Chapter 11
Convergence in Distribution

Chapter 11

Convergence in Distribution

1 Weak convergence in metric spaces

Suppose that (M, d) is a metric space, and let $\mathcal{M}$ denote the Borel sigma-field (the sigma field generated by the open sets in M). Let $C_b(M)$ denote the set of all real-valued, bounded continuous functions on M, and let $C_u(M)$ denote the set of all real-valued, bounded uniformly continuous functions on M.

Definition 1.1 (weak convergence) If $\{P_n\}$, P are probability measures on $(M, \mathcal{M})$ satisfying

$$\int f dP_n \to \int f dP \qquad \text{as} \;\; n \to \infty \qquad \text{for all} \;\; f \in C_b(M)$$

then we say that P_n converges in distribution (or law) to P, or that P_n converges weakly to P, and we write $P_n \to_d P$ or $P_n \Rightarrow P$. Similarly, if $\{X_n\}$ are random elements in M (i.e. measurable maps from some probability space(s) $(\Omega, \mathcal{A}, Pr)$ (or $(\Omega_n, \mathcal{A}_n, Pr_n)$) to $(M, \mathcal{M})$) with

$$Ef(X_n) \to Ef(X) \qquad \text{for all} \;\; f \in C_b(M)\,,$$

then we write $X_n \to_d X$ or $X_n \Rightarrow X$.

Definition 1.2 (boundary and P-continuity set) For any set $B \in \mathcal{M}$, the *boundary* of B is $\partial B \equiv \overline{B} \setminus B^o$ where $\overline{B}$ is the closure of B and B^o is the interior of B; i.e. the largest open set contained in B. A set B is called a *continuity set* of P if $P(\partial B) = 0$.

Definition 1.3 (Bounded Lipschitz functions) A real-valued function f on a metric space (M, d) is said to satisfy a *Lipschitz condition* if there exists a finite constant K for which

$$|f(x) - f(y)| \le K d(x, y) \qquad \text{for all} \;\; x, y \in M\,.$$

We write $BL(M)$ for the vector space of all bounded Lipshitz functions on M.

We can characterize the space $BL(M)$ in terms of a norm $\|f\|_{BL}$ defined for all real valued functions f on M as follows:

$$\|f\|_{BL} \equiv \max\{K_1(f), 2K_2(f)\}$$

where

$$K_1(f) \equiv \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)}, \qquad K_2(f) \equiv \sup_x |f(x)| .$$

Here we have followed Pollard (2002), who deviates from the usual definition of $\|f\|_{BL}$ in order to obtain the following nice inequality:

$$|f(x) - f(y)| \le \|f\|_{BL}\{1 \wedge d(x,y)\} \qquad \text{for all} \quad x, y \in M .$$

Definition 1.4 (Lower and upper semicontinuous functions) A function $f : M \to \mathbb{R}$ is said to be *lower semicontinuous* (or LSC) if $\{x : \ f(x) > t\}$ is an open set for each fixed t. A function f is said to be *upper semicontinuous* (or USC) if $\{x : f(x) < t\}$ is open for each fixed t.

Thus f is USC if and only if $-f$ is LSC. If f is both USC and LSC then it is continuous. The basic example of a lower semicontinuous function is the indicator function 1_B of an open set B; the basic example of an upper semicontinuous function is the indicator function 1_B of a closed set B. Our first theorem will use the following result connecting lower semicontinuous functions to functions in $BL(M)$.

Lemma 1.1 (LSC approximation) Let g be a lower semicontinuous function bounded from below on a metric space M. Then there exists a sequence $\{f_m\}_{m=1}^{\infty} \subset BL(M)$ satisfying $f_m(x) \uparrow g(x)$ for each $x \in M$.

Proof. We may assume that $g \ge 0$ without loss of generality (if not, replace g by $g + \sup_x(-g(x))$). For each $t > 0$ the set $B_t \equiv \{x : g(x) \le t\}$ is closed. The sequence of functions $f_{k,t}(x) \equiv t \wedge (kd(x, B_t))$ for $k \in \mathbb{N}$ are in $BL(M)$ and satisfy $f_{k,t}(x) \uparrow t1_{B_t^c}(x) = t1_{[g(x)>t]}$ since $d(x, B_t) > 0$ if and only if $g(x) > t$.

Now consider the countable collection $\mathcal{G} = \cup_{k \in \mathbb{N}} \cup_{t \in \mathbb{Q}^+} \{g_{k,t}\}$ where $\mathbb{Q}$ is the set of all rational numbers. The pointwise supremum of $\mathcal{G}$ is g. If we enumerate $\mathcal{G}$ as $\{g_1, g_2, \ldots\}$, and then define $f_m \equiv \max_{j \le m} g_j$, it follows that f_m is in $BL(M)$ for each m and $f_m \uparrow g$. $\square$

Our first result gives a number of equivalences to the definition of weak convergence given in Definition 1.1.

Theorem 1.1 (portmanteau theorem) For probability measures $\{P_n\}$, P on $(M, \mathcal{M})$ the following are equivalent:

(i) $\displaystyle\int f dP_n \to \int f dP \qquad \text{for all} \quad f \in C_b(M); \quad \text{i.e. } P_n \to_d P .$

(ii) $\displaystyle\int f dP_n \to \int f dP \qquad \text{for all} \quad f \in C_u(M) .$

(iii) $\displaystyle\int f dP_n \to \int f dP \qquad \text{for all} \quad f \in BL(M) .$

(iv) $\displaystyle\limsup_{n\to\infty} \int f dP_n \le \int f dP$ for every upper semicontinuous f bounded from above.

(v) $\displaystyle\liminf_{n\to\infty} \int f dP_n \ge \int f dP$ for every lower semicontinuous f bounded from below.

(vi) $\limsup_{n\to\infty} P_n(B) \le P(B)$ for all closed sets $B \in \mathcal{M}$.

(vii) $\liminf_{n\to\infty} P_n(B) \ge P(B)$ for all open sets $B \in \mathcal{M}$.

(viii) $\lim_{n\to\infty} P_n(B) = P(B)$ for all $P-$continuity sets $B \in \mathcal{M}$.

(ix) $\lim_{n\to\infty} \int f dP_n = \int f dP$ for all bounded measurable functions f with $P(C_f) = 1$.

Proof. Clearly (i) implies (ii) and (ii) implies (iii) since $BL(M) \subset C_u(M) \subset C_b(M)$. We also note that (iv) and (v) are equivalent since $-f$ is lower semicontinuous and bounded from below if f is upper semicontinuous and bounded from above. Similarly, (vi) and (vii) are equivalent by taking complements. Since the indicator function of an open set is lower semicontinuous and bounded from below, (v) implies (vii), (and similarly, (iv) implies (vi)).

Now we use Lemma 1.1 to show that (iii) implies (v): suppose that (iii) holds, and let g be a LSC function bounded from below. By Lemma 1.1 there exists a sequence $\{f_m\}$ in $BL(M)$ with $f_m \uparrow g$ pointwise. Then, for each fixed m we have

$$\liminf_n \int g dP_n \ge \liminf_n \int f_m dP_n = \int f_m dP \qquad \text{since} \quad \int f_m dP_n \to \int f_m dP$$

by (iii). Take the supremum over m; by the monotone convergence theorem the right side in the last display converges to $\int g dP$, and thus (v) holds.

To see that (vi) and (vii) imply (viii), let B be a $P-$continuity set. Then since B^o is open and $\overline{B}$ is closed,

$$P(B^o) \le \liminf P_n(B^o) \le \liminf P_n(B) \le \limsup P_n(B) \le \limsup P_n(\overline{B}) \le P(\overline{B}).$$

Since B is a $P-$continuity set $P(\partial B) = 0$ and $P(\overline{B}) = P(B^o)$, so the extreme terms in the last display are equal and hence $\lim P_n(B) = P(B)$.

Next we show that (viii) implies (vi): Let B be a closed set and suppose that (viii) holds. Since $\partial\{x : \ d(x,B) \le \delta\} \subset \{x : \ d(x,B) = \delta\}$, the boundaries are disjoint for different $\delta > 0$, and hence at most countably many of them can have positive $P-$measure. Therefore for some sequence $\delta_k \to 0$ the sets $B_k \equiv \{x : \ d(x,B) < \delta_k\}$ are $P-$continuity sets and $B_k \downarrow B$ if B is closed. It follows that

$$\limsup_n P_n(B) \le \limsup_n P_n(B_k) = P(B_k) \qquad \text{since} \quad P_n(B_k) \to P(B_k)$$

by (viii). By letting $k \to \infty$ this yields (vi).

Now we show that (vi) implies (i). Suppose that (vi) holds and fix $f \in C_b(M)$. Without loss of generality we can transform f so that $0 < f(x) \le 1$ for all $x \in M$. Fix $k \ge 1$ and define the closed sets

$$B_j \equiv \{x \in M : \ \frac{j}{k} \le f(x)\} \qquad \text{for} \quad j = 0, \ldots, k.$$

Then it follows that

$$\sum_{j=1}^k \frac{j-1}{k} P(B_{j-1} \cap B_j^c) \le \int f dP \le \sum_{j=1}^k \frac{j}{k} P(B_{j-1} \cap B_j^c).$$

Rewriting the sum on the right side and summing by parts gives

$$\sum_{j=1}^{k} \frac{j}{k} \{P(B_{j-1}) - P(B_j)\} = \frac{1}{k} + \frac{1}{k} \sum_{j=1}^{k} P(B_j)$$

which, together with a similar summation by parts on the left side yields

$$\frac{1}{k} \sum_{j=1}^{k} P(B_j) \le \int f dP \le \frac{1}{k} + \frac{1}{k} \sum_{j=1}^{k} P(B_j)\,.$$

Since the sets B_j are closed, it follows from the last display (also used with P replaced by P_n throughout) and (vi) that

$$\limsup_n \int f dP_n \le \limsup_n \left\{ \frac{1}{k} + \frac{1}{k} \sum_{j=1}^{k} P_n(B_j) \right\} \le \frac{1}{k} + \frac{1}{k} \sum_{j=1}^{k} P(B_j) \le \frac{1}{k} + \int f dP\,.$$

Letting $k \to \infty$ gives

$$\limsup_n \int f dP_n \le \int f dP\,.$$

Applying this last conclusion to $-f$ yields

$$\liminf_n \int f dP_n \ge \int f dP\,.$$

Combining these last two displays yields (i).

Since (ix) implies (viii) by taking $f = 1_B$, it remains only to show that (iv) (and (v) since (iv) and (v) are equivalent) implies (ix). Suppose that f is a bounded measurable function and suppose that (iv) holds; without loss of generality we may assume that $0 \le f \le 1$. Define the lower semicontinuous function $\overset{\circ}{f}$ and the upper semicontinuous function $\overline{f}$ by

$$\begin{aligned} \overset{\circ}{f} &\equiv \sup\{g:\ g \le f,\ g \ \ LSC\}\,, \\ \overline{f} &\equiv \inf\{g:\ g \ge f,\ g \ \ USC\}\,. \end{aligned}$$

Note that this notation is sensible: if we take $f = 1_B$ for a Borel set B, then

$$\overset{\circ}{(1_B)} = 1_{B^\circ}\,, \qquad \overline{(1_B)} = 1_{\overline{B}}\,.$$

Also note that

$$\overset{\circ}{f} \le f \le \overline{f}\,.$$

We claim that

$$E_f \equiv \{x : \overset{\circ}{f} = \overline{f}\} = \{x :\ f \ \text{ is continuous at } \ x\} \equiv C_f\,.$$

At any x for which $\overset{\circ}{f}(x) = f(x)$, the set $\{y : \overset{\circ}{f}(y) > f(x) - \epsilon\}$ is an open neighborhood of x, and on this neighborhood $f(y) > f(x) - \epsilon$. Similarly, if $\overline{f}(x) = f(x)$, there exists a neighborhood of x on which $f(y) < f(x) + \epsilon$. Putting these together shows that f is continuous at each point of

$\{x : \overline{f}(x) = \overset{\circ}{f}(x)\}$; i.e. $E_f \subset C_f$. To see the reverse inclusion, note that if f is continuous at x, the for each $\epsilon > 0$ there is an open set G for which $|f(y) - f(x)| < \epsilon$ for all $y \in G$. Then it follows that

$$(f(x) - \epsilon)1_G(y) - 21_{G^c}(y) \leq f(y) \leq (f(x) + \epsilon)1_G(y) + 21_{G^c}(y)$$

which differ by 2ϵ at x. Note that the upper bound in the last display is USC and the lower bound is LSC. This shows that $\overline{f}(x) - \overset{\circ}{f}(x) \leq \epsilon$ and hence that $\overline{f}(x) = \overset{\circ}{f}(x)$. This shows that $E_f \supset C_f$ and completes the proof of (a)

Now by (a) together with (iv) and (vi) we have (using the abbreviated notation $Pf \equiv \int f dP$)

$$P\overset{\circ}{f} \leq \liminf P_n \overset{\circ}{f} \leq \liminf P_n f \ \leq \limsup P_n f \ \leq \limsup_n P\overline{f} \ \leq P\overline{f}\,.$$

Since $P(C_f) = 1$ by hypothesis, it follows from (a) that $P(\overset{\circ}{f}) = P\overline{f} = Pf$. We thus conclude that (ix) holds. $\square$

The last part of the portmanteau Theorem, part (ix) has an important consequence: weak convergence is preserved under a map T to another metric space (M', d') which is continuous at a sufficiently large set of points with respect to the limit measure P. This is the Mann-Wald or continuous mapping theorem.

Theorem 1.2 (Continuous mapping) Suppose that T is a $\mathcal{M} \setminus \mathcal{M}'$ measurable mapping from (M, d) into another metric space (M', d') with Borel sigma-field $\mathcal{M}'$. Suppose that T is continuous at each point of a measurable subset $C_T \subset M$. If $P(C_T) = 1$, then $P_n^T \rightarrow_d P^T$; equivalently if $X_n \sim P_n$, $X \sim P$ are random elements in (M, d), then $T(X_n) \rightarrow_d T(X)$ in (M', d') provided $P(X \in C_T) = 1$.

Proof. Let $g \in C_b(M')$. Then

$$\int g dP_n^T = \int g(T) dP_n$$

where $g(T) = g \circ T : M \mapsto \mathbb{R}$ is bounded and continuous a.e. P since $P(C_T) = 1$. It therefore follows from (ix) of the portmanteau theorem that

$$\int g dP_n^T = \int g(T) dP_n \rightarrow \int g(T) dP = \int g dP^T\,.$$

$\square$

2 Weak convergence in $\mathbb{R}$ and $\mathbb{R}^k$

Weak convergence in $\mathbb{R}$

When the metric space M is $\mathbb{R}$, further equivalences can be added to those given in the portmanteau theorem, Theorem 1.1. In particular we can add smoothness restrictions to the functions f involved (that only make sense for functions defined on $\mathbb{R}$). The following proposition is one such result in this direction.

Proposition 2.1 Suppose that $\{X, X_n\}$, are real valued random variables, and suppose further that $Ef(X_n) \to Ef(X)$ for each $f \in C^\infty(\mathbb{R})$, the class of all bounded functions with bounded derivatives of all orders. Then $X_n \to_d X$.

Proof. Let $Z \sim N(0,1)$. For a fixed $f \in BL(\mathbb{R})$ and $\sigma > 0$, define a smoothed function f_σ by convolution:

$$f_\sigma(x) = Ef(x + \sigma Z) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x-y)^2\right) f(y)dy\,.$$

Note that $f_\sigma \in C^\infty(\mathbb{R})$ (since we can justify repeated integration via the dominated convergence theorem), and f_σ converges uniformly to f since

$$|f_\sigma(x) - f(x)| \le E|f(x+\sigma Z) - f(x)| \le \|f\|_{BL} E\{1 \wedge \sigma|Z|\} \to 0$$

as $\sigma \searrow 0$ by the dominated convergence theorem.

Suppose that $\epsilon > 0$ is given. Fix $\sigma > 0$ so that $\sup_x |f_\sigma(x) - f(x)| \le \epsilon$. Then

$$|Ef(X_n) - Ef(X)| \le |Ef_\sigma(X_n) - Ef_\sigma(X)| + 2\epsilon$$

so that

$$\limsup_n |Ef(X_n) - Ef(X)| \le 2\epsilon$$

since $f_\sigma \in C^\infty(\mathbb{R})$ and hence $Ef_\sigma(X_n) \to Ef_\sigma(X)$ by the hypothesis of the lemma. $\square$

Here is another proposition of this type giving further equivalences:

Proposition 2.2 Suppose that $\{X, X_n\}$ are real valued random variables. Then the following are equivalent:

(i) $F_n(x) = P(X_n \le x) \to P(X \le x) = F(x)$ for all x with $P(X = x) = 0$
(i.e. all $P-$continuity intervals of the form $(-\infty, x]$).
(ii) $X_n \to_d X$; i.e. $Ef(X_n) \to Ef(X)$ for all $f \in C_b(\mathbb{R})$.
(iii) $Ef(X_n) \to Ef(X)$ for all $f \in C^3(\mathbb{R})$.
(iv) $Ef(X_n) \to Ef(X)$ for all $f \in C^\infty(\mathbb{R})$.
(v) $E\exp(itX_n) \to E\exp(itX)$ for all $t \in \mathbb{R}$.

Proof. We have proved that (iv) implies (ii), and the reverse implication is trivially true. Since $C^\infty(\mathbb{R}) \subset C^3(\mathbb{R}) \subset C_b(\mathbb{R})$, the equivalences with (iii) follow easily.

For the equivalence of (i) and (ii) see Exercise xx.

The equivalence of (v) and (ii) will be established in Chapter 12. $\square$

On the real line $\mathbb{R}$ we can metrize weak convergence in terms of the distribution functions: the metric that does this is the *Lévy metric* λ.

Proposition 2.3 (Lévy metric) For any distribution functions F and G define

$$\lambda(F,G) \equiv \inf\{\epsilon > 0 : \ F(x-\epsilon) - \epsilon \le G(x) \le F(x+\epsilon) + \epsilon \ \text{ for all } \ x \in \mathbb{R}\}\,.$$

Then λ is a metric. Moreover, the set of all distribution functions under λ is a complete separable metric space. Also $F_n \to_d F$ as $n \to \infty$ if and only if $\lambda(F_n, F) \to 0$ as $n \to \infty$.

Proof. See Problem 6.5. □

Our goal now is to use part (iii) of Proposition 2.2 to prove several basic central limit theorems using the method of Lindeberg. The proofs will use the following "replacement inequality".

Proposition 2.4 (Lindeberg replacement inequality) Suppose that X and Y are independent random variables with $E|Y|^3 < \infty$, and suppose that W is another random variable independent of X with $E|W|^3 < \infty$. Suppose further that $EY = EW$ and $EY^2 = EW^2$. Then for $f \in C^3(\mathbb{R})$

$$|Ef(X+Y) - Ef(X+W)| \le C\left(E|Y|^3 + E|W|^3\right)$$

where $C = (1/6)\sup_x |f'''(x)|$. In particular when $W \sim N(\mu, \sigma^2)$, then

$$|Ef(X+Y) - Ef(X+W)| \le C_1 E|Y|^3$$

where $C_1 \equiv (5 + 4E|Z|^3)C \doteq (11.3831\ldots)C$ and $Z \sim N(0,1)$, and hence

$$E|Z|^3 = 2(2\pi)^{-1/2}\int_0^\infty z^3 e^{-z^2/2}dz = 4(2\pi)^{-1/2} \doteq 1.59577\ldots\,.$$

Proof. Fix $f \in C^3(\mathbb{R})$; by Taylor's theorem

$$f(x+y) = f(x) + yf'(x) + \frac{1}{2}y^2 f''(y) + R(x,y)$$

where $R(x,y) = y^3 f'''(x^*)/6$ for some x^* satisfying $|x^* - x| \le |y|$. Therefore it follows that

(a) $\quad |R(x,y)| \le C|y|^3 \qquad$ for all $x,\ y$.

Thus for any two random variables X and Y

$$Ef(X+Y) = Ef(X) + E(Yf'(X)) + \frac{1}{2}E(Y^2 f''(X)) + ER(X,Y)\,.$$

Using independence of X and Y and the bound (a) it follows that

$$|Ef(X+Y) - Ef(X) - E(Y)E(f'(X)) - \frac{1}{2}E(Y^2)E(f''(X))| \le CE|Y|^3\,.$$

Since the same inequality holds with Y replaced by W for another random variable W independent of X with $E|W|^3 < \infty$, if Y and W have $E(Y) = E(W)$ and $E(Y^2) = E(W^2)$, then we can subtract and via cancellation of the first and second moment terms conclude that

(b) $\quad |Ef(X+Y) - Ef(X+W)| \le C\left(E|Y|^3 + E|W|^3\right)\,.$

When $W \sim N(\mu, \sigma^2)$ we can further bound $E|W|^3$: since $Z \equiv (W-\mu)/\sigma \sim N(0,1)$ we can write $W = \mu + \sigma Z$. Then by the C_r-inequality (with $r = 3$)

$$\begin{aligned} E|W|^3 &\le 2^{3-1}\{|\mu|^3 + \sigma^3 E|Z|^3\} \\ &\le 4\{|E(Y)|^3 + \{E(Y^2)\}^{3/2} E|Z|^3\} \\ &\le 4\{E|Y|^3 + E|Y|^3 E|Z|^3\} = (4 + 4E|Z|^3)E|Y|^3 \end{aligned}$$

where the last inequality follows from Jensen's inequality used twice. Combining the last display with (b) yields the second inequality of the proposition. $\Box$

Now suppose that $\xi_1, \ldots, \xi_k$ are independent random variables with

$$\mu_i \equiv E\xi_i, \qquad \sigma_i^2 = Var(\xi_i), \qquad E|\xi_i|^3 < \infty\,.$$

Suppose that $\{\eta_i\}$ are independent and independent of the collection $\{\xi_i\}$ with $\eta_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, k$. Define

$$S_k = \xi_1 + \ldots + \xi_k\,, \qquad T_k = \eta_1 + \ldots + \eta_k\,.$$

Note that $T_k \sim N(E(T_k), Var(T_k)) = N(\sum_1^k \mu_j, \sum_1^k \sigma_j^2)$. Now we set up notation to apply Proposition 2.4: we define, for each i

$$\begin{aligned} X_i &\equiv \xi_1 + \ldots + \xi_{i-1} + \quad + \eta_{i+1} + \ldots + \eta_k\,, \\ Y_i &\equiv \xi_i \\ W_i &\equiv \eta_i\,. \end{aligned}$$

By independence of the $2k$ random variables $\{\xi_i\}$ and $\{\eta_i\}$ it follows that X_i, Y_i, and W_i are independent for each i. From the second bound of Proposition 2.4 it follows that

$$|Ef(X_i + Y_i) - Ef(X_i + W_i)| \le C_1 E|\xi_i|^3 \qquad 1 \le i \le k\,.$$

Also note that for $i = k$ the definitions yield $X_k + Y_k = S_k$ and $X_1 + W_1 = T_k$. Each replacement of a Y_i by a W_i gives sums $X_i + Y_i$ and $X_i + W_i$ with one more normal random variable η_i, and taken together the k replacements result in replacing all the non-Gaussian variables ξ_i by the Gaussian random variables η_i to get T_k. The total change in expected value is therefore bounded by a sum of third moment terms. Here are the details: since $X_j + W_j = X_{j-1} + Y_{j-1}$ for $j = 2, \ldots, k$,

$$\begin{aligned} |Ef(S_k) - Ef(T_k)| &= |Ef(X_k + Y_k) - Ef(X_1 + W_1)| \\ &= |\sum_{j=1}^k (Ef(X_j + Y_j) - Ef(X_j + W_j))| \\ &\le \sum_{j=1}^k |Ef(X_j + Y_j) - Ef(X_j + W_j)| \\ &\le C_1\left(E|\xi_1|^3 + \cdots + E|\xi_k|^3\right)\,. \qquad (1) \end{aligned}$$

We will state the resulting theorem in terms of a *triangular array* of row-wise independent random variables $\{\xi_{n,i} : i = 1, \ldots, k_n,\ n \in \mathbb{N}\}$ where $n \mapsto k_n$ is non-decreasing:

$$\xi_{1,1},\ \xi_{1,2}, \ldots, \xi_{1,k_1}$$

$$
\begin{array}{llll}
\xi_{2,1}, & \xi_{2,2}, & \ldots, \xi_{2,k_2} & \\
\xi_{3,1}, & \xi_{3,2}, & & \ldots, \xi_{3,k_3} \\
\cdot & & & \\
\cdot & & & \\
\cdot & & &
\end{array}
$$

We assume that the random variables in each row are independent, but nothing is assumed about relationships between different rows. As we will see, this formulation is convenient for dealing with centering and scaling constants.

Theorem 2.1 (Basic triangular array CLT) Suppose that $\{\xi_{n,i}: \ i = 1, \ldots, k_n\}_{n=1}^{\infty}$ is a triangular array of row-wise independent random variables such that:
(i) $\sum_1^{k_n} E\xi_{n,i} \to \mu$ where $\mu \in \mathbb{R}$ is finite.
(ii) $\sum_1^{k_n} Var(\xi_{n,i}) \to \sigma^2 < \infty$.
(iii) $\sum_1^{k_n} E|\xi_{n,i}|^3 \to 0$.
Then

$$
\sum_{i=1}^{k_n} \xi_{n,i} \to_d N(\mu, \sigma^2) \, .
$$

Proof. Fix $f \in C^3(\mathbb{R})$. Application of the inequality (1) yields

$$
|Ef(\sum_1^{k_n} \xi_{n,i}) - Ef(T_n)| \le C_1 \sum_1^{k_n} E|\xi_{n,i}|^3 \to 0
$$

where $T_n \sim N(\mu_n, \sigma_n^2)$ and where $\mu_n \to \mu$, $\sigma_n^2 \to \sigma^2$ by (i) and (ii). Since this implies that $T_n \to_d N(\mu, \sigma^2)$ (see Exercise 6.2), it follows that

$$
Ef(\sum_{i=1}^{k_n} \xi_{n,i}) \to Ef(N(\mu, \sigma^2)) = Ef(\mu + \sigma Z)
$$

where $Z \sim N(0, 1)$, and this implies (2) in view of Proposition 2.2. $\square$

The basic central limit theorem for triangular arrays, Theorem 2.1, can be extended to cover sums of independent random varibles without third moment hypotheses via truncation arguments. Our next result, the classical (Lindeberg) central limit theorem for independent identically distributed random variables with finite variances is a good example of the technique.

Theorem 2.2 (Classical CLT) Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables with $E(X_i) = 0$ and $E(X_i^2) = 1$. Then

$$
\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) = \sqrt{n}(\overline{X}_n - 0) \to_d Z \sim N(0, 1) \, .
$$

In fact, for $f \in C^3(\mathbb{R})$,

$$
|Ef(n^{1/2}\overline{X}_n) - Ef(Z)| \le C_1 E\left\{ X_1^2 \left(1 \wedge \frac{|X_1|}{\sqrt{n}} \right) \right\} + \|f\|_{BL}\{2 + 2E|Z|\}E\{|X_1|^2 1_{[|X_1| > \sqrt{n}]}\}
$$

where $C_1 \equiv (5 + 4E|Z|^3)C \doteq (11.3831\ldots)C$ and $C \equiv \sup_x |f'''(x)|/6$.

Corollary 1 (Berry-Esseen type bound) Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables with $E(X_i) = 0$, $E(X_i^2) = 1$, and $E|X_i|^3 < \infty$. Then, for $f \in C^3(\mathbb{R})$,

$$|Ef(n^{1/2}\overline{X}_n) - Ef(Z)| \le K_f \frac{E|X_1|^3}{\sqrt{n}}$$

where $K_f \equiv C_1 + 2\|f\|_{BL}(1 + E|Z|)$.

Proof. The argument proceeds by applying Theorem 2.1 to the truncated and rescaled variables

$$\xi_{n,i} = \frac{X_i 1_{[|X_i| \le \sqrt{n}]}}{\sqrt{n}}\,, \qquad i = 1, \ldots, n\,.$$

We compute

$$\text{(a)} \qquad \mu_n \equiv \sum_1^n E\xi_{n,i} = nE\xi_{n,1} = -nE\{X_1 1_{[|X_1| > \sqrt{n}]}\}/\sqrt{n}$$

since $E(X_1) = 0$, and this yields

$$|\mu_n| \le \sqrt{n} E\{|X_1| 1_{[|X_1| > \sqrt{n}]}\} \le E\{|X_1|^2 1_{[|X_1| > \sqrt{n}]}\} \to 0$$

by the dominated convergence theorem. For the sum of variances we have

$$\sigma_n^2 \equiv \sum_1^n Var(\xi_{n,i}) = E\{X_1^2 1_{[|X_1| \le \sqrt{n}]}\} - n(E\xi_{n,1})^2 \to 1$$

since $E\xi_{n,1} = \mu_n/n = o(1/n)$ and by using the dominated convergence theorem again. In fact, we can also conclude that

$$|\sigma_n^2 - 1| \le E\{X_1^2 1_{[|X_1| > \sqrt{n}]}\} + n(E\xi_{n,1})^2 \le 2E\{X_1^2 1_{[|X_1| > \sqrt{n}]}\}$$

by (a) and Jensen's inequality.

Finally the sum of third moments is controlled by

$$\sum_1^{k_n} E|\xi_{n,i}|^3 \le \frac{n}{n^{3/2}} E\{|X_1|^3 1_{[|X_1| \le \sqrt{n}]}\} \le E\left\{X_1^2 (1 \wedge \frac{|X_1|}{\sqrt{n}})\right\} \to 0$$

again by the dominated convergence theorem. In fact this argument shows that

$$|Ef(\sum_1^n \xi_{n,i}) - Ef(T_n)| \le C_1 E\left\{X_1^2 \left(1 \wedge \frac{|X_1|}{\sqrt{n}}\right)\right\}$$

To conclude the proof we need to show that for $f \in C^3(\mathbb{R})$

$$Ef(n^{1/2}\overline{X}_n) - Ef(\sum_1^n \xi_{n,i}) \to 0\,.$$

But since $C^3(\mathbb{R}) \subset BL(\mathbb{R})$ the inequality (1) yields

$$\begin{aligned}
&|Ef(n^{1/2}\overline{X}_n) - Ef(\sum_1^n \xi_{n,i})| \\
&\quad \le \|f\|_{BL} E\Big|\frac{1}{\sqrt{n}}\sum_1^n X_i - \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i 1_{[|X_i|\le\sqrt{n}]}\Big| \\
&\quad \le \|f\|_{BL}\frac{n}{\sqrt{n}} E\{|X_1| 1_{[|X_1|>\sqrt{n}]}\} \\
&\quad \le \|f\|_{BL} E\{|X_1|^2 1_{[|X_1|>\sqrt{n}]}\} \to 0\,.
\end{aligned}$$

This completes the proof of the first claim of the theorem. To finish the proof of the second claim, it remains to bound $Ef(T_n) - Ef(Z) = Ef(\mu_n + \sigma_n Z) - Ef(Z)$ where $T_n \sim N(\mu_n, \sigma_n^2)$ and $Z \sim N(0,1)$. Again, for $f \in C^3(\mathbb{R})$ the inequality (1) yields

$$\begin{aligned}
|Ef(\mu_n + \sigma_n Z) - Ef(Z)| &\le \|f\|_{BL} E|\mu_n + (\sigma_n - 1)Z| \\
&\le \|f\|_{BL}\{|\mu_n| + |\sigma_n - 1| E|Z|\} \\
&\le \|f\|_{BL}\{E\{|X_1|^2 1_{[|X_1|>\sqrt{n}]}\} + E|Z|\frac{1}{\sigma_n + 1}|\sigma_n^2 - 1|\} \\
&\le \|f\|_{BL}\{1 + 2E|Z|\} E\{|X_1|^2 1_{[|X_1|>\sqrt{n}]}\}
\end{aligned}$$

by (b). Collecting the bounds yields the second conclusion of the theorem. $\Box$

To prove the direct half of the classical Lindeberg-Feller central limit theorem, we will use the following lemma.

Lemma 2.1 Suppose that $\Delta_n(\epsilon) \to 0$ for each fixed $\epsilon > 0$. Then there exists a sequence $\epsilon_n \to 0$ such that $\Delta_n(\epsilon_n) \to 0$.

Proof. For each positive integer k there is an integer n_k such that $|\Delta_n(1/k)| < 1/k$ for $n \ge n_k$. We may assume, without loss of generality that $n_1 < n_2 < \ldots$. Set

$$\epsilon_n \equiv \begin{cases} 1/2 & \text{if } n < n_1 \\ 1/k & \text{if } n_k \le n < n_{k+1}\,. \end{cases}$$

Then for $n \ge n_1$ it follows that $\epsilon_n = 1/k_n$ where k_n satisfies $n_{k_n} \le n < n_{k_n+1}$. Note that $k_n \to \infty$ as $n \to \infty$, and for $n \ge n_1$ $|\Delta_n(\epsilon_n)| < 1/k_n \to 0$ as $n \to \infty$. $\Box$

Our next theorem gives the forward half of the Lindeberg-Feller central limit theorem.

Theorem 2.3 (Lindeberg-Feller) Suppose that $\{X_{n,i} : 1 \le i \le n;\ n \in \mathbb{N}\}$ is a triangular array of (row-wise independent) random variables with $E(X_{n,i}) = 0$ for all i and $n \in \mathbb{N}$ and $\sum_{i=1}^n E(X_{n,i}^2) = 1$. Then the following are equivalent:
(i) $\sum_1^n X_{n,i} \to_d Z \sim N(0,1)$ and $\max_{1\le i\le n} E(X_{n,i}^2) \to 0$;
(ii) $L_n(\epsilon) \equiv \sum_1^n E\{X_{n,i}^2 1_{[|X_{n,i}|>\epsilon]}\} \to 0$ for each $\epsilon > 0$.

Proof. Here we show that the Lindeberg condition (ii) implies (i). By (ii) it follows that $\Delta_n(\epsilon) \equiv L_n(\epsilon)/\epsilon^2 \to 0$ for each $\epsilon > 0$. By Lemma 2.1 we can find $\epsilon_n \to 0$ slowly enough that $\Delta_n(\epsilon_n) \to 0$. Now we truncate the $X_{n,i}$'s at ϵ_n: define a new triangular array $\{\xi_{n,i}\}$ by $\xi_{n,i} = X_{n,i}1_{[|X_{n,i}|\le \epsilon_n]}$. Note that

$$P(\xi_{n,i} \neq X_{n,i} \ \text{ for some } i) \le \sum_1^n P(|X_{n,i}| > \epsilon_n) \le L_n(\epsilon_n)/\epsilon_n^2 \to 0\,.$$

Thus it suffices to show that $\sum_1^n \xi_{n,i} \to_d Z$. To do this we use Theorem 2.1. Since the $X_{n,i}$ have mean zero,

$$\Big|\sum_1^n E(\xi_{n,i})\Big| = \Big| -\sum_1^n E\{X_{n,i}1_{[|X_{n,i}|>\epsilon_n]}\}\Big| \le L_n(\epsilon_n)/\epsilon_n = \epsilon_n L_n(\epsilon_n)/\epsilon_n^2 \to 0\,.$$

Furthermore,

$$\begin{aligned}\sum_1^n Var(\xi_{n,i}) &= \sum_1^n E\{X_{n,i}^2 1_{[|X_{n,i}|\le\epsilon_n]}\} - \sum_1^n (-E\{X_{n,i}1_{[|X_{n,i}|>\epsilon_n]}\})^2 \\ &= \sum_1^n E(X_{n,i}^2) - L_n(\epsilon_n) - o(1) = 1 - L_n(\epsilon_n) - o(1). \to 1\,.\end{aligned}$$

For the third moments we compute

$$\sum_1^n E|\xi_{n,i}|^3 \le \epsilon_n \sum_1^n E(X_{n,i}^2) \to 0\,.$$

Thus the hypotheses of Theorem 2.1 hold and we conclude that $\sum_1^n \xi_{n,i} \to_d Z$. To complete the proof that (ii) implies (i) we need to show that $\max_{1\le i\le n} E(X_{n,i}^2) \to 0$. But

$$\begin{aligned}E(X_{n,i}^2) &= E(X_{n,i}^2 1_{[|X_{n,i}|\le\epsilon_n]}) + E(X_{n,i}^2 1_{[|X_{n,i}|>\epsilon_n]}) \\ &\le \epsilon_n^2 + L_n(\epsilon_n)\,,\end{aligned}$$

and hence

$$\max_{1\le i\le n} E(X_{n,i}^2) \le \epsilon_n^2 + L_n(\epsilon_n) \to 0\,.$$

We will prove that (i) implies (ii) in Chapter 10 (PfS, lecture notes version; Chapter 13 PfS (2000)). □

A Converse CLT

Proposition 2.5 (Converse CLT) Suppose that $X_1, \ldots, X_n$ are i.i.d., and let $S_n \equiv n^{-1/2}\sum_{i=1}^n X_i$. If $S_n = O_p(1)$, then $E(X_1^2) < \infty$ and $E(X_1) = 0$.

Our proof of Proposition 2.5 will rely on the following three lemmas.

Lemma 2.2 (Symmetrization) For independent rv's $X_1, \ldots, X_n$ and $\epsilon_1, \ldots, \epsilon_n$ i.i.d. Rademacher rv's independent of the X_i's,

$$(2) \qquad P(|n^{-1/2}\sum_{i=1}^n \epsilon_i X_i| > 2t) \le 2\sup_n P(|n^{-1/2}\sum_{i=1}^n X_i| > t)\,.$$

Proof. By conditioning on the Rademacher's we see that

$$\begin{aligned}
P\left(n^{-1/2}|\sum_{i=1}^n \epsilon_i X_i| > 2t\right) &\le P\left(n^{-1/2}|\sum_{i:\epsilon_i=1} \epsilon_i X_i| + n^{-1/2}|\sum_{i:\epsilon_i=-1} \epsilon_i X_i| > 2t\right)\\
&\le E_\epsilon P_X\left(n^{-1/2}|\sum_{i:\epsilon_i=1} X_i| > t\right)\\
&\qquad + E_\epsilon P_X\left(n^{-1/2}|\sum_{i:\epsilon_i=-1} X_i| > t\right)\\
&\le 2\sup_{k\le n} P\left(n^{-1/2}|\sum_{i=1}^k X_i| > t\right)\\
&\le 2\sup_{k\le n} P\left(k^{-1/2}|\sum_{i=1}^k X_i| > t\right)\\
&\le 2\sup_{1\le k<\infty} P\left(k^{-1/2}|\sum_{i=1}^k X_i| > t\right)\,,
\end{aligned}$$

i.e. (2) holds. □

Lemma 2.3 (Khinchine's inequalities) There exist constants A_p, B_p, such that, for $a = (a_1, \ldots, a_n) \in R^n$, and $p \ge 1$,

$$A_p\left\{\sum_{i=1}^n a_i^2\right\}^{p/2} \le E|\sum_{i=1}^n a_i \epsilon|^p \le B_p\left\{\sum_{i=1}^n a_i^2\right\}^{p/2}\,.$$

Recall that we proved this for $p = 1$ and found that $A_1 = 1/\sqrt{3}$ and $B_1 = 1$ work.

Lemma 2.4 (Paley-Zygmund inequality) Suppose that Y is a non-negative random variable with mean EY and second moment $E(Y^2) = \|Y\|_2^2$. Then

$$(3) \qquad P(Y > t) \ge \left(\frac{(EY-t)^+}{\|Y\|_2}\right)^2\,.$$

Proof.

$$\begin{aligned}
E(Y) &= E(Y1_{[Y\le t]}) + E(Y1_{[Y>t]})\\
&\le t + \sqrt{E(Y^2)P(Y>t)}
\end{aligned}$$

by the Cauchy-Schwarz inequality. Rearranging this inequality yields (3). □

Proof. **(Proposition 2.5)** The following proof is from Giné and Zinn (1994). Lemma 2.2 yields

$$\sup_n P(|n^{-1/2}\sum_{i=1}^n \epsilon_i X_i| > 2t) \le 2\sup_n P(|n^{-1/2}\sum_{i=1}^n X_i| > t)\,.$$

Thus tightness of $\{S_n\}$ implies that

$$\{n^{-1/2}\sum_{i=1}^n \epsilon_i X_i\} \qquad \text{is tight}\,.$$

By Khinchine's inequality (Lemma 2.3), regarding the X_i's as fixed (conditioning on the X_i's), we find that

$$E_\epsilon\left|n^{-1/2}\sum_{i=1}^n \epsilon_i X_i\right| \ge A_1\left(n^{-1}\sum_{i=1}^n X_i^2\right)^{1/2} \equiv c[S_n]\,.$$

Thus by the Paley-Zygmund inequality (Lemma 2.4) applied with $Y = |n^{-1/2}\sum_{i=1}^n \epsilon_i X_i|$ and the X_i's held fixed (conditioning on the X_i's)

$$\begin{aligned} P_\epsilon(|n^{-1/2}\sum_{i=1}^n \epsilon_i X_i| > t) &\ge \left(\frac{(EY-t)^+}{(E(Y^2))^{1/2}}\right)^2 \\ &\ge \left(\frac{(c[S_n]-t)^+}{[S_n]}\right)^2 \\ &= c^2\left(1-\frac{t}{c[S_n]}\right)^2 \\ &\ge \frac{c^2}{4}1_{[[S_n]>2t/c]}\,. \end{aligned}$$

Taking expectations across this inequality with respect to the X_i's yields

$$P(|n^{-1/2}\sum_{i=1}^n \epsilon_i X_i| > t) \ge \frac{c^2}{4}P([S_n] > 2t/c)\,.$$

It follows that the sequence $\{[S_n]\}$ is tight. Now for fixed $M \in (0,\infty)$

$$\frac{1}{n}\sum_{i=1}^n X_i^2 1_{[X_i^2\le M]} \to_{a.s.} E(X_1^2 1_{[X_1^2\le M]}) \qquad \text{as} \ \ n\to\infty\,.$$

Thus in particular this convergence holds in probability and in distribution. Therefore, by the Portmanteau theorem 11.7.4 (f),

$$\begin{aligned} 1_{[E(X_1^2 1_{[X_1^2\le M]})>t]} &\le \liminf_{n\to\infty} P(\frac{1}{n}\sum_{i=1}^n X_i^2 1_{[X_i^2\le M]} > t) \\ &\le \sup_n P(\frac{1}{n}\sum_{i=1}^n X_i^2 1_{[X_i^2\le M]} > t)\,, \end{aligned}$$

so it follows that

$$\begin{aligned}\sup_{M>0} 1_{[E(X_1^2 1_{[X_1^2 \le M]})>t]} &\le \sup_{M>0}\sup_n P\left(\frac{1}{n}\sum_{i=1}^n X_i^2 1_{[X_i^2\le M]} > t\right)\\ &\le \sup_n P\left(\frac{1}{n}\sum_{i=1}^n X_i^2 > t\right)\\ &= \sup_n P([S_n]^2 > t)\,.\end{aligned}$$

By the tightness of $\{[S_n]\}$, we can make the right side of the last display as small as we please; in particular there exists a number $t_0 < \infty$ such that the right side is less than $1/2$. But this implies that for this t_0 the indicator on the left side of the inequality must be zero, uniformly in M; i.e.

$$\sup_{M>0} E(X_1^2 1_{[X_1^2\le M]}) \le t_0\,.$$

But the last supremum is just $E(X_1^2)$, and hence we have $E(X_1^2) \le t_0 < \infty$.

To complete the proof, note that $E(X_1^2) < \infty$ implies that $E|X_1| < \infty$, and hence by the strong law of large numbers we have

$$n^{-1}\sum_{i=1}^n X_i \to_{a.s.} E(X_1)\,.$$

But the hypothesis $n^{-1/2}\sum_{i=1}^n X_i = O_p(1)$ implies that

$$n^{-1}\sum_{i=1}^n X_i \to_p 0\,,$$

Combining these two displays yields $E(X_1) = 0$. $\square$

Giné and Zinn (1994) use similar methods to establish the corresponding theorem for U-statistics.

Theorem. (Giné and Zinn, 1994). If the sequence $\{n^{m/2}U_n(h)\}_{n=1}^\infty$ is tight (stochastically bounded), then $Eh^2(X_1,\ldots,X_m) < \infty$ and $Eh(X_1, x_2, \ldots, x_m) = 0$ for almost every $(x_2,\ldots,x_m) \in \mathcal{X}^{m-1}$.

Reference: Giné, E. and Zinn, J. (1994). A remark on convergence in distribution of U-statistics. *Ann. Probability* **22**, 117 - 125.

Weak convergence in $\mathbb{R}^k$

The next step is to extend the results for $M = \mathbb{R}$ to $M = \mathbb{R}^k$. We first state a set of equivalences for $\to_d$ in $\mathbb{R}^k$.

Proposition 2.6 Suppose that $\{X, X_n\}$ are random vectors with values in $\mathbb{R}^k$, and let $F_n(x) \equiv P(X_n \le x)$ and $F(x) \equiv P(X \le x)$ for $x \in \mathbb{R}^k$. Then the following are equivalent:
(i) $F_n(x) = P(X_n \le x) \to P(X \le x) = F(x)$ for all $x \in C_F \equiv \{y \in \mathbb{R}^k :\ F \text{ is continuous at } y\}$.
(ii) $X_n \to_d X$; i.e. $Ef(X_n) \to Ef(X)$ for all $f \in C_b(\mathbb{R}^k)$.
(iii) $Ef(X_n) \to Ef(X)$ for all $f \in C^\infty(\mathbb{R}^k)$.
(iv) $E\exp(it'X_n) \to E\exp(it'X)$ for all $t \in \mathbb{R}^k$.

In Proposition 2.6 the equivalence of (ii) and (iii) depends on the equivalence of (i) and (iii) in Theorem 1.1 and then a generalization of Proposition 2.1 to $\mathbb{R}^k$; see Exercise 6.6.

The replacement techniques of Lindeberg can be extended in a straightforward way to random vectors; see Exercises 6.7 and 6.7 for the start of this. One concrete result in this direction is the following central limit theorem for sums of independent random vectors.

Theorem 2.4 (Classical multivariate CLT) Suppose that $X_1, \ldots, X_n$ are i.i.d. random vectors in $\mathbb{R}^k$ with $E(X_1) = \mu$ and $E(|X_1|^2) < \infty$. Then

$$n^{-1/2}(X_1 + \cdots + X_n - n\mu) = \sqrt{n}(\overline{X}_n - \mu) \rightarrow_d Y \sim N_k(0, \Sigma)$$

where $\Sigma = E(X_1 X_1^T) = (Cov(X_{1j}, X_{1j'}))_{j,j'=1}^{\infty}$.

On the other hand, the usual approach to deriving limit theorems of this type is via the result of Cramér and Wold (1936) characterizing convergence in distribution of random vectors in terms of the convergence of linear combinations in $\mathbb{R}$.

Proposition 2.7 (Cramér - Wold device) Let X_n, X be random vectors in $\mathbb{R}^k$. Then $X_n \rightarrow_d X$ in $\mathbb{R}^k$ if and only if $a' X_n \rightarrow_d a' X$ in $\mathbb{R}$ for each $a \in \mathbb{R}^k$.

Proof. Suppose that $X_n \rightarrow_d X$ in $\mathbb{R}^k$ and let $a \in \mathbb{R}^k$. Then $g(x) = a'x$ is a continuous function on $\mathbb{R}$ and hence by the continuous mapping theorem $a' X_n = g(X_n) \rightarrow_d g(X) = a' X$.

To prove the reverse implication we use part (iv) of Proposition 2.6. Suppose that $a' X_n \rightarrow_d a' X$ for every $a \in \mathbb{R}^k$. Then by part (v) of Proposition 2.2 it follows that

$$E \exp(it(a' X_n)) \rightarrow E \exp(it(a' X))$$

for all $t \in \mathbb{R}$, and this holds for every $a \in \mathbb{R}^k$. In particular, when $t = 1$ we have

$$\varphi_{X_n}(a) = E \exp(ia' X_n) \rightarrow E \exp(ia' X) = \varphi_X(a)$$

for every $a \in \mathbb{R}^k$. But then by (iv) of Proposition 2.6 this implies that $X_n \rightarrow_d X$ in $\mathbb{R}^k$. $\square$

Walther (1997) gives a proof of the result of Cramér and Wold without use of characteristic functions, and notes that related results were established by Radon (1917).

3 Tightness and subsequences

It is often useful to argue using subsequences in arguments involving convergence in distribution. The following basic proposition gives a starting point for our discussion:

Proposition 3.1 If P_n and P are distributions (probability measures) on $(M, \mathcal{M})$ such that for every subsequence $\{P_{n'}\}$ with $\{n'\} \subset \mathbb{N}$ there is a further subsequence $\{P_{n''}\}$ such that $P_{n''} \rightarrow_d P$, then $P_n \rightarrow P$.

Proof. Suppose not. Then for some $f \in C_b(M)$ we have $P_n f \not\rightarrow Pf$. Thus for some $\epsilon > 0$ and subsequence n' it follows that $|P_{n'}f - Pf| > \epsilon$ for all $n' \in \{n'\}$. But then there is no further subsequence $\{n''\}$ for which $P_{n''}f \rightarrow Pf$, contradicting the hypothesis. $\Box$

To be able to extract convergent subsequences in general requires some appropriate notion of compactness. Here the right idea is to rule out "escape of mass". On the real line this "escape" is possible only toward $\pm\infty$, but in more complicated spaces it can happen in many ways. The following definitions are aimed at ruling out the "escape of mass" in quite general settings.

Definition 3.1 (Tightness) A probability measure P on $\mathcal{M}$ is said to be *tight* if for each $\epsilon > 0$ there exists a compact set $K = K_\epsilon$ such that $P(K_\epsilon) > 1 - \epsilon$.

The basic result concerning tightness of individual measures P is due to Ulam.

Theorem 3.1 (Ulam's theorem) If M is separable and complete, then each P on $(M, \mathcal{M})$ is tight.

Proof. Let $\epsilon > 0$. By the separability of M, for each $m \geq 1$ there is a sequence $A_{m1}, A_{m2}, \ldots$ of open $1/m$ spheres covering M. Choose i_m so that $P(\cup_{i \leq i_m} A_{mi}) > 1 - \epsilon/2^m$. Now the set $B \equiv \cap_{m=1}^{\infty} \cup_{i \leq i_m} A_{mi}$ is totally bounded in M: for each $\epsilon > 0$ it has a finite $\epsilon-$net (i.e. a set of points $\{x_k\}$ with $d(x, x_k) < \epsilon$ for some x_k for each $x \in B$). By completeness of M, $\overline{B}$ is complete and $\overline{B} \equiv K$ is compact. Since

$$P(K^c) = P(\overline{B}^c) \leq P(B^c) \leq \sum_{i=1}^{\infty} P\left\{(\cup_{i \leq i_m} A_{mi})^c\right\} < \sum_{m=1}^{\infty} \frac{\epsilon}{2^m} = \epsilon\,,$$

the conclusion follows. $\Box$

Definition 3.2 (Uniform tightness) If $\mathcal{P}$ is a set of probability measures on a metric space (M, d), then $\mathcal{P}$ is called *uniformly tight* if and only if for every $\epsilon > 0$ there is a compact set $K \subset M$ such that $P(K) > 1 - \epsilon$ for all $P \in \mathcal{P}$.

In the case of a sequence of measures $\{P_n\}$ it is convenient to relax the requirement in Definition 3.2 slightly.

Definition 3.3 (Asymptotic tightness (of a sequence)) If $\{P_n\}$ is a sequence of probability measures on (M, d), then $\{P_n\}$ is called *asymptotically tight* if and only if for every $\epsilon > 0$ there is a compact set $K = K_\epsilon$ such that $\limsup_n P_n(G^c) < \epsilon$ for every open set G containing K_ϵ

The main result for an asymptotically tight sequence is the following theorem due to Prohorov (1956) and Le Cam (1957).

Theorem 3.2 (Prohorov, 1956; Le Cam, 1957) Suppose that $\{P_n\}$ on $(M, \mathcal{M})$ is asymptotically tight. Then there exists a subsequence $\{P_{n'}\}$ that satisfies $P_{n'} \to_d$ (some) P where P is tight.

Pollard (2001) relaxes the definition of uniform tightness for a sequence still further, and proves the same result for arbitrary metric spaces.

The proof of the Prohorov - LeCam theorem 3.2 depends on the following auxiliary results. The first of these gives a correspondence between tight measures and tight linear functionals.

Theorem 3.3 (Correspondence theorem) A linear functional $T : BL(M)^+ \to \mathbb{R}^+$ with $T1 = 1$ defines a tight probability measure if and only if it is functionally tight: i.e. for each $\epsilon > 0$ there exists a compact set K_ϵ such that $T(l) < \epsilon$ for every $l \in BL(M)^+$ for which $l \le 1_{K_\epsilon^c}$.

Up to inconsequential constant multiples, asymptotic tightness is equivalent to: for each $\epsilon > 0$ there exists K_ϵ such that

$$\limsup_{n\to\infty} P_n l < 2\epsilon \qquad \text{for every} \;\; l \in BL(M)^+ \;\; \text{with} \;\; 0 \le l \le 1_{K_\epsilon^c}\,.$$

To see that asymptotic tightness implies this, note that for such a function l, the set $G_\epsilon = \{l < \epsilon\}$ is open and $G_\epsilon \supset K_\epsilon$. Then

$$P_n(l) \le \epsilon + P_n(G_\epsilon^c) < 2\epsilon$$

eventually.

The second analytic result we will use is:

Proposition 3.2 (Continuous partition of unity) For each $\delta > 0$, $\epsilon > 0$, and each compact set K, there exists a finite collection $\mathcal{G} = \{g_0, g_1, \ldots, g_k\} \subset BL(M)^+$ such that:
(i) $g_0(x) + g_1(x) + \cdots + g_k(x) = 1$ for each $x \in M$;
(ii) $\mathrm{diam}[g_i > 0] \le \delta$ for $i \ge 1$ where $\mathrm{diam}(A) \equiv \sup\{d(x,y) : \; x, y \in A\}$;
(iii) $g_0 < \epsilon$ on K.

Proof. Let $x_1, \ldots, x_k$ be the centers of open balls of radius $\delta/4$ whose union covers K. Define functions $f_0 \equiv \epsilon/2$, $f_i(x) = (1 - 2d(x, x_i)/\delta)^+$ for $i \ge 1$, so that $f_j \in BL(M)^+$ for $j = 0, \ldots, k$. Also note that $f_i(x) = 0$ if $d(x, x_i) > \delta/2$. Thus the set $\{f_i > 0\}$ has diameter less than δ for $i \ge 1$. The function $F(x) = \sum_{i=0}^k f_i(x)$ is everywhere greater than $\epsilon/2$ and is in $BL(M)^+$. The non-negative functions $g_i \equiv f_i/F$ are bounded by 1 and satisfy a Lipschitz condition:

$$\begin{aligned}
|g_i(x) - g_i(y)| &\le \frac{|F(y)f_i(x) - F(x)f_i(y)|}{F(x)F(y)} \\
&\le \frac{|f_i(x) - f_i(y)|}{F(x)} + \frac{|F(y) - F(x)|f_i(y)}{F(x)F(y)} \\
&\le \frac{\|f\|_{BL} d(x,y)}{\epsilon/2} + \frac{\|F\|_{BL} d(x,y)}{\epsilon/2}\,.
\end{aligned}$$

For each $x \in K$, there is an i for which $d(x, x_i) < \delta/4$. For this i, $f_i(x) > 1/2$ and $g_0(x) \le f_0(x)/f_i(x) < (\epsilon/2)/(1/2) = \epsilon$. Thus the functions g_i satisfy (i) - (iii). $\square$

Proof. **(Prohorov-LeCam theorem)**. Write K_i for the compact set corresponding to $\epsilon = 1/i$, $i \geq 1$. Write $\mathcal{G}_i$ for the finite collecton of functions in $BL(M)^+$ constructed in Proposition 3.2 with $\delta = \epsilon = 1/i$ and $K = K_i$. The collection $\mathcal{G} \equiv \cup_{i \in \mathbb{N}} \mathcal{G}_i$ is countable.

For each $g \in \mathcal{G}$ the sequence of real numbers $P_n g$ is bounded. It has a convergent subsequence. Via the Cantor-diagonalization argument we can construct a single sequence $\mathbb{N}_1 \subset \mathbb{N}$ for which $\lim_{n' \in \mathbb{N}_1} P_{n'} g$ exists for every $g \in \mathcal{G}$. The aproximation properties of $\mathcal{G}$ will allow us to show that $T(l) \equiv \lim_{n' \in \mathbb{N}_1} P_{n' \in \mathbb{N}_1} P_{n'} l$ exists for every $l \in BL(M)^+$. Without loss of generality, suppose that $\|l\|_{BL} \leq 1$. Given $\epsilon > 0$, choose an $i > 1/\epsilon$, then write $\mathcal{G}_i = \{g_0, g_1, \ldots, g_k\}$ for the finite collection guaranteed by Proposition 3.2. The open set $G_i = \{g_0 < \epsilon\}$ contains K_i which implies that $\limsup_{n \to \infty} P_n G_i^c < \epsilon$. For each $1 \leq j \leq k = k(i)$, let x_j be any point at which $g_j(x_j) > 0$. If x is any other point with $g_j(x) > 0$, then

$$|l(x) - l(x_j)| \leq d(x, x_j) \leq \epsilon\,.$$

It follows that for every $x \in M$

$$\begin{aligned} |l(x) - \sum_1^k l(x_j) g_j(x)| &\leq l(x) g_0(x) + \sum_{j=1}^k |l(x) - l(x_j)| g_j(x) \\ &\leq (\epsilon + 1_{G_i^c}) + \epsilon\,, \end{aligned}$$

and this integrates to give

$$|P_n l - \sum_{j=1}^k l(x_j) P_n(g_j)| \leq P_n G_i^c + 2\epsilon\,.$$

Since $\lim_{n' \in \mathbb{N}_1} P_{n'} g_j$ exists, it follows that

$$\limsup_{n' \in \mathbb{N}_1} P_{n'} l - \liminf_{n' \in \mathbb{N}_1} P_{n'} l \leq 6\epsilon\,.$$

This shows that $T(l) \equiv \lim_{n' \in \mathbb{N}_1} P_{n'} l$ exists for each $l \in BL(M)^+$.

Note that $T(1) = 1$ easily, and T inherits functional tightness from asymptotic tightness of $\{P_n\}$. From the correspondence Theorem 3.3 the functionally tight linear functional T corresponds to a tight probability measure P to which $\{P_{n'} :\ n' \in \mathbb{N}_1\}$ converges weakly. □

Definition 3.4 (Relative compactness) Let $\mathcal{P}$ be a set of probability measures on $(M, \mathcal{M})$. We say that $\mathcal{P}$ is *relatively compact* if every sequence $\{P_n\} \subset \mathcal{P}$ contains a weakly convergent subsequence. Thus every $\{P_n\} \subset \mathcal{P}$ contains a subsequence $\{P_{n'}\}$ with $P_{n'} \to_d$ some Q (not necessarily in $\mathcal{P}$).

Proposition 3.3 Let (M, d) be a separable metric space.
(i) (Le Cam). If $P_n \to_d P$, then $\{P_n\}$ is uniformly tight.
(ii) If $P_n \to_d P$, then $\{P_n\}$ is relatively compact.
(iii) If $\{P_n\}$ is relatively compact and the set of limit points is just the single point P, then $P_n \to_d P$.

Theorem 3.4 (Prohorov's theorem) Let $\mathcal{P}$ be a collection of probability measures on $(M, \mathcal{M})$.
(i) If $\mathcal{P}$ is uniformly tight, then it is relatively compact.
(ii) Suppose that (M, d) is separable and complete. If $\mathcal{P}$ is relatively compact it is uniformly tight.

4 Metrizing weak convergence

The Lévy metric on distribution functions defined in Proposition 2.3 extends in a nice way to give a metric for $\to_d$ more generally. For any set $B \in \mathcal{M}$ and $\epsilon > 0$ define

$$B^\epsilon \equiv \{y \in M : \ d(x,y) < \epsilon \ \text{ for some } x \in B\}.$$

Definition 4.1 (Prohorov metric) For P, Q two probability measures on $(M, \mathcal{M})$, the Prohorov distance $\rho(P,Q)$ between P and Q is defined by

$$\rho(P,Q) \equiv \inf\{\epsilon > 0 : P(B) \le Q(B^\epsilon) + \epsilon \ \text{ for all } \ B \in \mathcal{M}\}.$$

Another very useful metric on $\mathcal{P}$ is defined in terms of the bounded Lipschitz functions $BL(M)$ defined in Section 1.

Definition 4.2 (Bounded Lipschitz metric) For P, Q two probability measures on $(M, \mathcal{M})$, the *bounded Lipschitz distance* $\beta(P,Q)$ between P and Q is defined by

$$\beta(P,Q) \equiv \sup\left\{\left|\int f dP - \int f dQ\right| : \ \|f\|_{BL} \le 1\right\}.$$

Proposition 4.1 Both ρ and β are metrics on $\mathcal{P} \equiv \{\text{all probability measures on } (M,\mathcal{M})\}$.

Proof. See Exercise 6.10. $\square$

The following theorem says that both ρ and β metrize $\to_d$ just as the Lévy metric metrized convergence of distribution functions on $\mathbb{R}$.

Theorem 4.1 For any separable metric space (M,d) and Borel probability measures $\{P_n\}$, P on $(M, \mathcal{M})$ the following are equivalent:
(i) $P_n \to_d P$.
(ii) $\int f dP_n \to \int f dP$ for all $f \in BL(M)$.
(iii) $\beta(P_n, P) \to 0$.
(iv) $\rho(P_n, P) \to 0$.

Proof. We prove the result under the additional assumption that M is complete. The equivalence of (i) and (ii) has been proved in Theorem 1.1. Now we show that (ii) implies (iii): by Ulam's Theorem 3.1, for any $\epsilon > 0$ we can choose K compact so that $P(K) > 1 - \epsilon$. Now the set of functions $\mathcal{E} = \{f \in BL(M) : \ \|f\|_{BL} \le 1\}$ restricted to K form a compact set of functions for $\|\cdot\|_\infty$ (by the Arzela-Ascoli theorem; see e.g. Billingsley (1968) page 221). Thus for some finite k there are $f_1, \ldots, f_k \in BL(M)$ such for any $f \in \mathcal{E}$ there is an f_j with $\sup_{x\in K} |f(x) - f_j(x)| \le \epsilon$. Then, since f, $f_j \in BL(M)$,

$$\sup_{x \in K^\epsilon} |f(x) - f_j(x)| \le 3\epsilon.$$

Let $g(x) \equiv \max\{0, (1 - d(x,K)/\epsilon)\}$; then $g \in BL(M)$ and $1_K \le g \le 1_{K^\epsilon}$. For n sufficiently large we have

$$P_n(K^\epsilon) \ge \int g dP_n > 1 - 2\epsilon,$$

and hence for any $f \in \mathcal{E}$

$$
\begin{aligned}
\Big|\int f dP_n - \int f dP\Big| &= \Big|\int (f - f_j) d(P_n - P) + \int f_j d(P_n - P)\Big| \\
&\le \Big|\int (f - f_j) dP_n\Big| + \Big|\int (f - f_j) dP\Big| + \Big|\int f_j d(P_n - P)\Big| \\
&\le 3\epsilon + 2 \cdot 2\epsilon + 2\epsilon + 2\epsilon + + \Big|\int f_j d(P_n - P)\Big| \\
&\le 7\epsilon + 4\epsilon + \epsilon = 12\epsilon
\end{aligned}
$$

by choosing n large. Hence (iii) holds.

Now we show that (iii) implies (iv): given a Borel set B and $\epsilon > 0$, let $f_\epsilon(x) \equiv \max\{0, (1 - d(x, B)/\epsilon)\}$. Then $f_\epsilon \in BL(M)$, $\|f\|_{BL} \le 2 \vee \epsilon^{-1}$, and $1 < f_\epsilon \le 1_{B^\epsilon}$. Therefore, for any P and Q on M we have

$$
\begin{aligned}
Q(B) &\le \int f_\epsilon dQ \le \int f_\epsilon dP + (2 \vee \epsilon^{-1})\beta(P, Q) \\
&\le P(B^\epsilon) + (2 \vee \epsilon^{-1})\beta(P, Q)\,,
\end{aligned}
$$

and it follows that

$$\rho(P, Q) \le \max\{\epsilon, (2 \vee \epsilon^{-1})\beta(P, Q)\}\,.$$

Hence if $\beta(P, Q) \le \epsilon^2$, then

$$\rho(P, Q) < \max\{\epsilon, (2 \vee \epsilon^{-1})\epsilon^2\} = \max\{2\epsilon^2, \epsilon\} \le \epsilon(1 + 2\epsilon) \le 3\epsilon\,.$$

Hence for all P, Q we have $\rho(P, Q) \le 3\sqrt{\beta(P, Q)}$. Thus (iii) implies (iv). [It can also be shown that $2^{-1}\beta(P, Q) \le \rho(P, Q)$; see e.g. Dudley (1976), RAP, Corollary 11.6.5, page 411.]

Finally we show that (iv) implies (i): Suppose that (iv) holds, let B be a $P-$continuity set, and let $\epsilon > 0$. Then for $0 < \delta < \epsilon$ small, $P(B^\delta \setminus B) < \epsilon$ and $P((B^c)^\delta \setminus B^c) < \epsilon$. Then

$$P_n(B) \le P(B^\delta + \delta \le P(B) + 2\epsilon$$

and

$$P_n(B^c) \le P(((B^c)^\delta + \delta \le P(B^c) + 2\epsilon\,;$$

combining these yields

$$|P_n(B) - P(B)| \le 2\epsilon$$

and hence $P_n(B) \to P(B)$. By the portmanteau theorem 11.1.1 this yields (i). $\Box$

More Metrics on $\mathcal{P}$

There are other useful metrics on $\mathcal{P}$ that metrize topologies other than weak convergence. It is frequently useful to relate these to the Prohorov and bounded Lipschitz metrics ρ and β we have introduced earlier in this section.

Definition 4.3 For probability measures P, Q on $(M, \mathcal{M})$, the *total variation* distance from P to Q is defined by

$$d_{TV}(P, Q) \equiv \sup\{|P(A) - Q(A)| : \ A \in \mathcal{M}\}\,.$$

Proposition 4.2 The *total variation* distance $d_{TV}(P,Q)$ is given by

$$d_{TV}(P,Q) = \frac{1}{2}\int |p-q| d\mu = 1 - \int (p \wedge q)\, d\mu$$

where $p = dP/d\mu$, $q = dQ/d\mu$, and μ is any measure dominating both P and Q (e.g. $P+Q$).

Proof. See Exercise 6.11. □

Definition 4.4 The *Hellinger* distance $H(P,Q)$ is defined by

$$H^2(P,Q) \equiv \frac{1}{2}\int \{\sqrt{p} - \sqrt{q}\}^2\, d\mu = 1 - \int \sqrt{pq} d\mu\,,$$

where $p = dP/d\mu$, $q = dQ/d\mu$, and μ is any measure dominating both P and Q.

It is not hard to show (see Exercise 6.12) that $H(P,Q)$ does not depend on the choice of the dominating measure μ.

Here is a theorem relating these metrics to each other and to the Prohorov and bounded Lipschitz metrics.

Theorem 4.2 For P,Q probability measures on $(M,\mathcal{M})$ with (M,d) separable, the following inequalities hold:
(i) $2^{-1}\beta(P,Q) \le \rho(P,Q) \le 3\sqrt{\beta(P,Q)}$.
(ii) $H^2(P,Q) \le d_{TV}(P,Q) \le H(P,Q)\{1 - H^2(P,Q)/2\}^{1/2}$.
(iii) $\rho(P,Q) \le d_{TV}(P,Q)$.
For distribution functions F,G on $\mathbb{R}$ (or on $\mathbb{R}^k$) we have:
(iv) $\lambda(F,G) \le \rho(F,G) \le d_{TV}(F,G)$.
(v) $\lambda(F,G) \le d_K(F,G) \le d_{TV}(F,G)$
where $d_K(F,G) \equiv \|F-G\|_\infty \equiv \sup_x |F(x) - G(x)|$.

Proof. The right side of (i) was proved in the course of the proof of Theorem 4.1. For the left side, see Dudley (1976) section 18.6. We leave the remaining inequalities as exercises. □

Wasserstein metrics

These metrics, often denoted by $W_p(P,Q)$ or $W_r(P,Q)$ are also called Kantorovich distances, or Monge-Kantorovich distances, or the Mallows metric, or Wasserstein transport distances. To define these metrics, suppose that (M,d) is a separable metric space, and let $\mathcal{P}(\mathcal{M},d)$ be the collection of all Borel probability measures on (M,d). For $r \ge 1$ let $\mathcal{P}_r(M,d) \equiv \mathcal{P}_r(M)$ be the collection of all probability measures $P \in \mathcal{P}(\mathcal{M},d)$ such that

$$\int_M d(x,x_0)^r dP(x) < \infty \quad \text{for some } x_0 \in M$$

(and then equivalently for all $x_0 \in M$). For $P,Q \in \mathcal{P}_r(M,d)$ define $W_r(P,Q)$ by

$$W_r^r(P,Q) \equiv \inf\left\{\int_M\int_M d(x,y)^r d\pi(x,y) :\ \pi \text{ on } (M\times M, \mathcal{M}\times\mathcal{M}) \text{ has margins } P \text{ and } Q\right\}.$$

Thus $\pi(A\times M) = P(A)$ for all $A \in \mathcal{M}$ and $\pi(M\times B) = Q(B)$ for all $B \in \mathcal{M}$. The most important values of r are 1, 2, and ∞.

Here are statements of several important results concerning the Wasserstein metrics. The first result shows that W_1 is closely related to the bounded Lipschitz metric β.

Theorem 4.3 (Kantorivich duality theorem) If (M, d) is separable, then for all $P, Q \in \mathcal{P}_1(M)$

$$W_1(P, Q) = \sup_{f: \|f\|_L \le 1} \left| \int_M f dP - \int_M f dQ \right|$$

where $\|f\|_L = \sup_{x \ne y} |f(x) - f(y)|/d(x, y)$.

The second result characterizes convergence in the W_r metrics.

Theorem 4.4 (Convergence in W_r) Let $1 \le r < \infty$. Suppose that $P \in \mathcal{P}_r(M, d)$ and $\{P_n\}_{n \ge 1} \subset \mathcal{P}_r(M, d)$ where (M, d) is a complete separable metric space (or, slightly more generally, M is Polish). Then the following are equivalent:
(i) $W_r(P_n, P) \to 0$ as $n \to \infty$.
(ii) $P_n \to_d P$ and for some $x_0 \in M$

$$\int_M d(x, x_0)^r dP_n(x) \to \int_M d(x, x_0)^r dP(x).$$

(iii) For any $f : M \to \mathbb{R}$ satisfying, for some $x_0 \in M$,

$$|f(x)| \le C(1 + d(x, x_0))^r, \text{ for all } x \in M,$$

$\int_M f dP_n \to \int_M f dP$.

Thus $W_r(P_n, P) \to 0$ is stronger than $P_n \to_d P$.
The next result relates $W_\infty(P, Q)$ to $\rho(P, Q)$.

Theorem 4.5 (W_r metrics dominate ρ) For all $P, Q \in \mathcal{P}_r(M, d)$ and $r \ge 1$,

$$\rho(P, Q) \le W_r(P, Q)^{r/(r+1)}.$$

In particular,

$$\begin{aligned} \rho(P, Q) \le W_\infty(P, Q) &\equiv \lim_{r \to \infty} W_r(P, Q) \ \sup_{r \ge 1} W_r(P, Q) \\ &= \inf_\pi \|d\|_{L^\infty(\pi)} = \inf_\pi \operatorname{esssup}_\pi d(x, y). \end{aligned}$$

Another way to state this connection is in terms of the Ky Fan metric α for convergence in probability.

Definition 4.5 For a separable metric space (M, d), a probability space $(\Omega, \mathcal{A}, P)$ and $X, Y \in \mathcal{L}^0(\Omega, M)$, let

$$\alpha(X, Y) \equiv \inf\{\epsilon > 0 : \ P(d(X, Y) > \epsilon) < \epsilon\}.$$

This is called the *Ky Fan metric* for convergence in probability; see Dudley, RAP, section 9.2.

Theorem 4.6 (α metrizes convergence in probability) On $L^0(\Omega, M)$, α is a metric for convergence in probability. That is, $\alpha(X_n, X) \to 0$ if and only if $X_n \to X$ in probability (or equivalently, $d(X_n, X) \to_p 0$).

Proposition 4.3 (Connecting ρ and α) For any separable metric space, laws P and Q on M and $\epsilon > 0$ (or $\epsilon = 0$ if P and Q are tight), there is a probability space $(\Omega, \mathcal{A}, \mu)$ and random variables X, Y on Ω with

$$\alpha(X,Y) \le \rho(P,Q) + \epsilon \quad \text{where} \quad X \sim P, \quad Y \sim Q.$$

Thus

$$\rho(P,Q) = \inf\{\alpha(X,Y) : \; X \sim P, \; Y \sim Q\}.$$

When the metric space (M,d) is $(\mathbb{R}, |\cdot|)$, the W_r metrics take a special form:

Theorem 4.7 (Prohorov) Let P, Q be probability measures in $\mathcal{P}_r(\mathbb{R})$, $r \ge 1$, with respective distribution functions F and G. Then

$$W_r^r(P,Q) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^r dt.$$

In particular,

$$W_1(P,Q) = \int_0^1 |F^{-1} - G^{-1}| dt = \int_{-\infty}^{\infty} |F(x) - G(x)| dx.$$

To understand the convergence properties of empirical measures with respect to the W_r-metrics we first need a theorem of Varadarajan (1958).

Theorem 4.8 (Varadarajan) Suppose that (M,d) is separable and suppose that $X_1, \ldots, X_n$ are i.i.d. P on $(M, \mathcal{M})$. Then $Pr(\mathbb{P}_n \to_d P) = 1$. Equivalently, $\rho(\mathbb{P}_n, P) \to_{a.s.} 0$ and $\beta(\mathbb{P}_n, P) \to_{a.s.} 0$.

This leads naturally to the following theorem:

Theorem 4.9 (Glivenko-Cantelli theorem with respect to W_r) Suppose that (M,d) is separable and suppose that $X_1, \ldots, X_n$ are i.i.d. P on $\mathcal{P}_r(M)$ with $r \ge 1$. Then

$$W_r(\mathbb{P}_n, P) \to_{a.s.} 0 \quad \text{as} \quad n \to \infty.$$

Proof. By Varadarajan's theorem and the characterization of $W_r(P_n, P) \to 0$ given above, it suffices to show that

$$\int d(x, x_0)^r d\mathbb{P}_n(x) = \frac{1}{n} \sum_{i=1}^n d(X_i, x_0)^r \to_{a.s.} \int_M d(x, x_0)^r dP(x) = E\{d(X, x_0)^r\}.$$

But this last convergence holds by the strong law of large numbers. $\square$

Some references:

- Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196 - 1217.
- Bobkov, S. and Ledoux, M. (2017). *One dimensional empirical measures, order statistics, and Kantorovich transport distances.* Memoirs of the American Mathematical Society, AMS Providence, to appear.
- Dudley, R. M. (2002). *Real Analysis and Probability*, 2nd edition. Cambridge University Press.
- Dümbgen, L., Samworth, R., and Schuhmacher, D. (2011). Approximation by log-concave distributions with application to regression. *Ann. Statist.* **39**, 702-730.
- Mallows, C. (1972). A note on joint asymptotic normality. *Ann. Math. Statist.* **43**, 508 - 515.
- Villani, C. (2003). *Topics in Optimal Transportation.* American Mathematical Society Graduate Studies in Mathematics **58**, AMS, Providence.
- Villani, C. (2009). *Optimal Transport*. Springer, Berlin.

5 Characterizing weak convergence in spaces of functions

Suppose that T is a set, and suppose that $X_n(t)$, $t \in T$ are *stochastic processes* indexed by the set T; that is, $X_n(t) : \Omega \mapsto \mathbb{R}$ is a measurable map from each $t \in T$ and $n \in \mathbb{N}$. Assume that the processes X_n have bounded sample functions almost surely (or, have versions with bounded sample paths almost surely). Then $X_n(\cdot) \in \ell^\infty(T)$ almost surely where $\ell^\infty(T)$ is the space of all bounded real-valued functions on T. The space $\ell^\infty(T)$ with the sup norm $\|\cdot\|_T$ is a Banach space; it is separable only if T is finite. Hence we will *not* assume that the processes X_n induce tight Borel probability laws on $\ell^\infty(T)$.

Now suppose that $X(t)$, $t \in T$, is a sample bounded process that *does* induce a tight Borel probability measure on $\ell^\infty(T)$. then we say that X_n *converges weakly* to X (or, informally X_n converges in law to X uniformly in $t \in T$), and write

$$X_n \Rightarrow X \qquad \text{in} \qquad \ell^\infty(T)$$

if

$$E^* H(X_n) \to EH(X)$$

for all bounded continuous functions $H : \ell^\infty(T) \mapsto \mathbb{R}$. Here E^* denotes outer expectation.

It follows immediately from the preceding definition that weak convergence is preserved by continuous functions: if $g : \ell^\infty(T) \mapsto \mathbb{D}$ for some metric space $(\mathbb{D}, d)$ where g is continuous and $X_n \Rightarrow X$ in $\ell^\infty(T)$, then $g(X_n) \Rightarrow g(X)$ in $(\mathbb{D}, d)$. (The condition of continuity of g can be relaxed slightly; see e.g. Van der Vaart and Wellner (1996), Theorem 1.3.6, page 20.) While this is not a deep result, it is one of the reasons that the concept of weak convergence is important.

The following example shows why the outer expectation in the definition of $\Rightarrow$ is necessary.

Example 5.1 Suppose that U is a Uniform$(0,1)$ random variable, and let $X(t) = 1\{U \le t\} = 1_{[0,t]}(U)$ for $t \in T = [0,1]$. If we assume the axiom of choice, then there exists a nonmeasurable subset A of $[0,1]$. For this subset A, define $F_A = \{1_{[0,\cdot]}(s) : \; s \in A\} \subset \ell^\infty(T)$. Since F_A is a discrete set for the sup norm, it is closed in $\ell^\infty(T)$. But $\{X \in F_A\} = \{U \in A\}$ is not measurable, and therefore the law of X does not extend to a Borel probability measure on $\ell^\infty(T)$.

On the other hand, the following proposition gives a description of the sample bounded processes X that do induce a tight Borel measure on $\ell^\infty(T)$.

Proposition 5.1 (de la Peña and Giné (1999), Lemma 5.1.1; van der Vaart and Wellner (1996), Lemma 1.5.9)). Let $X(t)$, $t \in T$ be a sample bounded stochastic process. Then the finite-dimensional distributions of X are those of a tight Borel probability measure on $\ell^\infty(T)$ if and only if there exists a pseudometric ρ on T for which (T, ρ) is totally bounded and such that X has a version with almost all its sample paths uniformly continuous for ρ.

Proof. Suppose that the induced probability measure of X on $\ell^\infty(T)$ is a tight Borel measure P_X. Let K_m, $m \in \mathbb{N}$ be an increasing sequence of compact sets in $\ell^\infty(T)$ such that $P_X(\cup_{m=1}^\infty K_m) = 1$, and let $K = \cup_{m=1}^\infty K_m$. Then we will show that the pseudometric ρ on T defined by

$$\rho(s,t) = \sum_{m=1}^\infty 2^{-m}(1 \wedge \rho_m(s,t)),$$

where

$$\rho_m(s,t) = \sup\{|x(s) - x(t)| : x \in K_m\},$$

makes (T,ρ) totally bounded. To show this, let $\epsilon > 0$, and choose k so that $\sum_{m=k+1}^{\infty} 2^{-m} < \epsilon/4$ and let $x_1, \ldots, x_r$ be a finite subset of $\cup_{m=1}^{k} K_m = K_k$ that is $\epsilon/4-$dense in K_k for the supremum norm; i.e. for each $x \in \cup_{m=1}^{k} K_m$ there is an integer $i \le r$ such that $\|x - x_i\|_T \le \epsilon/4$. Such a finite set exists by compactness. The subset A of $\mathbb{R}^r$ defined by $\{(x_1(t), \ldots, x_r(t)) : t \in T\}$ is bounded (note that $\cup_{m=1}^{k} K_m$ is compact and hence bounded). Therefore A is totally bounded and hence there exists a finite set $T_\epsilon = \{t_j : 1 \le j \le N\}$ such that, for each $t \in T$, there is a $j \le N$ for which $\max_{1\le s\le r} |x_s(t) - x_s(t_j)| \le \epsilon/4$. It is easily seen that T_ϵ is $\epsilon-$dense in T for the pseudo-metric ρ: if t and t_j are as above, then for $m \le k$ it follows that

$$\rho_m(t,t_j) = \sup_{x\in K_m} |x(t) - x(t_j)| \le \max_{s\le r} |x_s(t) - x_s(t_j)| + \frac{\epsilon}{2} \le \frac{3\epsilon}{4},$$

and hence

$$\rho(t,t_j) \le \frac{\epsilon}{4} + \sum_{m=1}^{k} 2^{-m} \rho_m(t,t_j) \le \epsilon.$$

Thus we have proved that (T,ρ) is totally bounded. Furthermore, the functions $x \in K$ are uniformly $\rho-$continuous, since, if $x \in K_m$, then $|x(s) - x(t)| \le \rho_m(s,t) \le 2^m \rho(s,t)$ for all $s,t \in T$ with $\rho(s,t) \le 1$. Since $P_X(K) = 1$, the identity function of $(\ell^\infty(T), \mathcal{B}, P_X)$ yields a version of X with almost all of its sample paths in K, hence in $C_u(T,\rho)$, the space of bounded uniformly $\rho-$continuous functions on T. This proves the direct half of the proposition.

Conversely, suppose that $X(t)$, $t \in T$, is a stochastic process with a version whose sample functions are almost all in $C_u(T,\rho)$ for a metric or pseudometric ρ on T for which (T,ρ) is totally bounded. We will continue to use X to denote the version with these properties. We can clearly assume that all the sample functions are uniformly continuous. If $(\Omega, \mathcal{A}, P)$ is the probability space where X is defined, then the map $X : \Omega \mapsto C_u(T,\rho)$ is Borel measurable because the random vectors $(X(t_1), \ldots, X(t_k))$, $t_i \in T$, $k \in \mathbb{N}$, are measurable and the Borel $\sigma-$ algebra of $C_u(T,\rho)$ is generated by the "finite-dimensional sets" $\{x \in C_u(T,\rho) : (x(t_1), \ldots, x(t_k)) \in A\}$ for all Borel sets A of $\mathbb{R}^k$, $t_i \in T$, $k \in \mathbb{N}$. Therefore the induced probability law P_X of X is a tight Borel measure on $C_u(T,\rho)$ by Ulam's theorem; see e.g. Billingsley (1968), Theorem 1.4 page 10, or Dudley (1989), Theorem 7.1.4 page 176. But the inclusion of $C_u(T,\rho)$ into $\ell^\infty(T)$ is continuous, so P_X is also a tight Borel measure on $\ell^\infty(T)$. $\square$

Exhibiting convenient metrics ρ for which total boundedness and continuity holds is more involved. It can be shown that (see e.g. Hoffmann-Jørgensen (1984), (1991); Andersen (1985), Andersen and Dobric (1987)) that if any pseudometric works, then the pseudometric

$$\rho_0(s,t) = E \arctan |X(s) - X(t)|$$

will do the job. However, ρ_0 may not be the most natural or convenient pseudometric for a particular problem. In particular, for the frequent situation in which the process X is Gaussian, the pseudometrics ρ_r defined by

$$\rho_r(s,t) = (E|X(s) - X(t)|^r)^{1/(r\vee 1)}$$

for $0 < r < \infty$ are often more convenient, and especially ρ_2 in the Gaussian case; see Van der Vaart and Wellner (1996), Lemma 1.5.9, and the following discussion.

Proposition 5.1 motivates our next result which characterizes weak convergence $X_n \Rightarrow X$ in terms of *asymptotic equicontinuity* and convergence of finite-dimensional distributions.

Theorem 5.1 The following are equivalent:
(i) All the finite-dimensional distributions of the sample bounded processes X_n converge in law, and there exists a pseudometric ρ on T such that both:
(a) (T, ρ) is totally bounded, and
(b) the processes X_n are asymptotically equicontinuous in probability with respect to ρ: that is

$$\lim_{\delta\to 0} \limsup_{n\to\infty} Pr^* \left\{ \sup_{\rho(s,t)\le\delta} |X_n(s) - X_n(t)| > \epsilon \right\} = 0 \qquad \text{for all} \quad \epsilon > 0\,. \tag{1}$$

(ii) There exists a process X with tight Borel probability distribution on $\ell^\infty(T)$ and such that

$$X_n \Rightarrow X \qquad \text{in} \qquad \ell^\infty(T)\,.$$

If (i) holds, then the process X in (ii) (which is completely determined by the limiting finite-dimensional distributions of $\{X_n\}$), has a version with sample paths in $C_u(T, \rho)$, the space of all $\rho-$uniformly continuous real-valued functions on T. If X in (ii) has sample functions in $C_u(T, \gamma)$ for some pseudometric γ for which (T, γ) is totally bounded, then (i) holds with the pseudometric ρ taken to be γ.

Proof. Suppose that (i) holds. Let T_∞ be a countable $\rho-$dense subset of T, and let T_k, $k \in \mathbb{N}$, be finite subsets of T satisfying $T_k \nearrow T_\infty$. (Such sets exist by virtue of the hypothesis that (T, ρ) is totally bounded.) The limiting distributions of the processes X_n are consistent, and thus define a stochastic process X on T. Furthermore, by the portmanteau theorem for finite-dimensional convergence in distribution,

$$\begin{aligned}
&Pr\{\max_{\rho(s,t)\le\delta,\, s,t\in T_k} |X(s) - X(t)| > \epsilon\} \\
&\quad\le \liminf_{n\to\infty} Pr\{\max_{\rho(s,t)\le\delta,\, s,t\in T_k} |X_n(s) - X_n(t)| > \epsilon\} \\
&\quad\le \liminf_{n\to\infty} Pr\{\max_{\rho(s,t)\le\delta,\, s,t\in T_\infty} |X_n(s) - X_n(t)| > \epsilon\}\,.
\end{aligned}$$

Taking the limit in the last display as $k \to \infty$ and then using the asymptotic equicontinuity condition (1), it follows that there is a sequence $\delta_m \searrow 0$ such that

$$Pr\{\max_{\rho(s,t)\le\delta_m,\, s,t\in T_\infty} |X(s) - X(t)| > \epsilon\} \le 2^{-m}\,.$$

Hence it follows by Borel-Cantelli that there exist $m = m(\omega) < \infty$ a.s. such that

$$\sup_{\rho(s,t)\le\delta_m,\, s,t\in T_\infty} |X(s,\omega) - X(t,\omega)| \le 2^{-m}$$

for all $m > m(\omega)$. Therefore $X(t, \omega)$ is a $\rho-$uniformly continuous function of $t \in T_\infty$ for almost every ω. The extension to T by uniform continuity of the restriction of X to T_∞ yields a version of X with sample paths all in $C_u(T, \rho)$; note that it suffices to consider only the set of ω's upon which X is uniformly continuous. It then follows from Proposition 5.1 that the law of X exists as a tight Borel measure on $\ell^\infty(T)$.

Our proof of convergence will be based on the following fact (see Exercise 6.16): if $H : \ell^\infty(T) \mapsto \mathbb{R}$ is bounded and continuous, and $K \subset \ell^\infty(T)$ is compact, then for every $\epsilon > 0$ there exists $\tau > 0$ such that: if $x \in K$ and $y \in \ell^\infty(T)$ with $\|x - y\|_T < \tau$ then

(a) $\quad |H(x) - H(y)| < \epsilon\,.$

Now we are ready to prove the weak convergence part of (ii). Since (T, ρ) is totally bounded, for every $\delta > 0$ there exists a finite set of points $t_1, \ldots, t_{N(\delta)}$ that is $\delta-$dense in (T, ρ); i.e. $T \subset \cup_{i=1}^{N(\delta)} B(t_i, \delta)$ where $B(t, \delta)$ is the open ball with center t and radius δ. Thus, for each $t \in T$ we can choose $\pi_\delta(t) \in \{t_1, \ldots, t_{N(\delta)}\}$ so that $\rho(\pi_\delta(t), t) < \delta$. Then we can define processes $X_{n,\delta}$, $n \in \mathbb{N}$, and X_δ by

$$X_{n,\delta}(t) = X_n(\pi_\delta(t)) \qquad X_\delta(t) = X(\pi_\delta(t)), \qquad t \in T\,.$$

Note that $X_{n,\delta}$ and X_δ are approximations of the processes X_n and X respectively that can take on at most $N(\delta)$ different values. Convergence of the finite-dimensional distributions of X_n to those of X implies that

(b) $\quad X_{n,\delta} \Rightarrow X_\delta \qquad \text{in} \qquad l^\infty(T)\,.$

Furthermore, uniform continuity of the sample paths of X yields

(c) $\quad \lim_{\delta \to 0} \|X - X_\delta\|_T = 0 \qquad a.s.$

Let $H : \ell^\infty(T) \mapsto \mathbb{R}$ be bounded and continuous. Then it follows that

$$\begin{aligned} &|E^* H(X_n) - EH(X)| \\ &\quad \leq \ |E^* H(X_n) - EH(X_{n,\delta})| + |EH(X_{n,\delta}) - EH(X_\delta)| + |EH(X_\delta) - EH(X)| \\ &\quad \equiv \ I_{n,\delta} + II_{n,\delta} + III_\delta\,. \end{aligned}$$

To show the convergence part of (ii) we need to show that $\lim_{\delta \to 0} \limsup_{n \to \infty}$ of each of these three terms is 0. This follows for $II_{n,\delta}$ by (b). Now we show that $\lim_{\delta \to 0} III_\delta = 0$. Given $\epsilon > 0$, let $K \subset l^\infty(T)$ be a compact set such that $Pr\{X \in K^c\} < \epsilon/(6\|H\|_\infty)$, let $\tau > 0$ be such that (a) holds for K and $\epsilon/6$, and let $\delta_1 > 0$ be such that $Pr\{\|X_\delta - X\|_T \geq \tau\} < \epsilon/(6\|H\|_\infty)$ for all $\delta < \delta_1$; this can be done by virtue of (c). Then it follows that

$$\begin{aligned} |EH(X_\delta) - EH(X)| \ &\leq \ 2\|H\|_\infty Pr\{[X \in K^c] \cup [\|X_\delta - X\|_T \geq \tau]\} \\ &\qquad + \ \sup\{|H(x) - H(y)| : \ x \in K, \ \|x - y\|_T < \tau\} \\ &\leq \ 2\|H\|_\infty \left(\frac{\epsilon}{6\|H\|_\infty} + \frac{\epsilon}{6\|H\|_\infty} \right) + \frac{\epsilon}{6} < \epsilon\,, \end{aligned}$$

so that $\lim_{\delta \to 0} III_\delta = 0$ holds.

To show that $\lim_{\delta \to 0} \limsup_{n \to \infty} I_{n,\delta} = 0$, chose ϵ, τ, and K as above. Then we have

$$\begin{aligned} |E^* H(X_n) - H(X_{n,\delta})| \ &\leq \ 2\|H\|_\infty \left\{ Pr^*\{\|X_n - X_{n,\delta}\|_T \geq \tau/2\} + Pr\{X_{n,\delta} \in (K_{\tau/2})^c\} \right\} \\ &\qquad + \ \sup\{|H(x) - H(y)| : \ x \in K, \ \|x - y\|_T < \tau\} \end{aligned} \tag{d}$$

where $K_{\tau/2}$ is the $\tau/2$ open neighborhood of the set K for the sup norm. The inequality in the previous display can be checked as follows: if $X_{n,\delta} \in K_{\tau/2}$ and $\|X_n - X_{n,\delta}\|_T < \tau/2$, then there exists $x \in K$ such that $\|x - X_{n,\delta}\|_T < \tau/2$ and $\|x - X_n\|_T < \tau$. Now the asymptotic equicontinuity hypothesis implies that there is a δ_2 such that

$$\limsup_{n\to\infty} Pr^*\{\|X_{n,\delta} - X_n\|_T \geq \tau/2\} < \frac{\epsilon}{6\|H\|_\infty}$$

for all $\delta < \delta_2$, and finite-dimensional convergence yields

$$\limsup_{n\to\infty} Pr\{X_{n,\delta} \in (K_{\tau/2})^c\} \leq Pr\{X_\delta \in (K_{\tau/2})^c\} \leq \frac{\epsilon}{6\|H\|_\infty}\,.$$

Hence we conclude from (d) that, for $\delta < \delta_1 \wedge \delta_2$,

$$\limsup_{n\to\infty} |E^* H(X_n) - EH(X_{n,\delta})| < \epsilon\,,$$

and this completes the proof that (i) implies (ii).

The converse implication is an easy consequence of the "closed set" part of the portmanteau theorem: if $X_n \Rightarrow X$ in $\ell^\infty(T)$, then, as for usual convergence in law,

$$\limsup_{n\to\infty} Pr^*\{X_n \in F\} \leq Pr\{X \in F\}$$

for every closed set $F \subset \ell^\infty(T)$; see e.g. Van der Vaart and Wellner (1996), page 18. If (ii) holds, then by Proposition 5.1 there is a pseudometric ρ on T which makes (T, ρ) totally bounded and such that X has (a version with) sample paths in $C_u(T, \rho)$. Thus for the closed set $F = F_{\delta,\epsilon}$ defined by

$$F_{\epsilon,\delta} = \{x \in \ell^\infty(T) : \sup_{\rho(s,t)\leq\delta} |x(s) - x(t)| \geq \epsilon\}\,,$$

we have

$$\begin{aligned}
&\limsup_{n\to\infty} Pr^*\left\{\sup_{\rho(s,t)\leq\delta} |X_n(s) - X_n(t)| \geq \epsilon\right\} \\
&\quad= \limsup_{n\to\infty} Pr^*\{X_n \in F_{\epsilon,\delta}\} \leq Pr\{X \in F_{\epsilon,\delta}\} = Pr\{\sup_{\rho(s,t)\leq\delta} |X(s) - X(t)| \geq \epsilon\}\,.
\end{aligned}$$

Taking limits across the resulting inequality as $\delta \to 0$ yields the asymptotic equicontinuity in view of the $\rho-$uniform continuity of the sample paths of X. Thus (ii) implies (i) $\square$

We conclude this section by stating an obvious corollary of Theorem 5.1 for the empirical process $\mathbb{G}_n$ indexed by a class of measurable real-valued functions $\mathcal{F}$ on the probability space $(\mathcal{X}, \mathcal{A}, P)$, and let ρ_P be the pseudo-metric on $\mathcal{F}$ defined by $\rho_P^2(f,g) = Var_P(f(X) - g(X)) = P(f-g)^2 - [P(f-g)]^2$.

Corollary 1 Let $\mathcal{F}$ be a class of measurable functions on $(\mathcal{X}, \mathcal{A})$. Then the following are equivalent:
(i) $\mathcal{F}$ is $P-$Donsker: $\mathbb{G}_n \Rightarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$.
(ii) $(\mathcal{F}, \rho_P)$ is totally bounded and $\mathbb{G}_n$ is asymptotically equicontinuous with respect to ρ_P in probability: i.e.

$$\lim_{\delta\searrow 0} \limsup_{n\to\infty} Pr^*\left\{\sup_{f,g\in\mathcal{F}:\, \rho_P(f,g)<\delta} |\mathbb{G}_n(f) - \mathbb{G}_n(g)| > \epsilon\right\} = 0 \tag{2}$$

for all $\epsilon > 0$.

We close this section with another equivalent formulation of the asymptotic equicontinuity condition in terms of partitions of the set T.

A sequence $\{X_n\}$ in $\ell^\infty(T)$ is said to be *asymptotically tight* if for every $\epsilon > 0$ there exists a compact set $K \subset \ell^\infty(T)$ such that

$$\liminf_{n\to\infty} P_*(X_n \in K^\delta) \geq 1 - \epsilon \qquad \text{for every} \quad \delta > 0\,.$$

Here $K^\delta = \{y \in \ell^\infty(T) : d(y, K) < \delta\}$ is the "$\delta-$enlargement" of K.

Theorem 5.2 The sequence $\{X_n\}$ in $\ell^\infty(T)$ is asymptotically tight if and only if $X_n(t)$ is asymptotically tight in $\mathbb{R}$ for every $t \in T$ and, for every $\epsilon > 0$, $\eta > 0$, there exists a finite partition $T = \cup_{i=1}^k T_i$ such that

$$\limsup_n P^* \left(\sup_{1\leq i\leq k} \sup_{s,t\in T_i} |X_n(s) - X_n(t)| > \epsilon \right) < \eta\,.$$

Proof. See Van der Vaart and Wellner (1996), Theorem 1.5.6, page 36. □

Example 5.2 (Partial sum process) Suppose that $X_1, X_2, \ldots$ are i.i.d. random variables with $E(X_1) = 0$, $Var(X_1) = 1$. The partial sum process $\mathbb{S}_n$ is defined by

$$\mathbb{S}_n(t) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} X_i \qquad \text{for} \quad 0 \leq t < \infty\,.$$

We will consider the process $\{\mathbb{S}_n(t) : \; 0 \leq t \leq 1\}$. Note that $\mathbb{S}_n$ takes values in $D[0, 1]$ since it has jumps of size $X_i/\sqrt{n}$ at the points $t = i/n$, $i = 1, \ldots, n$. The linearly interpolated version of the process $\overline{\mathbb{S}}_n$ is given by $\overline{\mathbb{S}}_n(k/n) = \mathbb{S}_n(k/n)$ and

$$\overline{\mathbb{S}}_n(t) = \mathbb{S}_n(k/n) + \sqrt{n}(t - k/n)X_{k+1}, \qquad k/n \leq t \leq (k + 1)/n\,.$$

Note that $\overline{\mathbb{S}}_n$ takes values in $C[0, 1]$, and that

$$(3) \qquad \|\overline{\mathbb{S}}_n - \mathbb{S}_n\|_\infty \leq n^{-1/2} \max_{1\leq i\leq n} |X_i| \to_{a.s.} 0$$

since $E(X_1^2) < \infty$.

To show that the finite-dimensional distributions of $\overline{\mathbb{S}}_n$ converge in distribution, we will show that the finite dimensional distributions of $\mathbb{S}_n$ converge in distribution. By (3) the same will hold for $\overline{\mathbb{S}}_n$. Let $0 < t_1 < \cdots < t_k \leq 1$, and consider the random vectors $Y_n \equiv (\mathbb{S}_n(t_1), \ldots, \mathbb{S}_n(t_k))$ in $\mathbb{R}^k$. Define $g : \mathbb{R}^k \to \mathbb{R}^k$ by $g(y) = (y_1, y_2 - y_1, y_3 - y_2, \ldots, y_k - y_{k-1})$. Then

$$g(Y_n) = (\mathbb{S}_n(t_1), \mathbb{S}_n(t_2) - \mathbb{S}_n(t_1), \ldots, \mathbb{S}_n(t_k) - \mathbb{S}_n(t_{k-1}))$$

has components which are independent (by independence of the X_i's), and

$$\begin{aligned} \mathbb{S}_n(t_j) - \mathbb{S}_n(t_{j-1}) &= \frac{1}{\sqrt{n}} \sum_{i=\lfloor nt_{j-1}\rfloor+1}^{\lfloor nt_j\rfloor} X_i \\ &= \frac{\sqrt{\lfloor nt_j\rfloor - \lfloor nt_{j-1}\rfloor}}{\sqrt{n}} \frac{1}{\sqrt{\lfloor nt_j\rfloor - \lfloor nt_{j-1}\rfloor}} \sum_{i=\lfloor nt_{j-1}\rfloor+1}^{\lfloor nt_j\rfloor} X_i \\ &\to_d \sqrt{t_j - t_{j-1}} Z_j \stackrel{d}{=} \mathbb{S}(t_j) - \mathbb{S}(t_{j-1}) \sim N(0, t_j - t_{j-1}), \qquad j = 1, \ldots, k \end{aligned}$$

where $Z = (Z_1, \ldots, Z_k)$ is a vector of independent $N(0,1)$ random variables. Thus it follows that $g(Y_n) \to_d g(Y)$ where $Y \equiv (\mathbb{S}(t_1), \mathbb{S}(t_2) - \mathbb{S}(t_1), \ldots, \mathbb{S}(t_k) - \mathbb{S}(t_{k-1}))$. Now $g^{-1} = h$ where $h : \mathbb{R}^k \to \mathbb{R}^k$ given by $h(x) \equiv (x_1, x_1 + x_2, \ldots, x_1 + \cdots + x_k)$ is continuous. Hence by the continuous mapping theorem $Y_n = h(g(Y_n)) \to_d h(g(Y)) = Y$; i.e.

$$(\mathbb{S}_n(t_1), \ldots, \mathbb{S}_n(t_k)) = Y_n \to_d Y \stackrel{d}{=} (\mathbb{S}(t_1), \ldots, \mathbb{S}(t_k)) .$$

Since $([0,1], |\cdot|)$ is clearly totally bounded, it remains to verify the asymptotic equicontinuity condition:

$$\lim_{\delta \searrow 0} \limsup_{n \to \infty} P(\sup_{|t-s| \le \delta} |\overline{\mathbb{S}}_n(t) - \overline{\mathbb{S}}_n(s)| > \epsilon) = 0 \qquad \text{for every } \ \epsilon > 0 .$$

To do this, let $t_j = j\delta$, $j = 0, \ldots, k \equiv k(\delta)$, and $t_{k+1} = 1$ where k is the largest integer strictly less than $1/\delta$, $k = \lceil 1/\delta \rceil - 1$. Then $t_j - t_{j-1} \le \delta$, $j = 1, \ldots, k+1$, and, by letting $t_j(t)$ denote the largest point t_j to the left of $t \in [0,1]$ we find that

$$\begin{aligned}
&\sup_{|t-s| \le \delta} |\mathbb{S}_n(t) - \mathbb{S}_n(s)| \\
&= \sup_{|t-s| \le \delta} |\mathbb{S}_n(t) - \mathbb{S}_n(t_j(t)) + \mathbb{S}_n(t_j(t)) - \mathbb{S}_n(t_{j'}(s)) + \mathbb{S}_n(t_{j'}(s)) - \mathbb{S}_n(s)| \\
&\le \max_{0 \le j \le k} \left\{ \sup_{t_j \le t \le t_{j+1}} |\mathbb{S}_n(t) - \mathbb{S}_n(t_j)| \right. \\
&\qquad \left. + |S_n(t_{j+1}) - \mathbb{S}_n(t_j)| + \sup_{t_j \le s \le t_{j+1}} |\mathbb{S}_n(s) - \mathbb{S}_n(t_j)| \right\} \\
&\le 3 \max_{0 \le j \le k} \sup_{t_j \le t \le t_{j+1}} |\mathbb{S}_n(t) - \mathbb{S}_n(t_j)| .
\end{aligned}$$

Therefore it follows that, by choosing δ so that, by $\sqrt{\delta} < \epsilon/12$ and using the Ottaviani-Skorokod inequality,

$$\begin{aligned}
&P(\sup_{|t-s| \le \delta} |\mathbb{S}_n(t) - \mathbb{S}_n(s)| > \epsilon) \\
&\le P(\max_{0 \le j \le k} \sup_{t_j \le t \le t_{j+1}} |\mathbb{S}_n(t) - \mathbb{S}_n(t_j)| > \epsilon/3) \\
&\le \sum_{j=0}^{k} P(\sup_{t_j \le t \le t_{j+1}} |\mathbb{S}_n(t) - \mathbb{S}_n(t_j)| > \epsilon/3) \\
&= \sum_{j=0}^{k} P \left(\max_{1 \le l \le \lfloor nt_{j+1} \rfloor - \lfloor nt_j \rfloor} \Big| \sum_{i=\lfloor nt_j \rfloor + 1}^{\lfloor nt_j \rfloor + l} X_i \Big| > (\epsilon/3)\sqrt{n} \right) \\
&\le \sum_{j=0}^{k} 2P \left(\Big| \sum_{i=1}^{\lfloor nt_{j+1} \rfloor - \lfloor nt_j \rfloor} X_i \Big| > (\epsilon/6)\sqrt{n} \right) \\
&\le \sum_{j=0}^{k} 2P \left(\Big| \sum_{i=1}^{\lfloor nt_{j+1} \rfloor - \lfloor nt_j \rfloor} X_i \Big| \ge (\epsilon/6)\sqrt{n} \right) .
\end{aligned}$$

Since

$$\frac{1}{\sqrt{\lfloor nt_{j+1}\rfloor - \lfloor nt_j\rfloor}} \sum_{i=1}^{\lfloor nt_{j+1}\rfloor - \lfloor nt_j\rfloor} X_i \to_d N(0,1)\,,$$

it follows from the portmanteau theorem that

$$\begin{aligned}
&\limsup_{n\to\infty} P(\sup_{|t-s|\le\delta} |\mathbb{S}_n(t) - \mathbb{S}_n(s)| > \epsilon) \\
&\qquad \le\ 2\frac{2}{\delta} P(|Z| \ge (\epsilon/12)\delta^{-1/2}) \\
&\qquad \le\ \frac{4}{\delta}\frac{12\sqrt{\delta}}{\epsilon}\phi((\epsilon/12)\delta^{-1/2}) \qquad \text{by Mills' ratio} \\
&\qquad =\ \frac{48}{\epsilon\sqrt{2\pi}}\delta^{-1/2}\exp\left(-\frac{\epsilon^2}{288}\delta^{-1}\right) \to 0 \qquad \text{as} \quad \delta \searrow 0\,.
\end{aligned}$$

It follows from Theorem 5.1 that $\overline{\mathbb{S}}_n \Rightarrow \mathbb{S}$ in $C[0,1]$. We can also conclude, via (3) that $\mathbb{S}_n \Rightarrow \mathbb{S}$ in $D[0,1]$.

Example 5.3 (Uniform empirical process) Suppose that $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ are i.i.d. Uniform$[0,1]$ random variables. Let $\mathbb{G}_n(t) = n^{-1}\sum_{i=1}^n 1_{[0,t]}(\xi_i)$ for $0 \le t \le 1$ be the empirical distribution function. Then $\mathbb{U}_n(t) = \sqrt{n}(\mathbb{G}_n(t) - t$ for $0 \le t \le 1$ is the *uniform empirical process*.

Now $\mathbb{U}_n \to_{f.d.} \mathbb{U}$ where $\mathbb{U}$ is a standard Brownian bridge process on $[0,1]$ (i.e. $\mathbb{U}$ is a mean 0 Gaussian process with $E\{\mathbb{U}(s)\mathbb{U}(t)\} = s\wedge t - st$ for $0 \le s,t \le 1$). That is, for $0 < t_1 < \ldots < t_k < 1$,

$$\begin{pmatrix} \mathbb{U}_n(t_1) \\ \mathbb{U}_n(t_2) \\ \cdot \\ \cdot \\ \cdot \\ \mathbb{U}_n(t_k) \end{pmatrix} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{pmatrix} 1_{[0,t_1]} - t_1 \\ 1_{[0,t_2]} - t_2 \\ \cdot \\ \cdot \\ \cdot \\ 1_{[0,t_k]} - t_k \end{pmatrix} \to_d \begin{pmatrix} \mathbb{U}(t_1) \\ \mathbb{U}(t_2) \\ \cdot \\ \cdot \\ \cdot \\ \mathbb{U}(t_k) \end{pmatrix} \sim N_k(0, (t_j \wedge t_{j'} - t_j t_{j'}))$$

by the multivariate central limit theorem.

To show that $\mathbb{U}_n \Rightarrow \mathbb{U}$ in $l^\infty([0,1])$, we need to show that $\mathbb{U}_n$ is asymptotically equicontinuous in probability; i.e. that

$$\lim_{\delta\searrow 0}\limsup_{n\to\infty} P(\sup_{|t-s|\le\delta} |\mathbb{U}_n(t) - \mathbb{U}_n(s)| > \epsilon) = 0$$

for every $\epsilon > 0$. Just as we argued in the case of the partial sum process $\mathbb{S}_n$,

$$\sup_{|t-s|\le\delta} |\mathbb{U}_n(t) - \mathbb{U}_n(s)| \le 3 \max_{0\le j\le k} \sup_{t_j \le t \le t_{j+1}} |\mathbb{U}_n(t) - \mathbb{U}_n(t_j)|$$

where again $t_j = j\delta$, $j = 0, \ldots, k(\delta)$, and $k \equiv k(\delta) = \lceil 1/\delta\rceil - 1$. Thus, using a Bennett type exponential bound obtained via Doob's maximal inequality and the martingale $\{\mathbb{U}_n(t)/(1-t) : 0 \le t < 1\}$,

$$\begin{aligned}
&P(\sup_{|t-s|\le\delta} |\mathbb{U}_n(t) - \mathbb{U}_n(s)| > \epsilon) \\
&\quad \le\ P(\max_{0\le j\le k} \sup_{t_j\le t\le t_{j+1}} |\mathbb{U}_n(t) - \mathbb{U}_n(t_j)| > \epsilon/3)
\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{j=0}^{k} P(\sup_{t_j \le t \le t_{j+1}} |\mathbb{U}_n(t) - \mathbb{U}_n(t_j)| > \epsilon/3) \\
&= (k+1) P(\sup_{0 \le t \le \delta} |\mathbb{U}_n(t)| > \epsilon/3) \\
&\le 4k \exp\left(-\frac{\epsilon^2/9}{2\delta} (1-\delta) \psi\left(\frac{\epsilon(1-\delta)}{3\delta\sqrt{n}} \right) \right) .
\end{aligned}
$$

Hence it follows, using $\psi(0) = 1$ that

$$
\begin{aligned}
&\limsup_{n\to\infty} P(\sup_{|t-s|\le\delta} |\mathbb{U}_n(t) - \mathbb{U}_n(s)| > \epsilon) \\
&\le \frac{8}{\delta} \exp\left(-\frac{\epsilon^2}{18\delta}(1-\delta)\psi(0) \right) \\
&= \frac{8}{\delta} \exp\left(-\frac{\epsilon^2}{18\delta}(1-\delta) \right) \\
&\to 0 \qquad \text{as} \;\; \delta \searrow 0 .
\end{aligned}
$$

Thus the asymptotic equicontinuity condition in probability holds, and $\mathbb{U}_n \Rightarrow \mathbb{U}$ in $l^\infty([0,1])$.

6 Problems and Complements

Exercise 6.1 Prove the equivalence of (i) and (ii) in Proposition 11.2.2.

Exercise 6.2 Suppose that $\mu_n \to \mu$ and $\sigma_n^2 \to \sigma^2$ where both μ and σ^2 are finite. Suppose that $Z \sim P_0$ on $\mathbb{R}$.
(a) Show that $X_n \stackrel{d}{=} \mu_n + \sigma_n Z \to_d \mu + \sigma Z \stackrel{d}{=} X$.
(b) Show that for $f \in BL(\mathbb{R})$

$$|Ef(X_n) - Ef(X)| \le \|f\|_{BL} E\{1 \wedge (|\mu_n - \mu| + |\sigma_n - \sigma||Z|)\}\,.$$

Exercise 6.3 Suppose that $X_n \sim N(\mu_n, \sigma_n^2)$ and $X_n \to_d$ (some rv) X. Show that $\mu \equiv \lim_n \mu_n$ and $\sigma^2 \equiv \lim_n \sigma_n^2$ must exist as finite limits, and that $X \sim N(\mu, \sigma^2)$. Hint: choose M with $P(\{M\}) = P(\{-M\}) = 0$ and $P[-M, M] > 3/4$. Then show that if $|\mu_n| > M$ or if σ_n is large enough, show that $P(|X_n| > M) \ge 1/2$. Show that all convergent subsequences of $\{(\mu_n, \sigma_n)\}$ must converge to the same limit.

Exercise 6.4 Give a direct proof of the equivalence of (i) and (iv) in Proposition 2.2. Hint: Consider the functions $\psi_\epsilon(y) = \psi(y/\epsilon)$ where ψ is defined as follows: $\psi(y) = 1$ if $y \le 0$, $\psi(y) = 0$ if $y \ge 1$, and

$$\psi(y) = \frac{\int_y^1 \exp(-1/(u(1-u)))du}{\int_0^1 \exp(-1/(u(1-u)))du} \qquad \text{for } \ 0 \le y \le 1\,.$$

Exercise 6.5 Prove Proposition 2.3.

Exercise 6.6 Formulate and prove an extension of Proposition 2.1 to $\mathbb{R}^k$.

Exercise 6.7 Suppose that X and Y are independent random vectors, and that W is another random vector independent of X with $E(Y) = E(W)$ and $Cov(Y) = Cov(W)$ and satisfying $E|Y|^3 < \infty$ and $E|W|^3 < \infty$. Show that if $f \in C^3(\mathbb{R}^k)$ (define carefully what you mean by this latter class of functions), then

$$|Ef(X+Y) - Ef(X+W)| \le C(E|Y|^3 + E|W|^3)$$

where C is a constant depending only on (third derivatives) of f.

Exercise 6.8 Let Y be a random vector in $\mathbb{R}^k$ with $\mu = E(Y)$ and

$$\Sigma = Cov(Y) = E\{(Y-\mu)(Y-\mu)'\}\,.$$

Thus we can write $\Sigma = A\Lambda^2 A'$ where A is an orthogonal matrix (so $AA' = I$) and Λ is diagonal with each diagonal entry non-negative. Define $B = A\Lambda$. Let Z be a random vector with independent $N(0,1)$ coordinates; thus $Z \sim N_k(0, I)$.
(a) Show that $|\mu| \le E|Y|$. Hint: Note that $u'Y \le |Y|$ for all unit vectors u, and in particular for $u = \mu/|\mu|$.
(b) Show that $E|BZ|^3 = E|\sum_{i=1}^k \lambda_i Z_i|^3 \le (\text{trace}(\Sigma))^{3/2} E|Z_1|^3$.
(c) Show that $E|\mu + AZ|^3 \le 8E|Y|^3 + 8(E|Y|^2)^{3/2} E|Z_1|^3$. Can the factor 8 be improved to 4?

Exercise 6.9 Use the Cramér - Wold device to prove the multivariate CLT from the classical CLT in $\mathbb{R}$, Theorem 2.2.

Exercise 6.10 Prove Proposition 4.1.

Exercise 6.11 Prove Proposition 4.2.

Exercise 6.12 Prove that the Hellinger distance $H(P,Q)$ does not depend on the choice of the dominating measure μ.

Exercise 6.13 Show that (ii) of Theorem 4.2 holds.

Exercise 6.14 Show that (iii) of Theorem 4.2 holds.

Exercise 6.15 (Statistical interpretation of the total variation metric) Consider testing P versus Q. Find the test that minimizes the sum of the error probabilities, and show that the minimum sum of errors is $\|P \wedge Q\| \equiv \int p \wedge q \; d\mu$. in the notation of Proposition 4.2. Note that P and Q are orthogonal as measures if and only if $d_{TV}(P,Q) = 1$ if and only if $\|P \wedge Q\| = 0$ if and only if $\int \sqrt{pq} d\mu \equiv \int \sqrt{dPdQ} = 0$.

Exercise 6.16 Show the basic fact used in the proof of (i) implies (ii) for Theorem 5.1: i.e. if $H : \ell^\infty(T) \mapsto \mathbb{R}$ is bounded and continuous, and $K \subset \ell^\infty(T)$ is compact, then for every $\epsilon > 0$ there is a $\delta > 0$ such that: if $x \in K$ and $y \in \ell^\infty(T)$ with $\|y - x\|_T < \delta$, then $|H(x) - H(y)| < \epsilon$.