of A, in the sense that $f \approx 1$ on A and $f \approx 0$ outside $A^{3\delta}$, we will be able to deduce inequality <11> from the multivariate form of Lindeberg's method (Section 7.3), which gives a third moment bound for a difference in expectations,

<12> $$|\mathbb{P} f(S) - \mathbb{P} f(T)| \le C\left(\mathbb{P}|\xi_1|^3 + \ldots + \mathbb{P}|\xi_k|^3\right) = C\beta.$$

More precisely, if the constant C_f is such that

<13> $$\left| f(x+y) - f(x) - y' \dot{f}(x) - \tfrac{1}{2} y' \ddot{f}(x) y \right| \le C_f |y|^3 \qquad \text{for all } x \text{ and } y,$$

then we may take $C = \left(9 + 8\mathbb{P}|N(0,1)|^3\right) C_f \le 15 C_f$.

For a fixed Borel set A, Lemma <18> at the end of the Section will show how to construct a smooth function f for which approximation <13> holds with $C_f = (\sigma^2\delta)^{-1}$ and for which, if $\delta > \sigma\sqrt{k}$,

<14> $$(1-\epsilon)\{x \in A\} \le f(x) \le \epsilon + (1-\epsilon)\{x \in A^{3\delta}\} \qquad \text{where } \begin{cases} \epsilon := \left(\dfrac{1+\alpha}{e^{\alpha}}\right)^{k/2}, \\ 1+\alpha := \dfrac{\delta^2}{k\sigma^2}. \end{cases}$$

The Lindeberg bound <12>, with $C\beta = 15\beta/(\sigma^2\delta) = 15B(1+\alpha)$, then gives

$$\begin{aligned} \mathbb{P}\{S \in A\} &\le (1-\epsilon)^{-1} \mathbb{P} f(S) \\ &\le (1-\epsilon)^{-1}\left(\mathbb{P} f(T) + 15B(1+\alpha)\right) \\ &\le \mathbb{P}\{T \in A^{3\delta}\} + \epsilon' \qquad \text{where } \epsilon' := \frac{\epsilon + 15B(1+\alpha)}{(1-\epsilon)}. \end{aligned} \tag{<15>}$$

We need to choose α, as a function of k and B, to make ϵ' small.

Clearly the bound <15> is useful only when ϵ is small, in which case the $(1-\epsilon)$ factor in the denominator contributes only an extra contant factor to the final bound. We should concentrate on the numerator. Similarly, the assertion of the Theorem is trivial if B is not small. Provided we make sure $C_0 \ge e$, we may assume $B \le e^{-1}$, that is, $\log(1/B) \ge 1$.

To get within a factor 2 of minimizing a sum of two nonnegative functions, one increasing and the other decreasing, it suffices to equate the two contributions. This fact suggests we choose α to make

$$\alpha - \left(1 - \frac{2}{k}\right)\log(1+\alpha) \approx \frac{2}{k}\log(1/B) + O(k^{-1}).$$

If B is small then α will be large, which would make $\log(1+\alpha)$ small compared with α. If we make α slightly larger than $2k^{-1}\log(1/B)$ we should get close to equality. Actually, we can afford to have α a larger multiple of $\log(1/B)$, because extra multiplicative factors will just be absorbed into constant C_0. With these thoughts, it seems to me I cannot do much better than choose

$$\alpha := 3\left(1+\frac{2}{k}\log(1/B)\right),$$

which at least has the virtue of giving a clean bound:

$$\log\epsilon \le \frac{k}{2}\left(\log(1+\alpha)-\frac{2\alpha}{3}\right)-\frac{k\alpha}{6} \le -\log(1/B) \le -1.$$

and hence

$$\epsilon' = \frac{\epsilon+15B(1+\alpha)}{(1-\epsilon)} \le \frac{90}{1-e^{-1}}B\left(1+\frac{\log(1/B)}{k}\right) \qquad \text{when } B \le e^{-1}.$$

The proof is complete, except for the construction of the smooth function f satisfying <14>. □

Before moving on to the construction of f, let us see what we can do with the coupling from the Theorem in the case of identically distributed random vectors. For convenience of notation write $\psi_k(x)$ for the function $C_0 x\,(1+|\log(1/x)|/k)$.

<16> **Example.** Let $\xi_1, \xi_2, \ldots$ be independent, identically distributed random k-vectors with $\mathbb{P}\xi_1 = 0$, $\operatorname{var}(\xi_1) := V$, and $\mu_3 := \mathbb{P}|\xi_1|^3 < \infty$. Write S_n for $\xi_1 + \ldots \xi_n$. The central limit theorem asserts that $S_n/\sqrt{n} \rightsquigarrow N(0,V)$. Theorem <10>, asserts existence of a sequence of random vectors W_n, each distributed $N(0,V)$ for which

$$\mathbb{P}\left\{\left|\frac{S_n}{\sqrt{n}} - W_n\right| \ge 3\delta\right\} \le \psi_k\left(\frac{kn\mu_3}{(\delta\sqrt{n})^3}\right).$$

For fixed k, we can make the right-hand side as small as we please by choosing δ as a large enough enough multiple of $n^{-1/6}$. Thus, with finite third moments,

$$\left|\frac{S_n}{\sqrt{n}} - W_n\right| = O_p(n^{-1/6}) \qquad \text{via the Yurinskii coupling.}$$

For $k = 1$, this coupling is not the best possible. For example, under an assumption of finite third moments, a theorem of Major (1976) gives a sequence of independent random variables $Y_1, Y_2, \ldots$, each distributed $N(0,V)$, for which

$$\left|\frac{S_n}{\sqrt{n}} - \frac{Y_1+\ldots+Y_n}{\sqrt{n}}\right| = o_p(n^{-1/6}) \qquad \text{almost surely.}$$

Major's result has the correct joint distributions for the approximating normals, as n changes, as well as providing a slightly better rate. □

<17> **Example.** Yurinskii's coupling (and its refinements: see, for example, the discussion near Lemma 2.12 of Dudley & Philipp 1983) is better suited to situations where the dimension k can change with n.

Consider the case of a sequence of independent, identially distributed stochastic processes $\{X_i(t) : t \in T\}$. Suppose $\mathbb{P}X_1(t) = 0$ and $|X_1(t)| \le 1$ for every t. Under suitable regularity conditions on the sample paths, we might try to show that the standardized partial sum processes, $Z_n(t) := (X_1(t)+\ldots+X_n(t))/\sqrt{n}$, behave like a

centered Gaussian process $\{Z(t) : t \in T\}$, with the same covariance structure as X_1. We might even try to couple the processes in such a way that $\sup_t |Z_n(t) - Z(t)|$ is small in some probabilistic sense.

The obvious first step towards establishing a coupling of the processes is to consider behavior on large finite subsets $T(k) := \{t_1, \ldots, t_k\}$ of T, where k is allowed to increase with n. The question becomes: How rapidly can k tend to infinity?

For fixed k, write ξ_i for the random k-vector with components $X_i(t_j)$, for $j = 1, \ldots, k$. We seek to couple $(\xi_1 + \ldots + \xi_n)/\sqrt{n}$ with a random vector W_n, distributed like $\{Z(t_j) : j = 1, \ldots, k\}$. The bound is almost the same as in Example <16>, except for the fact that the third moment now has a dependence on k,

$$\mathbb{P}|\xi_1|^3 = k^{3/2}\mathbb{P}\left(\frac{1}{k}\sum_{j=1}^{k} X_1(t_j)^2\right)^{3/2} \le k^{3/2}\mathbb{P}\left(\frac{1}{k}\sum_{j=1}^{k} |X_1(t_j)|^3\right) \le k^{3/2}.$$

Via the general fact that $\max_j |x_j| \le \left(\sum_j x_j^2\right)^{1/2}$, the coupling bound becomes

$$\begin{aligned}\mathbb{P}\left\{\max_{j\le k}\left|Z_n(t_j) - W_{n,j}\right| \ge 3\delta\right\} &\le \mathbb{P}\left\{\left|\frac{\xi_1 + \ldots + \xi_n}{\sqrt{n}} - W_n\right| \ge 3\delta\right\} \\ &\le \mathcal{Y}_k\left(\frac{nk^{5/2}}{(\delta\sqrt{n})^3}\right) \\ &\to 0 \qquad \text{if } k = o\left(n^{1/5}\right) \text{ and } \delta \to 0 \text{ slowly enough.}\end{aligned}$$

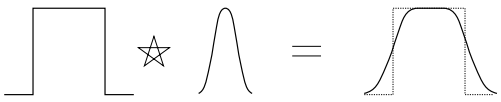
□ That is, $\max_{j\le k}\left|Z_n(t_j) - W_{n,j}\right| = o_p(1)$ if k increases more slowly than $n^{1/5}$.

Smoothing of indicator functions

There are at least two methods for construction of a smooth approximation f to a set A. The first uses only the metric:

$$f(x) = (1 - d(x, A)/\delta)^+ .$$

For an interval in one dimension, the approximation has the effect of replacing the discontinuity at the boundary points by linear functions with slope $1/\delta$. The second method treats the indicator function of the set as an element of an $\mathcal{L}^1$ space, and constructs the approximation by means of convolution smoothing,



$$f(x) = \mathfrak{m}^w\left(\{w \in A\}\phi_\sigma(w - x)\right),$$

where ϕ_σ denotes the $N(), \sigma^2 I_k)$ density and $\mathfrak{m}$ denotes Lebesgue measure on $\mathcal{B}(\mathbb{R}^k)$. (Any smooth density with rapidly decreasing tails would suffice.) A combination of the two methods of smoothing will give the best bound:

<18> **Lemma.** *Let A be a Borel subset of $\mathbb{R}^k$. let Z have a $N(0, I_k)$ distribution. For positive constants δ and σ define*

$$g(x) := \left(1 - \frac{d(x, A^\delta)}{\delta}\right)^+ \qquad \text{and} \qquad f(x) := \mathbb{P}g(x + \sigma Z) = \mathfrak{m}^w \left(g(w)\phi_\sigma(w - x)\right).$$

Then f satisfies <13> with $C := (\sigma^2\delta)^{-1}$, and approximation <14> holds.

Proof. The function f inherits some smoothness from g and some from the convolving standard normal density ϕ_σ, which has derivatives

$$\frac{\partial}{\partial z}\phi_\sigma(z) = -\frac{z}{\sigma^2}\phi_\sigma(z) \qquad \text{and} \qquad \frac{\partial^2}{\partial z^2}\phi_\sigma(z) = \left(\frac{zz'}{\sigma^4} - \frac{I_k}{\sigma^2}\right)\phi_\sigma(z).$$

For fixed x and y, the function $h(t) := f(x + ty)$, for $0 \le t \le 1$, has second derivative

$$\begin{aligned}\ddot{h}(t) &= \mathfrak{m}^w \left(g(w)\left(\frac{(y'(w - x - ty))^2}{\sigma^4} - \frac{|y|^2}{\sigma^2}\right)\phi_\sigma(w - x - ty)\right) \\ &= \sigma^{-2}\mathbb{P}\left(g(x + ty + \sigma Z)\left((y'Z)^2 - |y|^2\right)\right).\end{aligned}$$

The Lipschitz property $|g(x + ty + \sigma Z) - g(x + \sigma Z)| \le t|y|/\delta$ then implies

$$|\ddot{h}(t) - \ddot{h}(0)| \le \frac{t|y|}{\sigma^2\delta}\mathbb{P}\left((y'Z)^2 + |y|^2\right) \le \frac{2|y|^3}{\sigma^2\delta}.$$

The asserted inequality <13> then follows from a Taylor expansion,

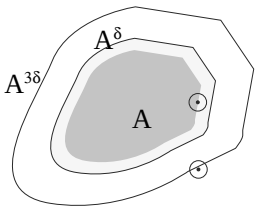
$$|h(1) - h(0) - \dot{h}(0) - \tfrac{1}{2}\ddot{h}(0)| = \tfrac{1}{2}\left|\ddot{h}(t^*) - \ddot{h}(0)\right| \qquad \text{where } t^* \in (0, 1).$$

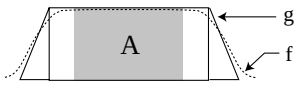
For approximation <14>, first note that $A^\delta \le g \le A^{2\delta}$ and $0 \le f \le 1$ everywhere. Also $\mathbb{P}\{|Z| > \delta/\sigma\} \le \epsilon$, from Problem [7]. Thus

$$f(x) \ge \mathbb{P}g(x + \sigma Z)\{|\sigma Z| \le \delta\} = \mathbb{P}\{|Z| \le \delta/\sigma\} \ge 1 - \epsilon \qquad \text{if } x \in A,$$

and

$$f(x) = \mathbb{P}g(x + \sigma Z)\{|\sigma Z| \le \delta\} + \mathbb{P}g(x + \sigma Z)\{|\sigma Z| > \delta\} \le \epsilon \qquad \text{if } x \notin A^{3\delta}.$$





□

5. Quantile coupling of Binomial with normal

As noted in Section 1, if η is distributed $N(0, 1)$, with distribution function Φ, and if q denotes the Bin$(n, 1/2)$ quantile function, then the random variable $X := q(\Phi(\eta))$ has exactly a Bin$(n, 1/2)$ distribution. In a sense made precise by the following Lemma, X is very close to the random variable $Y := n/2 + \eta\sqrt{n/4}$, which has a $N(n/2, n/4)$ distribution. The coupling of the Bin$(n, 1/2)$ with its

approximating $N(n/2, n/4)$ has been the starting point for a growing collection of striking approximation results, inspired by the publication of the fundamental paper of Komlós, Major & Tusnády (1975).

<19> **Tusnády's Lemma.** *For each positive integer n there exists a deterministic, increasing function $\tau(n,\cdot)$ such that the random variable $X := \tau(n,\eta)$ has a* $\text{Bin}(n, 1/2)$ *distribution whenever η has a $N(0,1)$ distribution. The random variable X satisfies the inequalities*

$$|X - Y| \le 1 + \frac{\eta^2}{8} \qquad \text{and} \qquad \left|X - \frac{n}{2}\right| \le 1 + \frac{\sqrt{n}|\eta|}{2},$$

where $Y := \frac{n}{2} + \eta\sqrt{\frac{n}{4}}$, which has a $N\left(\frac{n}{2}, \frac{n}{4}\right)$ distribution.

At first glance it is easy to underestimate the delicacy of these two inequalities. Both X and Y have mean $n/2$ and standard deviation of order $\sqrt{n}$. It would be no challenge to construct a coupling for which $|X - Y|$ is of order $\sqrt{n}$; the Lemma gives a coupling for which $|X - Y|$ is bounded by a quantity whose distribution does not even change with n.

The original proof (Tusnády 1977) of the Lemma is challenging. Appendix D contains an alternative derivation of similar inequalities. To simplify the argument, I have made no effort to derive the best constants for the bound. In fact, the precise constants appearing in the Lemma will have no importance for us. It will be enough for us to have a universal constant C_0 for which there exists couplings such that

<20> $$|X - Y| \le C_0\left(1 + \eta^2\right) \qquad \text{and} \qquad \left|X - \frac{n}{2}\right| \le C_0\sqrt{n}\,(1 + |\eta|)\,,$$

a weaker bound that follows easily from the inequalities in Appendix D.

6. Haar coupling—the Hungarian construction

Let $x_1, \ldots, x_n$ be n independent observations from the uniform distribution P on $(0,1]$. The ***empirical measure*** P_n is defined as the discrete distribution that puts mass $1/n$ at each of $x_1, \ldots, x_n$. That is, $P_n f := \sum_{i=1}^n f(x_i)/n$, for each function f on $(0,1]$. Notice that nP_nD has a $\text{Bin}(n, PD)$ distribution for each Borel set D. The standardized measure $\nu_n := \sqrt{n}\,(P_n - P)$ is called the ***uniform empirical process***. For each square integrable function f,

$$\nu_n f = n^{-1/2}\sum_{i=1}^n (f(x_i) - Pf) \rightsquigarrow N(0, \sigma_f^2) \qquad \text{where } \sigma_f^2 = Pf^2 - (Pf)^2.$$

More generally, for each finite set of square integrable functions $f_1, \ldots, f_k$, the random vector $(\nu_n f_1, \ldots, \nu_n f_k)$ has a limiting multivariate normal distribution with zero means and covariances $P(f_i f_j) - (Pf_i)(Pf_j)$. These finite dimensional distributions identify a Gaussian process that is closely related to the isonormal process $\{G(f) : f \in L^2(P)\}$ from Section 9.3.

Recall that G is a centered Gaussian process, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\text{cov}(G(f), G(g)) = \langle f, g\rangle = P(fg)$, the $L^2(P)$ inner product. The

Haar basis, $\Psi = \{1\} \cup \{\psi_{i,k} : 0 \le i < 2^k,\ k \in \mathbb{N}_0\}$, for $L^2(P)$ consists of rescaled differences of indicator functions of intervals $J_{i,k} := J(i,k) := (i2^{-k}, (i+1)2^{-k}]$,

$$\psi_{i,k} := 2^{k/2}\left(J_{2i,k+1} - J_{2i+1,k+1}\right) = 2^{k/2}\left(2J_{2i,k+1} - J_{i,k}\right) \qquad \text{for } 0 \le i < 2^k.$$

REMARK. For our current purposes, it is better to replace $L^2(P)$ by $\mathcal{L}^2(P)$, the space of square-integrable real functions whose P-equivalence classes define $L^2(P)$. It will not matter that each $G(f)$ is defined only up to a P-equivalence. We need to work with the individual functions to have $P_n f$ well defined. It need not be true that $P_n f = P_n g$ when f and g differ only on a P-negligible set.

Each function in $\mathcal{L}^2(P)$ has a series expansion,

$$f = (Pf) + \sum_{k=0}^{\infty}\sum_i \psi_{i,k}\langle f, \psi_{i,k}\rangle,$$

which converges in the $\mathcal{L}^2(P)$ sense. The random variables $\overline{\eta} := G(1)$ and $\eta_{i,k} := G(\psi_{i,k})$ are independent, each with a $N(0,1)$ distribution, and

$$G(f) = (Pf)\overline{\eta} + \sum_{k=0}^{\infty}\sum_i \eta_{i,k}\langle f, \psi_{i,k}\rangle,$$

with convergence in the $L^2(\mathbb{P})$ sense. If we center each function f to have zero expectation, we obtain a new Gaussian process, $\nu(f) := G(f - Pf) = G(f) - (Pf)G(1)$, indexed by $\mathcal{L}^2(P)$, whose covariances identify it as the limit process for ν_n. Notice that $\nu(\psi_{i,k}) = G(\psi_{i,k}) = \eta_{i,k}$ almost surely, because $P\psi_{i,k} = 0$. Thus we also have a series representation for ν,

<21> $$\nu(f) = G(f) - (Pf)\overline{\eta} = \sum_{k=0}^{\infty}\sum_i \eta_{i,k}\langle f, \psi_{i,k}\rangle = \sum_{k=0}^{\infty}\sum_i \nu(\psi_{i,k})\langle f, \psi_{i,k}\rangle.$$

At least in a heuristic sense, we could attempt a similar series expansion of the empirical process,

<22> $$\nu_n(f) \stackrel{?}{=} \sum_{k=0}^{\infty}\sum_i \nu_n(\psi_{i,k})\langle f, \psi_{i,k}\rangle.$$

REMARK. Don't worry about the niceties of convergence: when the heuristics are past I will be truncating the series at some finite k.

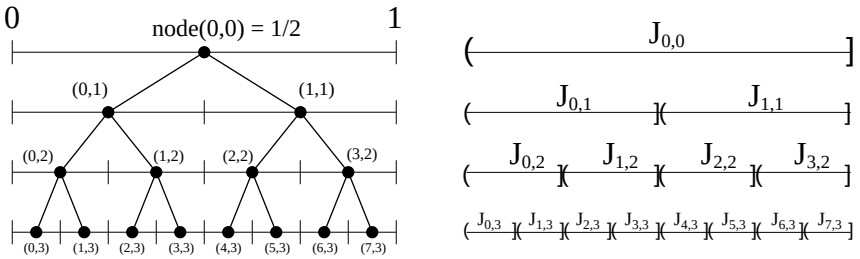
The expansion suggests a way of coupling the process ν_n and ν, namely, find a probability space on which $\nu_n(\psi_{i,k}) \approx \nu(\psi_{i,k}) = \eta_{i,k}$ for a large subset of the basis functions. Such a coupling would have several advantages. First, the peculiarities of each function f would be isolated in the behavior of the coefficients $\langle f, \psi_{i,k}\rangle$. Subject to control of those coefficients, we could derive simultaneous couplings for many different f's. Second, because the $\psi_{i,k}$ functions are rescaled differences of indicator functions of intervals, the $\nu_n(\psi_{i,k})$ are rescaled differences of Binomial counts. Tusnády's Lemma offers an excellent means for building Binomials from standard normals. With some rescaling, we can then build versions of the $\nu_n(\psi_{i,k})$ from the $\eta_{i,k}$.

The secret to success is a recursive argument, corresponding to the nesting of the $J_{i,k}$ intervals. Write node(i,k) for $(i + 1/2)/2^k$, the midpoint of $J_{i,k}$. Regard node$(2i, k+1)$ and node$(2i+1, k+1)$ as the children of node(i,k), corresponding to the decomposition of $J_{i,k}$ into the disjoint union of the two subintervals $J_{2i,k+1}$ and $J_{2i+1,k+1}$. The parent of node(i,k) is node$(\lfloor i/2 \rfloor, k-1)$.

For each integer i with $0 \le i < 2^k$ there is a path back through the tree,

$$\text{path}(i,k) := \{(i_0,0),\ (i_1,1),\ \ldots,\ (i_k,k)\} \qquad \text{where } i_0 = 0 \text{ and } i_k = i,$$

for which $J(i_k,k) \subset J(i_{k-1},k-1) \subset \ldots \subset J(0,0) = (0,1]$. That is, the path traces through all the ancestors (parent, grandparent, ...) back to the root of the tree.



The recursive argument constructs successively refined approximations to P_n by assigning the numbers of observations $X_{i,k}$ amongst $x_1, x_2, \ldots, x_n$ that land in each interval $J_{i,k}$. Notice that, conditional on $X_{i,k} = N$, the two offspring counts must sum to N, with $X_{2i,k+1}$ having a conditional $\text{Bin}(N, 1/2)$ distribution. Via Lemma <19> define

$$X_{0,1} := \tau(n, \eta_{0,0}) =: n - X_{1,1},$$
$$X_{0,2} := \tau(X_{0,1}, \eta_{0,1}) =: X_{0,1} - X_{1,2}, \qquad X_{2,2} := \tau(X_{1,1}, \eta_{1,1}) =: X_{1,1} - X_{3,2},$$

and so on. That is, recursively divide the count $X_{i,k}$ at each node(i,k) between the two children of the node, using the normal variable $\eta_{i,k}$ to determine the $\text{Bin}(X_{i,k}, 1/2)$ count assigned to the child at node$(2i, k+1)$. The joint distribution for the $X_{i,k}$ variables is the same as the joint distribution for the empirical counts $nP_nJ_{i,k}$, because we have used the correct conditional distributions.

If we continued the process forever then, at least conceptually, we would identify the locations of the n observations, without labelling. Each point would be determined by a nested sequence of intervals. To avoid difficulties related to pointwise convergence of the Haar expansion, we need to stop at some finite level, say the mth, after which we could independently distribute the $X_{i,m}$ observations (if any) within $J_{i,m}$.

The recursive construction works well because Tusnády's Lemma, even in its weakened form <20>, provides us with a quadratic bound in the normal variables for the difference between $\nu_n(\psi_{i,k})$ and the corresponding $\eta_{i,k}$.

<23> **Lemma.** *There exists a universal constant C such that, for each k and $0 \le i_k < 2^k$,*

$$|\nu_n(\psi_{i,k}) - \eta_{i,k}| \le \frac{C}{\sqrt{n}} \sum_{j=0}^{k} 2^{j/2} \left(1 + \eta^2_{i_j,j}\right),$$

where $\{(i_j, j) : j = 0, 1, \ldots, k\}$ is a path from the root down to node(i_k, k).

Proof. Abbreviate $J(i_j, j)$ to J_j, and $\eta_{i_j,j}$ to η_j, and so on, for $j = 0, 1, \ldots, k$. Notice that the random variable P_nJ_j has expected value $PJ_j = 2^{-j}$, and a small variance, so we might hope that all of the random variables $\Delta_j := 2^j P_n J_j$ should be close to 1. Of course $\Delta_0 \equiv 1$.

Consider the effect of the split of J_j into its two subintervals, $J' := J(2i_j, j+1)$ and $J'' := J(2i_j+1, j+1)$. Write N for nP_nJ_j and X for nP_nJ', so that $\Delta_j = 2^jN/n$ and $\Delta' := 2^{j+1}P_nJ' = 2^{j+1}X/n$ and $\Delta'' := 2^{j+1}P_nJ'' = 2\Delta_j - \Delta'$. From inequality <20>, we have $X = N/2 + \sqrt{N}\eta_j/2 + R$, where

<24> $$|R| \le C_0(1+\eta_j^2) \qquad \text{and} \qquad |X - N/2| \le C_0\sqrt{N}\left(1+|\eta_j|\right).$$

By construction,

$$P_nJ' = \frac{X}{n} = \frac{N+\sqrt{N}\eta_j + 2R}{2n} = \tfrac{1}{2}\left(P_nJ_j + \sqrt{\frac{\Delta_j}{n2^j}}\,\eta_j + \frac{2R}{n}\right),$$

and hence

$$\nu_n\psi_j = \sqrt{n2^j}\,P_n\left(2J' - J_j\right) = \sqrt{\Delta_j}\eta_j + 2\sqrt{\frac{2^j}{n}}R.$$

From the first inequality in <24>,

<25> $$|\nu_n\psi_j - \eta_j| \le \left|\left(\sqrt{\Delta_j}-1\right)\eta_j\right| + 2C_0\sqrt{\frac{2^j}{n}}\left(1+\eta_j^2\right).$$

From the second inequality in <24>,

$$|\Delta'' - \Delta_j| = |\Delta' - \Delta_j| = \frac{2^{j+1}}{n}|X - N/2| \le C_0\sqrt{\frac{2^{j+2}}{n}}\sqrt{\Delta_j}\left(1+|\eta_j|\right).$$

Invoke the inequality $|\sqrt{a}-\sqrt{b}| \le |a-b|/\sqrt{b}$, for positive a and b, to deduce that

$$|\sqrt{\Delta_{j+1}} - \sqrt{\Delta_j}| \le \max\left(|\sqrt{\Delta'} - \sqrt{\Delta_j}|, |\sqrt{\Delta''} - \sqrt{\Delta_j}|\right) \le 2C_02^{j/2}\left(1+|\eta_j|\right)/\sqrt{n}.$$

From <25> with $j = k$, and the inequality from the previous line, deduce that

$$\begin{aligned}\sqrt{n}|\nu_n\psi_k - \eta_k| &\le 2C_02^{k/2}\left(1+\eta_k^2\right) + \sqrt{n}|\eta_k|\textstyle\sum_{j=0}^{k-1}|\sqrt{\Delta_{j+1}} - \sqrt{\Delta_j}| \\ &\le 2C_02^{k/2}\left(1+\eta_k^2\right) + 2C_0\textstyle\sum_{j=0}^{k-1}2^{j/2}\left(|\eta_k| + |\eta_k\eta_j|\right).\end{aligned}$$

Bound $|\eta_k| + |\eta_k\eta_j|$ by $1 + \eta_k^2 + \frac{1}{2}\eta_k^2 + \frac{1}{2}\eta_j^2$, then collect terms involving η_k^2, to complete the proof. □

7. The Komlós-Major-Tusnády coupling

The coupling method suggested by expansions <21> and <22> works particularly well when restricted to the set of indicator functions of intervals, $f_t(x) = \{0 < x \le t\}$, for $0 < t \le 1$. For that case, the limit process $\{\nu(0,t] : 0 \le t \le 1\}$, which can be chosen to have continuous sample paths, is called the ***Brownian Bridge***, or ***tied-down Brownian motion***, often written as $\{B^\circ(t) : 0 \le t \le 1\}$.

<26> **Theorem.** *(KMT coupling) There exists a Brownian Bridge $\{B^\circ(t) : 0 \le t \le 1\}$ with continuous sample paths, and a uniform empirical process ν_n, for which*

$$\mathbb{P}\left\{\sup_{0\le t\le 1}|\nu_n(0,t] - B^\circ(t)| \ge C_1\frac{x+\log n}{\sqrt{n}}\right\} \le C_0\exp(-x) \qquad \text{for all } x \ge 0,$$

with constants C_1 and C_0 that depend on neither n nor x.

REMARK. Notice that the exponent on the right-hand side is somewhat arbitrary; we could change it to any other positive multiple of $-x$ by changing the constant C_1 on the left-hand side. By the same reasoning, it would suffice to get a bound like $C_2 \exp(-c_2 x) + C_3 \exp(-c_3 x) + C_4 \exp(-c_4 x)$ for various positive constants C_i and c_i, for then we could recover the cleaner looking version by adjusting C_1 and C_0. In my opinion, the exact constants are unimportant; the form of the inequality is what counts. Similarly, it would suffice to consider only values of x bounded away from zero, such as $x \geq c_0$, because the asserted inequality is trivial for $x < c_0$ if $C_0 \geq e^{c_0}$.

It is easier to adjust constants at the end of an argument, to get a clean-looking inequality. When reading proofs in the literature, I sometimes find it frustrating to struggle with a collection of exquisitely defined constants at the start of a proof, eventually to discover that the author has merely been aiming for a tidy final bound.

Proof. We will build ν_n from B°, allocating counts down to intervals of length 2^{-m}, as described in Section 6. It will then remain only to control the behavior of both processes over small intervals. Let $T(m)$ denote the set of grid points $\{i/2^m : i = 0, 1, \ldots, 2^m\}$ in $[0, 1]$. For each t in $T(m)$, both series <21> and <22> terminate after $k = m$, because $[0, t]$ is orthogonal to each $\psi_{i,k}$ for $k > m$. That is, using the Hungarian construction we can determine $P_n J_{i,m}$ for each i, and then calculate

$$\nu_n(0, t] = \sum_{k=0}^{m} \sum_i \nu_n(\psi_{i,k}) \langle f_t, \psi_{i,k} \rangle \qquad \text{for } t \text{ in } T(m),$$

which we need to show is close to

$$B^\circ(t) := \nu(0, t] = \sum_{k=0}^{m} \sum_i \eta_{i,k} \langle f_t, \psi_{i,k} \rangle \qquad \text{for } t \text{ in } T(m).$$

Notice that $B^\circ(0) = B^\circ(1) = 0 = \nu_n(0, 0] = \nu_n(0, 1]$. We need only consider t in $T(m) \backslash \{0, 1\}$. For each k, at most one coefficient $\langle f_t, \psi_{i,k} \rangle$ is nonzero, corresponding to the interval for which $t \in J_{i,k}$, and it is bounded in absolute value by $2^{-k/2}$. The corresponding nodes determine a path $(0, 0), \ldots, (i_j, j), \ldots, (i_m, m)$ down to the mth level. The difference between the processes at t is controlled by the quadratic function,

$$S_m(t) := \sum_{j=0}^{m} \eta_{i_j,j}^2 \qquad \text{where } t \in J(i_j, j) \text{ for each } j,$$

of the normal variables at the nodes of this path:

$$\begin{aligned}
\sqrt{n} |\nu_n(0, t] - B^\circ(t)| &\leq \sum_{k=0}^{m} \sqrt{n} |\nu_n(\psi_{i_k,k}) - \eta_{i_k,k}| 2^{-k/2} \\
&\leq \sum_{j,k} \{0 \leq j \leq k \leq m\} C 2^{(j-k)/2} \left(1 + \eta_{i_j,j}^2\right) \qquad \text{by Lemma <23>} \\
&\leq 4C \sum_{j=0}^{m} \left(1 + \eta_{i_j,j}^2\right) \qquad \text{summing the geometric series} \\
&= 4C \left(m + 1 + S_m(t)\right). \qquad \text{<27>}
\end{aligned}$$

As t ranges over $T(m)$, or even over the whole of $(0, 1)$, the path defining $S_m(t)$ ranges over the set of all 2^m paths from the root down to the mth level. We bound the maximum difference between the two processes if we bound the maximum of $S_m(t)$. The maximum grows roughly linearly with m, the same rate as the contribution from a single t. More precisely

$$\mathbb{P}\{\max_t S_m(t) \geq 5m + x\} \leq 2 \exp(-x/4) \qquad \text{for each } x \geq 0. \qquad \text{<28>}$$

I postpone the proof of this result, in order not to break the flow of the main argument.

REMARK. The constants 5 and 4 are not magical. They could be replaced by any other pair of constants for which $\mathbb{P}\exp\left((N(0,1)^2 - c_1)/c_2\right) \le 1/2$.

From inequalities <27> and <28> we have

<29> $$\mathbb{P}\left\{\max_{t\in T(m)} |\nu_n(0,t] - B^\circ(t)| \ge 4C\frac{1+x+6m}{\sqrt{n}}\right\} \le \mathbb{P}\left\{\max_t S_m(t) \ge x+5m\right\} \le \exp(-x/4).$$

Provided we choose m smaller than a constant multiple of $x + \log n$, this term will cause us no trouble.

We now have the easy task of extrapolating from the grid $T(m)$ to the whole of $(0,1)$. We can make 2^{-m} exceedingly small by choosing m close to a large enough multiple of $x + \log n$. In fact, when $x \ge 2$, the choice of m such that

<30> $$2n^2e^x > 2^m \ge n^2e^x$$

will suffice. As an exercise, you might want to play around with other m and the various constants to get a neater statement for the Theorem.

We can afford to work with very crude estimates. For each s in $(0,1)$ write t_s for the point of $T(m)$ for which $t_s \le s < t_s + 2^{-m}$. Notice that

$$|\nu_n(0,s] - \nu_n(0,t_s]| \le \#\text{ points in } (t_s,s]/\sqrt{n} + \sqrt{n}2^{-m}.$$

The supremum over s is larger than $3/\sqrt{n}$ only when at least one $J_{i,m}$ interval, for $0 \le i < 2^m$ contains 2 or more observations, an event with probability less than

$$2^m\binom{n}{2}(2^{-m})^2 \le n^2 2^{-m} \le e^{-x} \qquad \text{for } m \text{ as in } <30>.$$

Similarly,

$$\begin{aligned}\sup_s |B^\circ(s) - B^\circ(t_s)| &\le \sup_s |G[0,s] - G[0,t_s]| + \sup_s |(s-t_s)\overline{\eta}| \\ &\le \max_{0\le i<2^m} \sup_{s\in J_{i,m}} |G[0,s] - G[0,i/2^m]| + 2^{-m}|\overline{\eta}|\end{aligned}$$

from which it follows that

$$\mathbb{P}\left\{\sup_s |B^\circ(s) - B^\circ(t_s)| \ge \frac{2x}{\sqrt{n}}\right\} \le 2^m\mathbb{P}\left\{\sup_{0\le s\le 2^{-m}} |B(s)| \ge \frac{x}{\sqrt{n}}\right\} + \mathbb{P}\left\{|N(0,1)| \ge \frac{2^m x}{\sqrt{n}}\right\}$$

where B is a Brownian motion. The second term on the right-hand side is less than $\exp\left(-4^m x^2/2n\right)$. By the reflection principle for Brownian motion (Section 9.5), the first term equals

$$2^m\mathbb{P}\left\{|B(2^{-m})| \ge \frac{x}{\sqrt{n}}\right\} = 2^{m+1}\mathbb{P}\left\{|N(0,1)| \ge \frac{2^{m/2}x}{\sqrt{n}}\right\} \le 2^{m+1}\exp\left(-\frac{2^m x^2}{2n}\right).$$

For $x \ge 2$ and m as in <30>, the sum of the two contributions from the Brownian Bridge is much smaller than e^{-x}.

From <29>, and the inequality

$$|\nu_n(0,s] - B^\circ(s)| \le |\nu_n(0,s] - \nu_n(0,t_s]| + |\nu_n(0,t_s] - B^\circ(t_s)| + |B^\circ(t_s) - B^\circ(s)|,$$

together with the bounds from the previous paragraph, you should be able to complete the argument. □

Proof of inequality <28>. Write R_m for $\max_t S_m(t)$. Think of the binary tree of depth m as two binary trees of depth $m-1$ rooted at node$(0,1)$ and node$(1,1)$, to see that that R_m has the same distribution as $\eta_{0,0}^2 + \max(T, T')$, where T and T' both have the same distribution as R_{m-1}, and $\eta_{0,0}$, T, and T' are independent. Write D_k for $\mathbb{P}\exp((R_k - 5k)/4)$. Notice that

$$e^{-5/4} D_0 = \mathbb{P}\exp\left(\tfrac{1}{4}\eta_{0,0}^2 - \tfrac{5}{4}\right) = \sqrt{2}\exp(-5/4) < 1/2.$$

For $m \geq 1$, independence lets us bound D_m by

$$\begin{aligned}\mathbb{P}\exp\left(\tfrac{1}{4}\eta_{0,0}^2 - \tfrac{5}{4}\right) \mathbb{P}\left(\tfrac{1}{4}\max\left(T - 5(m-1), T' - 5(m-1)\right)\right)& \\ < \tfrac{1}{2}\Big(\mathbb{P}\exp\left(\tfrac{1}{4}T - \tfrac{5}{4}(m-1)\right) + \mathbb{P}\exp\left(\tfrac{1}{4}T' - \tfrac{5}{4}(m-1)\right)\Big) = D_{m-1}.&\end{aligned}$$

By induction, $\mathbb{P}\exp(R_m - 5m)/4 = D_m \leq D_0 = \sqrt{2}$. Thus

$$\mathbb{P}\{R_m \geq 5m + x\} \leq \mathbb{P}\exp\big((R_m - 5m)/4\big)\exp(-x/4) \leq \sqrt{2}\exp(-x/4),$$

as asserted. □

By means of the quantile transformation, Theorem <26> extends immediately to a bound for the empirical distribution function F_n generated from a sample $\xi_1, \ldots, \xi_n$ from a probability measure on the real line with distribution function F. Again writing q_F for the quantile function, and recalling that we can generate the sample as $\xi_i = q_F(x_i)$, we have

$$nF_n(t) = \sum\nolimits_{i \leq n}\{\xi_i \leq t\} = \sum\nolimits_{i \leq n}\{q_F(x_i) \leq t\} = \sum\nolimits_{i \leq n}\{x_i \leq F(t)\},$$

which implies $\sqrt{n}\,(F_n(t) - F(t)) = \nu_n\big(0, F(t)\big]$. Notice that $F(t)$ ranges over a subset of $[0,1]$ as t ranges over $\mathbb{R}$; and when F has no discontinuities, the range covers all of $(0,1)$. Theorem <26> therefore implies

$$\text{<31>} \qquad \mathbb{P}\left\{\sup_t |\sqrt{n}\,(F_n(t) - F(t)) - B^\circ(F(t))| \geq C_1 \frac{x + \log n}{\sqrt{n}}\right\} \leq C_0 e^{-x} \qquad \text{for } x \geq 0.$$

Put another way, we have an almost sure representation $F_n(t) = F(t) + n^{-1/2}B^\circ(F(t)) + R_n(t)$, where, for example, $\sup_t |R_n(t)| = O_p\left(n^{-1}\log n\right)$.

> REMARK. From a given Brownian Bridge B° and a given n we have constructed a sample $x_1, \ldots, x_n$ from the uniform distribution. From the same B°, we could also generate a sample $x_1', \ldots, x_n', x_{n+1}'$ of size $n+1$. However, it is not true that $x_i = x_i'$ for $i \leq n$; it is not true that $x_1, \ldots, x_n, x_{n+1}'$ are mutually independent. If we wished to have the samples relate properly to each other we would have to change the Brownian Bridge with n. There is a version of KMT called the ***Kiefer*** coupling, which gets the correct joint distributions between the samples at the cost of a weaker error bound. See Csörgő & Révész (1981, Chapter 4) for further explanation.

Inequality <31> lets us deduce results about the empirical distribution function F_n from analogous results about the Brownian Bridge. For example, it implies $\sup_t \sqrt{n}|F_n(t) - F(t)| \rightsquigarrow \sup_t |B^\circ(F(t))|$. If F has no discontinuities, the limit distribution is the same as that of $\sup_s |B^\circ(s)|$. That is, we have an instant derivation of the Kolmogorov-Smirov theorem. The Csörgő & Révész book describes other consequences that make much better use of all the hard work that went into establishing the KMT inequality.

8. Problems

[1] Suppose F and F_n, for $n \in \mathbb{N}$ are distribution functions on the real line for which $F_n(x) \to F(x)$ for each x in a dense subset D of the real line. Show that the corresponding quantile functions Q_n converge pointwise to Q at all except (at worst) a countable subset of points in $(0,1)$. Hint: Prove convergence at each continuity point u_0 of Q. Given points x', x'' in D with $x' < x_0 = Q(u_0) < x''$, find $\delta > 0$ such that $x' < Q(u_0 - \delta)$ and $Q(u_0 + \delta) \le x''$. Deduce that

$$F_n(x') < F(x') + \delta < u_0 \le F(x'') - \delta \le F_n(x'') \qquad \text{eventually,}$$

in which case $x' < Q_n(u_0) \le x''$.

[2] Let P and Q be two probability measures defined on the same sigma-field $\mathcal{A}$ of a set $\mathcal{X}$. The total variation distance $v = v(P,Q)$ is defined as $\sup_{A\in\mathcal{A}} |PA - QA|$.

(i) Suppose X are Y are random elements of $\mathcal{X}$, defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with distributions P and Q. Show that $\mathbb{P}^*\{X \ne Y\} \ge v(P,Q)$. Hint: Choose a measurable set $D \supseteq \{X \ne Y\}$ with $\mathbb{P}D = \mathbb{P}^*\{X \ne Y\}$. Note that $\mathbb{P}\{X \in A\} - \mathbb{P}\{Y \in A\} = \mathbb{P}\{X \in A\} \cap D - \mathbb{P}\{Y \in A\} \cap D$.

(ii) Suppose the diagonal $\Delta := \{(x,y) \in \mathcal{X} \times \mathcal{X} : x = y\}$ is product measurable. Recall from Section 3.3 that $v = 1 - (P \wedge Q)(\mathcal{X}) = (P-Q)^+(\mathcal{X}) = (Q-P)^+(\mathcal{X})$. Define a probability measure $\mathbb{P} = \frac{1}{v}(P-Q)^+ \otimes (Q-P)^+ + \lambda$, where λ is the image of $P \wedge Q$ under the map $x \mapsto (x,x)$. Let X and Y be the coordinate maps. Show that X has distribution P and Y has distribution Q, and $\mathbb{P}\{X \ne Y\} = v$.

[3] Show that the Prohorov distance is a metric. Hint: For the triangle inequality, use the inclusion $(B^\epsilon)^{\epsilon'} \subseteq B^{\epsilon+\epsilon'}$. For symmetry, consider $\rho(P,Q) < \delta < \epsilon$. Put $D^c = B^\epsilon$. Prove that $D^\delta \subseteq B^c$, then deduce that $1 - PB^\epsilon \le QD^\delta + \delta \le 1 - QB + \delta$.

[4] Let P be a Borel probability measure concentrated on the closure of a countable subset $S = \{x_i : i \in \mathbb{N}\}$ of a metric space $\mathcal{X}$. For fixed $\epsilon > 0$, follow these steps to show that there exist a partition of $\mathcal{X}$ into finitely many P-continuity sets $C_0, C_1, \ldots, C_m$ such that $PC_0 < \epsilon$ and diameter$(C_i) < \epsilon$ for $i \ge 1$.

(i) For each x in $\mathcal{X}$, show that there are at most countably many closed balls B centered at x with $P(\partial B) > 0$.

(ii) For each x_i in S, find a ball B_i centered at x_i with radius between $\epsilon/4$ and $\epsilon/2$ and $P(\partial B_i) = 0$.

(iii) Show that $\cup_{i\in\mathbb{N}} B_i$ contains the closure of S. Hint: Each point of the closure lies within $\epsilon/4$ of at least one x_i.

(iv) Show that $P\left(\cup_{i\le m} B_i\right) > 1 - \epsilon$ when m is large enough.

(v) Show that the sets $C_i := B_i \setminus \cup_{1\le j<i} B_j$ and $C_0 := \left(\cup_{i\le m} B_i\right)^c$ have the desired properties.

[5] (Ye Olde Marriage Lemma) Suppose S is a finite set of princesses. Suppose each princess, σ, has a list, $K(\sigma)$, of frogs desirable for marriage. For each collection $A \subseteq S$, the combined list of frogs equals $K(A) = \bigcup\{K(\sigma) : \sigma \in A\}$. If

each princess is to find a frog on her list to marry, then clearly the "Desirable Frog Condition" (DFC), $\#K(A) \geq \#A$, for each $A \subseteq S$, must be satisfied. Show that DFC is also sufficient for happy princesses: under the DFC there exists a one-to-one map π from S into $K(S)$ such that $\pi(\sigma) \in K(\sigma)$ for every σ in S. Hint: Translate the following mathematical fairy tale into an inductive argument.

(i) Once upon a time there was a princess σ_0 who proposed to marry a frog τ_0 from her list. That would have left a collection $S\backslash\{\tau_0\}$ of princesses with lists $K(\sigma)\backslash\{\tau_0\}$ to choose from. If the analog of the DFC had held for those lists, an induction hypothesis would have made everyone happy.

(ii) Unfortunately, a collection $A_0 \subseteq S\backslash\{\sigma_0\}$ of princesses protested, on the grounds that $\#\mathcal{K}(A_0)\backslash\{\tau_0\} < \#A_0$; clearly not enough frogs to go around. They pointed out that the DFC held with equality for A_0, and that their happiness could be assured only if they had exclusive access to the frogs in $K(A_0)$.

(iii) Everyone agreed with the assertion of the A_0. They got their exclusive access, and, by induction, lived happily ever after.

(iv) The other princesses then got worried. Each collection B in $S\backslash A_0$ asked, "$\#K(B)\backslash K(A_0) \geq \#B$?" They were reassured, "Don't worry. Originally $\#K(B \cup A_0) \geq \#B + \#A_0$, and we all know that $\#K(A_0) = \#A_0$, so of course

$$\#K(B)\backslash K(A_0) = \#K(B \cup A_0) - \#K(A_0) \geq \#B.$$

You too can live happily ever after, by induction." And they did.

[6] Prove Lemma <9> by carrying out on the following steps. Write R_A for $\cup_{\alpha\in A} R_\alpha$. Argue by induction on the size of S. With no loss of generality, suppose $S = \{1, 2, \ldots, m\}$. Check the case $m = 1$. Work from the inductive hypothesis that the result is true for $\#S < m$.

(i) Suppose there exists a proper subset A_0 of S for which $\nu A_0 = \mu R_{A_0}$. Define $R'_\alpha = R_\alpha\backslash R_{A_0}$ for $\alpha \notin A_0$. Show that $\nu A \leq \mu R'_A$ for all $A \subseteq S\backslash A_0$. Construct K by invoking the inductive hypothesis separately for A_0 and $S\backslash A_0$. (Compare with part (iv) of Problem [5].)

Now suppose $\nu A < \mu R_A$ for all proper subsets A of S. Write L_α for the probability distribution $\mu(\cdot \mid R_\alpha)$, which concentrates on R_α.

(ii) Show that $\mu \geq \nu\{1\}L_1$. Hint: Show $\mu B \geq \nu\{1\}\mu(BR_1)/\mu R_1$ for all $B \subseteq R_1$.

Write ϵ_1 for the unit mass at 1. Let θ_0 be the largest value in $[0, \nu\{1\}]$ for which $(\mu - \theta_0 L_1)R_A \geq (\nu - \theta_0\epsilon_1)A$ for every $A \subseteq S$.

(iii) If $\theta_0 = \nu\{1\}$, use the inductive hypothesis to find a probability kernel from $S\backslash\{1\}$ into T for which $(\mu - \nu\{1\}L_1) \geq \sum_{\alpha\geq 2} \nu\{\alpha\}K_\alpha$. Define $K_1 = L_1$.

(iv) If $\theta_0 < \nu\{1\}$, show that there exists an $A_0 \subseteq S$ for which $(\mu - \theta L_1)R_{A_0} < (\nu - \theta\epsilon_1)A_0$ when $\nu\{1\} \geq \theta > \theta_0$. Deduce that A_0 must be a proper subset of S for which $(\mu - \theta_0 L_1)R_{A_0} = (\nu - \theta_0\epsilon_1)A_0$. Invoke part (i) to find a probability kernel M for which $\mu - \theta_0 L_1 \geq (\nu\{1\} - \theta_0)M_1 + \sum_{\alpha\geq 2} \nu\{\alpha\}M_\alpha$. Define $K_1 := (\theta_0/\nu\{1\})L_1 + (1 - \theta_0/\nu\{1\})M_1$.

[7] Establish the bound $\mathbb{P}\{|N(0, I_k)| > \sqrt{kx}\} \le \left(xe^{1-x}\right)^{k/2}$, for $x > 1$, as needed (with $\sqrt{kx} = \delta/\sigma$) for the proof of Lemma <18>. Hint: Show that

$$\mathbb{P}\{|N(0, I_k)|^2 > kx\} \le \exp(-tkx)(1-2t)^{-k/2} \qquad \text{for } 0 < t < 1/2$$

which is minimized at $t = \frac{1}{2}(1 - x^{-1})$.

[8] Let F_m and G_n be empirical distribution functions, constructed from independent samples (of sizes m and n) from the same distribution function F on the real line. Show that

$$\sqrt{\frac{mn}{m+n}}\,\sup_t |F_m(t) - G_n(t)| \rightsquigarrow \sup_t |B^\circ(F(t))| \qquad \text{as } \min(m,n) \to \infty.$$

Hint: Use <31>. Show that $\alpha B_1^\circ(s) + \beta B_1^\circ(s)$ is a Brownian Bridge if $\alpha^2 + \beta^2 = 1$ and B_1°, B_1° are independent Brownian Bridges.

9. Notes

In increasing degrees of generality, representations as in Theorem <4> are due to Skorohod (1956), Dudley (1968), Wichura (1970), and Dudley (1985).

Prohorov (1956) defined his metric for probability measures on complete, separable metric spaces. Theorem <8> is due to Strassen (1965). I adapted the proof from Dudley (1976, Section 18), who used the Marriage Lemma (Problem [5]) to prove existence of the desired coupling in a special discrete case. Lemma <9> is a continuous analog of the Marriage Lemma, slightly extending the method of Pollard (1984, Lemma IV.24).

The discussion in Section 4 is adapted from an exposition of Yurinskii (1977)'s method by Le Cam (1988). I think the slightly weaker bound stated by Yurinskii may be the result of his choosing a slightly different tail bound for $|N(0, I_k)|$, with a correspondingly different choice for the smoothing parameter.

The idea for Example <17> comes from the construction used by Dudley & Philipp (1983) to build strong approximations for sums of independent random processes taking values in a Banach space. Massart (1989) refined the coupling technique, as applied to empiricial processes, using a Hungarian coupling in place of the Yurinskii coupling.

The proof of the KMT approximation in the original paper (Komlós et al. 1975) was based on the analog of the first inequality in <20>, for $|X - n/2|$ smaller than a tiny multiple of n. The proof of the elegant refinement in Lemma <19> appeared in a 1977 dissertation of Tusnády, in Hungarian. I have seen an annotated extract from the dissertation (courtesy of Sándor Csörgő). Csörgő & Révész (1981, page 133) remarked that Tusnády's proof is "elementary" but not "simple". I agree. Bretagnolle & Massart (1989, Appendix) published another proof, an exquisitely delicate exercise in elementary calculus and careful handling of Stirling's approximation. The method used in Appendix D resulted from a collaboration between Andrew Carter and me.

Lemma <23> repackages a construction from Komlós et al. (1975) that has been refined by several authors, most notably Bretagnolle & Massart (1989), Massart (1989), and Koltchinskii (1994).

References

Bretagnolle, J. & Massart, P. (1989), 'Hungarian constructions from the nonasymptotic viewpoint', *Annals of Probability* **17**, 239–256.

Csörgő, M. & Révész, P. (1981), *Strong Approximations in Probability and Statistics*, Academic Press, New York.

Dudley, R. M. (1968), 'Distances of probability measures and random variables', *Annals of Mathematical Statistics* **39**, 1563–1572.

Dudley, R. M. (1976), 'Convergence of laws on metric spaces, with a view to statistical testing'. Lecture Note Series No. 45, Matematisk Institut, Aarhus University.

Dudley, R. M. (1985), 'An extended Wichura theorem, definitions of Donsker classes, and weighted empirical distributions', *Springer Lecture Notes in Mathematics* **1153**, 141–178. Springer, New York.

Dudley, R. M. & Philipp, W. (1983), 'Invariance principles for sums of Banach space valued random elements and empirical processes', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **62**, 509–552.

Kim, J. & Pollard, D. (1990), 'Cube root asymptotics', *Annals of Statistics* **18**, 191–219.

Koltchinskii, V. I. (1994), 'Komlós-Major-Tusnády approximation for the general empirical process and Haar expansion of classes of functions', *Journal of Theoretical Probability* **7**, 73–118.

Komlós, J., Major, P. & Tusnády, G. (1975), 'An approximation of partial sums of independent rv-s, and the sample df. I', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **32**, 111–131.

Le Cam, L. (1988), On the Prohorov distance between the empirical process and the associated Gaussian bridge, Technical report, Department of Statistics, U.C. Berkeley. Technical report No. 170.

Major, P. (1976), 'The approximation of partial sums of independent rv's', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **35**, 213–220.

Massart, P. (1989), 'Strong approximation for multivariate empirical and related processes, via KMT constructions', *Annals of Probability* **17**, 266–291.

Pollard, D. (1984), *Convergence of Stochastic Processes*, Springer, New York.

Pollard, D. (1990), *Empirical Processes: Theory and Applications*, Vol. 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*, Institute of Mathematical Statistics, Hayward, CA.

Prohorov, Yu. V. (1956), 'Convergence of random processes and limit theorems in probability theory', *Theory Probability and Its Applications* **1**, 157–214.

Skorohod, A. V. (1956), 'Limit theorems for stochastic processes', *Theory Probability and Its Applications* **1**, 261–290.

Strassen, V. (1965), 'The existence of probability measures with given marginals', *Annals of Mathematical Statistics* **36**, 423–439.

Tusnády, G. (1977), A study of Statistical Hypotheses, PhD thesis, Hungarian Academy of Sciences, Budapest. In Hungarian.

Wichura, M. J. (1970), 'On the construction of almost uniformly convergent random variables with given weakly convergent image laws', *Annals of Mathematical Statistics* **41**, 284–291.

Yurinskii, V. V. (1977), 'On the error of the Gaussian approximation for convolutions', *Theory Probability and Its Applications* **2**, 236–247.