BIO 244: Unit 4

Convergence of Stochastic Processes

In this section we will consider convergence results for sequences, $X_1(\cdot),\ X_2(\cdot),\cdots$ of stochastic processes. Unlike the usual setting, where the n^{th} element in the sequence was a random variable X_n, here the n^{th} element is a stochastic process $X_n(\cdot)$. For example, $X_n(\cdot)$ might denote the empirical c.d.f., say $\hat{F}_n(\cdot)$, based on a sample of n patients.

We will introduce the concepts of almost sure convergence, convergence in probability, and weak convergence as generalizations of these concepts for sequences of random variables which require a type of uniformity (in t) for convergence almost surely and in probability, and a tightness condition for weak convergence. We use examples to motivate and illustrate the various definitions. Throughout we assume that the processes $X_n(\cdot)$ and $X(\cdot)$ are defined on a common probability space and are cadlag (continu à droite, limite à gauche); that is, continuous on the right, limits on the left (corlol), where the latter is defined as $\lim_{h\downarrow 0} X_n(a-h)$ existing for each n and a, and similarly for $X(\cdot)$. Also, we use $\|\cdot\|$ to denote the supremum (or "sup") norm; that is,

$$\|X(\cdot)\| = \sup_{t\in\mathcal{T}} \mid X(t) \mid .$$

4.1 Definitions:

4.1.1 Definitions: convergence in probability and almost surely

Definition: The sequence $(X_n(\cdot))$ converges in probability to $X(\cdot)$ as $n\to\infty$, denoted $X_n(\cdot) \xrightarrow{P} X(\cdot)$, if for any $\epsilon>0$,

$$P\left(\|X_n(\cdot)-X(\cdot)\,\|>\epsilon\right)\to 0$$

as $n\to\infty$; that is, if $\|X_n(\cdot)-X(\cdot)\| \xrightarrow{P} 0$.

Note that $Y_n \stackrel{def}{=} \|X_n(\cdot) - X(\cdot)\|$ is just a scalar random variable, and thus we have defined converge in probability for a sequence of processes in terms of ordinary convergence in probability for random variables, using the sup norm. Use of the sup norm here means a uniform convergence (in t) of the function $X_n(\cdot\,;\omega)$ to the function $X(\cdot\,;\omega)$; that is, for any given ϵ and δ, we can find an N such that $n > N$ implies that $P\,(\mid X_n(t) - X(t) \mid < \epsilon \;\text{ for every }\; t) > 1-\delta$.

The rationale for requiring the stronger condition of uniform convergence for convergence in probability instead of pointwise convergence lies in how we sometimes use convergence in probability. For example, if $X_n(\cdot) \stackrel{P}{\to} m(\cdot)$, for some deterministic function $m(\cdot)$, and g is some functional, we would like to be able to conclude that $g(X_n(\cdot)) \stackrel{P}{\to} g(m(\cdot))$ assuming some smoothness on $g(\cdot)$. The following example illustrates that defining $\stackrel{P}{\to}$ in terms of pointwise convergence does not ensure this.

Example 4.1: Suppose that $\mathcal{T} = [0,1]$ and that the $X_n(\cdot)$ are defined on the unit interval probability space. Specifically, define $X_n(\cdot)$ by

$$X_n(t\,;\omega) = \omega n^2 t e^{-nt}\ ,$$

for $\omega \in [0,1]$, $t \in [0,1]$, and $n = 1, 2, \cdots$ (see Figure 4.1).

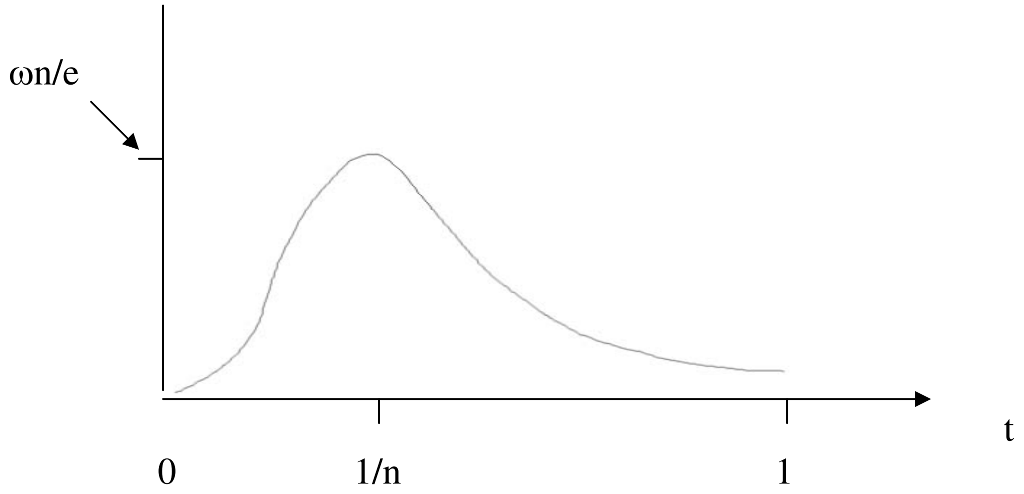

Figure 4.1: Plot of $X_n(t; \omega)$ for Example 4.1

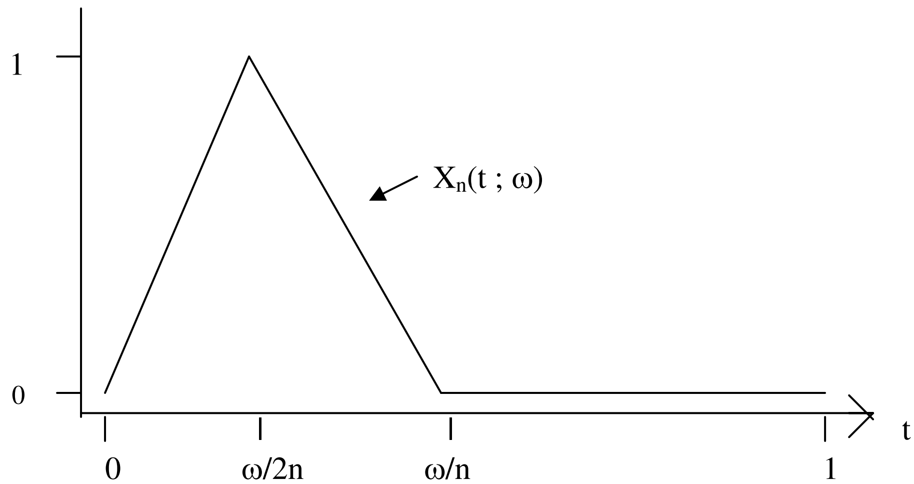

Figure 4.2: Plot of $X_n(t; \omega)$ for Example 4.2

Let $t \in [0, 1]$ be arbitrary but fixed and consider the random variables $Z_n = X_n(t)$ for $n = 1, 2, \cdots$. Then $Z_n \xrightarrow{P} 0$ because

$$P(\mid Z_n \mid > \epsilon) = P\left(\omega : \omega n^2 t e^{-nt} > \epsilon\right) = P\left(\omega : \omega > \frac{\epsilon e^{nt}}{n^2 t}\right) \to 0$$

as $n \to \infty$. Thus, $Z_n = X_n(t)$ converges pointwise in probability to zero.

We could hope that $X_n(\cdot)$ converges in probability to 0, or equivalently, that $\|X_n(\cdot)\|$ converges in probability to 0. So, consider the random variable

$$Z_n^* = \|X_n(\cdot)\|.$$

We investigate whether $Z_n^* \stackrel{P}{\to} 0$. To do this, fix $\omega \in [0,1]$, and note that $X_n(\cdot;\omega)$ has a unique maximum as a function of t since

$$\frac{\partial X_n(t;\omega)}{\partial t} = \omega n^2(-nte^{-nt} + e^{-nt}),$$

which is positive for $t < 1/n$ and negative for $t > 1/n$. The maximum value of $X_n(\cdot;\omega)$ therefore occurs at $t = 1/n$, and equals $\omega n/e$. Thus, $Z_n^*(\omega) = \omega n/e$ and so (assuming without loss of generality that $\epsilon \le 1/e$)

$$P[\mid Z_n^* \mid > \epsilon] = P[\|X_n(\cdot)\| > \epsilon] = P[\omega : \omega n/e > \epsilon] = (1 - \epsilon e/n)\ ,$$

which converges to one as $n \to \infty$. Thus, even though for each t the sequence of random variables $(X_n(t))$ converges in probability to zero, the sequence of processes $(X_n(\cdot))$ does not converge in probability. That is, we have demonstrated a functional (in this case the sup norm) which, with just a pointwise definition of convergence in probability, would not converge in probability when the underlying sequence of processes converges in probability. As we see later, defining convergence in probability the way we have will ensure that continuous functionals will also converge.

Definition: The sequence $(X_n(\cdot))$ converges <u>almost surely</u> to $X(\cdot)$, denoted $X_n(\cdot) \stackrel{a.s.}{\to} X(\cdot)$, if

$$\|X_n(\cdot) - X(\cdot)\| \stackrel{a.s.}{\to} 0$$

as $n \to \infty$. Thus, $X_n(\cdot) \stackrel{a.s.}{\to} X(\cdot)$ if $P(A) = 1$, where

$$A = \{\omega : \sup_t \mid X_n(t;\omega) - X(t;\omega) \mid \to 0 \text{ as } n \to \infty\}.$$

As with the definition of convergence in probability for processes, this definition requires that the sample paths $X_n(\cdot;\omega)$ converge to $X(\cdot;\omega)$ uniformly in t. Indeed, in Example 4.1, it is easy to see that for each fixed t, the sequence $(X_n(t))$ converges almost surely to 0, yet $X_n(\cdot)$ does not converge almost surely to 0.

Note also that almost sure convergence of a sequence of processes implies convergence in probability of the sequence; that is,

$$X_n(\cdot) \stackrel{a.s.}{\to} X(\cdot) \qquad \Rightarrow \qquad X_n(\cdot) \stackrel{P}{\to} X(\cdot), \tag{4.1}$$

This follows since $X_n(\cdot) \stackrel{a.s.}{\to} X(\cdot) \quad \Leftrightarrow \quad \|X_n(\cdot) - X(\cdot)\| \stackrel{a.s.}{\to} 0$, which implies that $\|X_n(\cdot) - X(\cdot)\| \stackrel{P}{\to} 0 \quad \Leftrightarrow \quad X_n(\cdot) \stackrel{P}{\to} X(\cdot)$. Furthermore, since convergence in probability of a sequence of random variables implies convergence in distribution and convergence in distribution to a constant implies convergence in probability to that constant, we also have

$$X_n(\cdot) \stackrel{P}{\to} X(\cdot) \quad \Leftrightarrow \quad \|X_n(\cdot) - X(\cdot)\| \stackrel{P}{\to} 0 \quad \Leftrightarrow \quad \|X_n(\cdot) - X(\cdot)\| \stackrel{\mathcal{L}}{\to} 0.$$

4.1.2 Definitions: weak convergence and tightness

Let us now consider <u>weak convergence</u> of sequences of stochastic processes. For sequences of random variables, "weak convergence" and "convergence in distribution" are used interchangeably. However, when discussing sequences of stochastic processes, the term "weak convergence" tends to be used, and not "convergence in distribution". To emphasize this distinction, we will use "$\stackrel{\mathcal{L}}{\to}$" to denote weak convergence of a sequence of random variables and "$\stackrel{w}{\to}$" to denote weak convergence of a sequence of stochastic processes.

When discussing weak convergence of a sequence of processes, it is tempting to define this in terms of the convergence of the corresponding sequences of finite-dimensional distributions; that is, for any k, $\mathbf{x} = (x_1, \cdots, x_k)$ and

$\mathbf{t} = (t_1, \cdots, t_k)$, let $F_n(\mathbf{x}; \mathbf{t})$ denote the c.d.f. of $(X_n(t_1), \cdots, X_n(t_k))$, and define $F(\mathbf{x}; \mathbf{t})$ analogously for the process $X(\cdot)$. Then we might say that $X_n(\cdot)$ converges weakly to $X(\cdot)$ if, for every k and $\mathbf{t} = (t_1, \cdots, t_k)$,

$$F_n(\mathbf{x}; \mathbf{t}) \to F(\mathbf{x}; \mathbf{t}) \quad \text{as} \quad n \to \infty \ , \tag{4.2}$$

for all $\mathbf{x} = (x_1, \cdots, x_k)$ at which $F(\mathbf{x}; \mathbf{t})$ is continuous. However, we will illustrate in the following example that requiring only (4.2), that is, convergence of finite-dimensional distributions, is not adequate for defining weak convergence of processes because it can lead to undesirable results. We will then go on to argue that an additional condition, "tightness", is needed. The following example is a modification of one in Fleming & Harrington (1991, page 337).

Example 4.2: Let $(\Omega, \mathcal{A}, P)$ denote the unit interval probability space, define $X(\cdot)$ to be identically zero, and define (see Figure 4.2) $X_n(\cdot)$ by

$$X_n(t;\omega) \quad = \quad \begin{cases} 2nt/\omega & \text{if } \ 0 \le t \le \omega/(2n) \\ 2(1 - nt/\omega) & \text{if } \ \omega/(2n) < t \le \omega/n \\ 0 & \text{if } \ \omega/n < t \le 1 \end{cases}$$

Note that the maximum value of $X_n(\cdot;\omega)$ for any n equals 1, and that this occurs when $t = \omega/(2n)$. As n get large, the point at which this maximum is attained moves towards zero. Also, for a given n, the point at which the maximum is attained moves from $t = 0$ to $t = 1/(2n)$ as ω goes from 0 to 1. Also, $X_n(t, \omega) = 0$ from $t = 1/n$ upwards. Thus, for any k and $\mathbf{t} = (t_1, \cdots, t_k)$ (where $t_1 < t_2 < \cdots < t_k$), there exists an N such that

$$n > N \qquad \Rightarrow \qquad X_n(t_1;\omega) = \cdots = X_n(t_k;\omega) = 0 \quad \text{ for all } \omega \ ,$$

and so, trivially,

$$(X_n(t_1), \cdots, X_n(t_k)) \xrightarrow{\mathcal{L}} (X(t_1), \cdots, X(t_k)).$$

That is, all finite dimensional distributions of the sequence $(X_n(\cdot))$ converge in distribution to those of $X(\cdot)$. However, now consider the sequence of random variables (Y_n) where $Y_n \stackrel{def}{=} \|X_n(\cdot)\|$, and let $Y = \|X(\cdot)\|$. Note that $Y_n = 1$ for every n yet $Y \equiv 0$. Thus, even though the finite dimensional distributions of $(X_n(\cdot))$ converge in distribution to those of $X(\cdot)$, the corresponding sequence, (Y_n), of functionals of the $X_n(\cdot)$ does not converge in distribution to Y, the corresponding functional of $X(\cdot)$. To the extent that we would want any definition of weak convergence of a sequence of stochastic processes to satisfy the property that

$$X_n(\cdot) \stackrel{w}{\to} X(\cdot) \quad \Rightarrow \quad g\left(X_n(\cdot)\right) \stackrel{\mathcal{L}}{\to} g\left(X(\cdot)\right),$$

where $\stackrel{w}{\to}$ denotes weak convergence and where $g(\cdot)$ is some continuous functional, then this example shows that defining weak convergence solely by (4.2) is inadequate. Indeed, the $X_n(\cdot;\omega)$ in this example are bounded and continuous in both t and ω, yet convergence of the finite-dimensional distributions is still insufficient to ensure that the sup functional will converge in distribution.

The additional condition (over and above (4.2)) that is imposed on the sequence $(X_n(\cdot))$ to define weak convergence is <u>tightness</u>. Unfortunately, this concept is not as simple for stochastic processes as it is for random variables.

There is more than one **definition of weak convergence** in the literature. Some of these definitions are equivalent, others are not. It always starts with random variables taking values in some metric space M (metric denoted by d):

Definition of metric. A metric d on a space M is a function $d : M \times M \to \mathbb{R}$ satisfying, for all $m_i \in M$:

1. $d(m_1, m_2) \geq 0$ (non-negative)
2. $d(m_1, m_2) = 0$ if and only if $m_1 = m_2$ (identity of indiscernibles)
3. $d(m_1, m_2) = d(m_2, m_1)$ (symmetry)
4. $d(m_1, m_3) \leq d(m_1, m_2) + d(m_2, m_3)$ (triangle inequality) .

A metric space is a space that has a metric. The open sets of a metric space are generated by the open balls: $\{m \in M : d(m, m_0) < \delta\}$, for $m_0 \in M$ and $\delta > 0$. The σ-algebra on M is usually taken to be the Borel σ-algebra: the smallest σ-algebra containing all the open sets.

Perhaps the most used definition of weak convergence is: X_n converges weakly to X if

$$P(X_n \in B) \to P(X \in B) \text{ as } n \to \infty$$

for all Borel sets B with $P(X \in \partial B) = 0$. ∂B is the boundary of B: the difference between the smallest closed set including B and the largest open set included in B. For an open or closed ball, it is its boundary: a circle. For an interval, it is also its boundary: two points for a bounded interval, and for $(-\infty, t]$, it is $\{t\}$. Compare this with the definition of weak convergence in $\mathbb{R}^k$!

This turns out to be the same as another much used definition of weak convergence: X_n converges weakly to X if for all bounded, continuous, real valued functions f,

$$Ef(X_n) \to Ef(X) \text{ as } n \to \infty.$$

That these two definitions are the same is Portmanteau's lemma, see e.g. Billingsley (1968).

Notice that both these definitions of weak convergence depend on the topology on the space M. This leads to the main difference between definitions of weak convergence on $D[0,\tau]$, the space of cadlag functions on $[0,\tau]$. Van der Vaart and Wellner take d the supremum metric:

$$d(f_1, f_2) = \sup_{t\in[0,\tau]} |f_1(t) - f_2(t)|.$$

Billingsley (1968) and Fleming and Harrington take d the so-called Skorohod metric. Intuitively, in the Skorohod metric, functions that are close when you shift the t-axis a little bit are close to each other. For the definition, suppose that Λ is the set of strictly increasing continuous functions λ such that $\lambda(0) = 0$ and $\lambda(\tau) = \tau$. Then

$$d_{Sk}(f_1, f_2) = \inf_{\lambda\in\Lambda} \left\{ \sup_{t\in[0,\tau]} |f_1(\lambda(t)) - f_2(t)| \vee \sup_{t\in[0,\tau]} |\lambda(t) - t| \right\},$$

where $a \vee b$ indicates the maximum of a and b.

For limiting processes with continuous sample paths, like the Wiener process (=Brownian Motion) or the Brownian Bridge, the latter definition turns out to be the same as the definition in Andersen et al., which in general is yet different from the two concepts of weak convergence we introduced before. Andersen et al. take as the σ-algebra on $D[0, \tau]$ the σ-algebra generated by the supremum-norm open balls. This σ-algebra turns out to be smaller than the supremum-norm Borel σ-algbra. Their definition: X_n converges weakly to X if for all bounded, real, continuous, measurable, real valued functions f,

$$Ef(X_n) \to Ef(X) \text{ as } n \to \infty.$$

We will follow Billingsley (1968) and Fleming and Harrington and use the Skorohod metric as the basis for our definition of weak convergence. This turns out to be a weaker definition than the definition in Van der Vaart and Wellner (see below).

Definition: A sequence of random processes X_n **converges weakly** to X if for all bounded, real valued functions f which are continuous with respect to the Skorohod topology,

$$Ef(X_n) \to Ef(X) \text{ as } n \to \infty.$$

Equivalently, X_n converges weakly to X if

$$P(X_n \in B) \to P(X \in B) \text{ as } n \to \infty$$

for all Skorohod-Borel sets B with $P(X \in \partial B) = 0$.

To make the connection with the finite dimensional distributions, let's first have a look at the definition of tightness of a sequence of random processes:

Definition: Tightness of a sequence X_n is usually defined as: for every $\epsilon > 0$ there exists a compact set K with $P(X_n \in K) > 1 - \epsilon$ for all n.

Then we have the following theorem from Fleming and Harrington, Appendix B:

Theorem 4.1: A sequence of random processes X_n in $D[0, \tau]$ converges weakly to X if and only if the finite dimensional distributions converge and the sequence X_n is tight.

Tightness is sometimes hard to check. A sufficient condition is given in the next theorem, from Fleming and Harrington Appendix B, due to Stone (1963).

Theorem 4.2: If for all $\epsilon > 0$

$$lim_{\delta \to 0} limsup_{n \to \infty} P[sup_{|t-s|<\delta} \mid X_n(t) - X_n(s) \mid > \epsilon] = 0 \tag{4.3}$$

then X_n is tight.

Thus, we have:

Corollary 4.1 If the finite dimensional distributions of X_n converge to those of X and (4.3) holds, then X_n converges weakly to X.

Let us assess condition (4.3) in Example 4.2. For any $\omega \in (0,1)$ and $\delta > 0$, it can be seen from Figure 4.2 that

$$sup_{|t-s|<\delta} \mid X_n(t) - X_n(s) \mid = 2n\delta/\omega \cdot 1[\delta \leq \omega/(2n)] \quad + \quad 1 \cdot 1[\delta > \omega/(2n)].$$

Therefore, for $\epsilon \in (0,1)$,

$$\{\omega : sup_{|t-s|<\delta} \mid X_n(t) - X_n(s) \mid > \epsilon\}$$

$$= \{\omega : 2n\delta/\omega > \epsilon \text{ and } \delta \leq \omega/(2n)\} \;\cup\; \{\omega : \delta > \omega/(2n)\}$$

$$= \{\omega : 2n\delta \leq \omega < 2n\delta/\epsilon\} \cup \{\omega : \omega < 2n\delta\} = \{\omega : \omega < 2n\delta/\epsilon\}.$$

For any δ and ϵ, the probability of the term on the right goes to 1 as $n \to \infty$, and hence condition (4.3) doesn't hold.

Example 4.3: Consider the following modification to the process defined in Example 4.2: Let $X(\cdot) \equiv 0$ as before, but now define $X_n(\cdot)$ by

$$X_n(t;\omega) \;=\; \begin{cases} 2nth(n)/\omega & \text{if } \; 0 \leq t \leq \omega/(2n) \\ 2h(n)(1 - nt/\omega) & \text{if } \; \omega/(2n) < t \leq \omega/n \\ 0 & \text{if } \; \omega/n < t \leq 1 \end{cases}$$

See Figure 4.2: the only difference between this $X_n(\cdot)$ and the one in Example 4.2 is that the maximum value of $X_n(\cdot;\omega)$ is $h(n)$ instead of 1. It is clear that, as in Example 4.2, the finite-dimensional distributions of $X_n(\cdot)$ converge to those of $X(\cdot)$. Let's examine whether tightness holds. We check (4.3), hoping to apply Corollary 4.1.

$$sup_{|t-s|<\delta} \mid X_n(t) - X_n(s) \mid = 2nh(n)\delta/\omega \cdot 1[\delta \le \omega/(2n)] \; + \; h(n) \cdot 1[\delta > \omega/(2n)]$$

$$\le h(n)1[\delta \le \omega/(2n)] \; + \; h(n)1[\delta > \omega/(2n)] \; = \; h(n) \; ,$$

so that

$$P[sup_{|t-s|<\delta} \mid X_n(t) - X_n(s) \mid > \epsilon] \; \le \; P[h(n) > \epsilon] \; .$$

Thus, (4.3) will hold if $h(n) = o(1)$. Thus, for example, if we had taken $h(n) = 1/n$ in this example, it would follow that (4.3) holds and thus that $X_n(\cdot) \stackrel{w}{\to} 0$. Comparing this to Example 4.2, we see that requiring the maximum value of the function $X_n(\cdot;\omega)$ to converge to zero, as we did, ensures weak convergence. Graphically, this prevents the 'spike' in $X_n(\cdot;\omega)$ at 1 that occurs in Example 4.2 when $t = \omega/(2n)$. Note also that if we define $Y_n = \|X_n(\cdot)\|$, then $Y_n \equiv h(n)$; if $Y \equiv 0$ then $Y_n \stackrel{\mathcal{L}}{\to} Y$ as long as $h(n) \to 0$; hence, indeed, when $h(n) = o(1)$, $\|X_n(\cdot)\| \stackrel{\mathcal{L}}{\to} \|X(\cdot)\|$.

4.2 Some Key Results Relating to Convergence of Processes:

We begin this subsection with the Glivenko-Cantelli Theorem, which shows that the ecdf converges almost surely to the corresponding c.d.f. that it estimates. We then consider the weak convergence of the empirical c.d.f., and conclude the section with stochastic process analogs of other useful convergence results for random variables, including generalizations of Slutsky's theorem, the continuous mapping theorem, and the delta method.

Recall the empirical cumulative distribution function (ecdf), denoted $\hat{F}_n(\cdot)$, based on a sample of n i.i.d. random variables $X_1,\ X_2, \cdots,\ X_n$. At time t this is defined by

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} 1[X_i \leq t] \ . \tag{4.4}$$

Thus for fixed t, the ecdf is just the average of n i.i.d. random variables. The strong law of large numbers (for sequences of random variables) ensures that $\hat{F}_n(t) \overset{a.s.}{\to} F(t)$ for every t. The Glivenko-Cantelli Theorem can be viewed as a stochastic process analog of the strong Law of Large Numbers.

Glivenko-Cantelli Theorem: Suppose that $X_1, X_2, \cdots$ are i.i.d. random variables with common distribution function $F(\cdot)$. Then

$$\hat{F}_n(\cdot) \overset{a.s.}{\to} F(\cdot) \qquad \text{as } n \to \infty \ . \tag{4.5}$$

That is, $P(A) = 1$, where

$$A = \{\omega : \sup_t \mid \hat{F}_n(t;\omega) - F(t) \mid \to 0 \ \text{ as } \ n \to \infty\} \ .$$

Proof: See Appendix.

Let us continue to focus on the ecdf. We earlier showed that for any fixed s and t ($s \le t$),

$$E\left(\hat{F}_n(t)\right) = F(t) \qquad \text{and} \qquad Cov\left(\hat{F}_n(s), \hat{F}_n(t)\right) = F(s)[1 - F(t)]/n.$$

From the central limit theorem for i.i.d. random variables, we also know that

$$\sqrt{n}\left(\hat{F}_n(t) - F(t)\right) \xrightarrow{\mathcal{L}} N(0, F(t)(1 - F(t))).$$

Using the Cramer-Wold device or the Multivariate Central Limit Theorem, it is easily shown that for any k and $t_1, t_2, \cdots, t_k$, the k-vector $\sqrt{n}\left(\hat{F}_n(t_1) - F(t_1), \cdots, \hat{F}_n(t_k) - F(t_k)\right)$ converges in distribution to the multivariate normal distribution with mean $(0, \cdots, 0)$ and with the covariance matrix having (s, t) element $F(s)(1 - F(t))$ for $s \le t$. Thus, the finite dimensional distributions of the process

$$Z_n(\cdot) = \sqrt{n}[\hat{F}_n(\cdot) - F(\cdot)]$$

converge to those of a zero-mean Gaussian process, say $Z(\cdot)$, with covariance function $c(s, t) = F(s)(1-F(t))$ for $s \le t$. Note also that such a Gaussian process has the same finite dimensional distributions as $W_0(F(\cdot))$, where $W_0(\cdot)$ is a Brownian Bridge (this follows by verifying that $W_0[F(\cdot)]$ is a zero-mean Gaussian process with the same covariance function as $Z(\cdot)$). Therefore, if we could show that the sequence $(Z_n(\cdot))$ is <u>tight</u>, we could conclude that

$$Z_n(\cdot) = \sqrt{n}\left(\hat{F}_n(\cdot) - F(\cdot)\right) \xrightarrow{w} W_0(F(\cdot)) \tag{4.6}$$

as $n \to \infty$. The proof that the sequence $(Z_n(\cdot))$ is tight is complicated and can be found in Billingsley (page 105-108), so we will not repeat it here. However, because it holds, then the weak convergence of the ecdf, that is, equation (4.6), follows.

We note that the result in (4.6) can be viewed as a CLT for stochastic processes. The connection with the i.i.d. central limit theorem for sequences of random variables may be more clear if we rewrite $Z_n(\cdot)$ as

$$Z_n(\cdot) \;=\; \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\{1[X_i \le \cdot] - F(\cdot)\}\right).$$

The n processes that comprise the summand are independent with zero mean, and just as with the CLT for sequences of i.i.d. random variables, standardizing the average by $\sqrt{n}$ leads to a stable large-sample result. Since the i.i.d. terms above are bounded, it is not necessary to impose any additional regularity conditions.

We also note that proving tightness is often difficult and thus can complicate proving that a sequence of processes converges weakly. Often, rather than algebraically demonstrating (4.3), it is more convenient to resort to demonstrating other sufficient conditions for tightness. Several of these are discussed in the text by van der Vaart & Wellner (1996) and the monograph by Pollard (1990).

We now state and illustrate several key convergence results for stochastic processes that can be viewed as generalizations of well-known results for sequences of random variables. The first is a generalization of Slutsky's theorem.

Theorem 4.3:

(a) Suppose that $X_n(\cdot) \xrightarrow{w} X(\cdot)$, (4.3) holds for X_n, and $Y_n(\cdot) \xrightarrow{P} m(\cdot)$ as $n \to \infty$, where $m(\cdot)$ is some nonrandom uniformly continuous function. Then

$$X_n(\cdot) + Y_n(\cdot) \xrightarrow{w} X(\cdot) + m(\cdot) \ .$$

(b) Suppose that $X_n(\cdot) \xrightarrow{w} X(\cdot)$, (4.3) holds for X_n, and $Y_n(\cdot) \xrightarrow{P} m(\cdot)$, where $m(\cdot)$ is some nonrandom uniformly continuous function. Suppose further that $(X_n(\cdot))$ and $(Y_n(\cdot))$ are bounded. Then

$$X_n(\cdot)Y_n(\cdot) \xrightarrow{w} m(\cdot)X(\cdot) \ .$$

Proof: See Appendix (a) and Exercises (b).

The following result is a version of the Continuous Mapping Theorem for stochastic processes:

Theorem 4.4 (Continuous Mapping Theorem). Suppose that $X_n(\cdot) \stackrel{w}{\to} X(\cdot)$ and that $g(\cdot)$ is a functional from $D[0,\tau]$ to some metric space which is continuous with respect to the topology generated by the Skorohod metric. Then

$$g(X_n(\cdot)) \stackrel{w}{\to} g(X(\cdot))$$

as $n \to \infty$.

With the equivalent definition of weak convergence in terms of convergence of expectations of bounded continuous functions, the proof of this theorem is trivial: if f is a continuous, bounded function, then $f \circ g$ is continuous and bounded, too. However, we have to interpret the definition of "continuous" here as "continuous with respect to the topology generated by the Skorohod metric". This turns out to be a stronger requirement than "continuous with respect to the supremum norm" (see below).

Alternatively, if the limiting process has continuous sample paths, we can use the characterization of weak convergence as in Andersen et al. and get a continuous mapping theorem for g continuous with respect to the supremum-norm topology and measurable with respect to the σ-algebra generated by the open balls in the uniform metric.

As one example of the Continuous Mapping Theorem, note that the supremum norm (or sup norm), given by $g(f) = \|f(\cdot)\| = sup_t \mid f(t) \mid$, is continuous with respect to the topology generated by the Skorohod metric. Thus $X_n(\cdot) \stackrel{w}{\to} X(\cdot)$ implies that the sequence, $(sup_t \mid X_n(t) \mid)$, of random variables converges in distribution to the random variable $sup_t\, |X(t)|$:

Lemma: $f \to \sup_{t\in[0,\tau]} |f(t)|$ is a continuous functional from $D[0,\tau]$ with the topology induced by Skorohod metric to $\mathbb{R}$.

Proof: Let $f_0 \in D[0,\tau]$ be given. We show that the above functional is continuous in f_0 with respect to the topology induced by the Skorohod metric.

So, let $\epsilon > 0$ be given. Choose $\delta = \epsilon$, and suppose that $d_{Sk}(f, f_0) < \delta$. Then there exists λ strictly increasing and continuous with $\lambda(0) = 0$ and $\lambda(\tau) = \tau$ and such that

$$\sup_{t\in[0,\tau]} |\lambda(t) - t| < \delta \quad \text{and} \quad \sup_{t\in[0,\tau]} |f(t) - f_0(\lambda(t))| < \delta.$$

But then

$$\begin{aligned} \left| \sup_{t\in[0,\tau]} |f(t)| - \sup_{t\in[0,\tau]} |f_0(t)| \right| &= \left| \sup_{t\in[0,\tau]} |f(t)| - \sup_{t\in[0,\tau]} |f_0(\lambda(t))| \right| \\ &\leq \left| \sup_{t\in[0,\tau]} |f(t) - f_0(\lambda(t))| \right| < \delta = \epsilon. \end{aligned}$$

That concludes the proof. □

It is harder to check that a functional is continuous with respect to the Skorohod topology than to the topology generated by the uniform metric:

Lemma: If a map g from $D[0, \tau] \to \mathbb{R}$ is continuous with respect to the topology induced by the Skorohod metric, it is also continuous with respect to the topology induced by the supremum norm.

Proof: Let g as in the lemma be given. Let $f_0 \in D[0, \tau]$ be given. We show that g is continuous at f_0 with respect to the topology induced by the supremum norm. So, let $\epsilon > 0$ be given. Then because of the continuity of g with respect to d_{Sk}, there exists $\delta > 0$ such that $d_{Sk}(f, f_0) < \delta$ implies $|g(f) - g(f_0)| < \epsilon$. We show that this same δ works for the uniform metric. Notice that $d_{Sk}(f, f_0) \leq \sup |f - f_0|$ (take λ the identity function). Hence, if $\sup |f - f_0| < \delta$, $d_{Sk}(f, f_0) < \delta$, so $|g(f) - g(f_0)| < \epsilon$.

□

The other way around, this lemma does not hold: there are functions that are continuous with respect to the topology generated by the supremum norm that are not continuous with respect to the topology generated by the Skorohod metric. Notice that for continuity with respect to the topology generated by the Skorohod metric, more "close" elements have to be mapped "close" together.

Note: this lemma implies that the notion of weak convergence in Van der Vaart and Wellner is stronger than the notion of weak convergence in Fleming and Harrington and Billingsley (1968) that we adopt in this class. Our definition requires the convergence of the expectation of less functions than Van der Vaart and Wellner.

Example 4.4: We return to the ecdf to illustrate these results. For simplicity, let's suppose that the ecdf is based on n *i.i.d.* non-negative continuous random variables so that $F(\cdot)$ is continuous and satisfies $F(0) = 0$. Suppose that we want to find a 95% confidence band for $(F(t) : 0 < t < \infty)$. Then since

$$\sqrt{n}(\hat{F}_n(\cdot) - F(\cdot)) \xrightarrow{w} W_0(F(\cdot)) \qquad \text{as } n \to \infty ,$$

it follows from the Continuous Mapping Theorem that

$$\begin{aligned} sup_{0<t<\infty} \left| \sqrt{n}(\hat{F}_n(t) - F(t)) \right| &\xrightarrow{\mathcal{L}} sup_{0<t<\infty} \left| W_0(F(t)) \right| \\ &= sup_{0<u<1} \left| W_0(u) \right| \ . \end{aligned} \tag{4.8}$$

Now suppose that k is some number satisfying

$$P\,[\ sup_{0<u<1} |W_0(u)| < k\ ] = 0.95 \ .$$

Then for large n,

$$\begin{aligned} 0.95 &\approx P\left[sup_{0<t<\infty} \left| \sqrt{n}[\hat{F}_n(t) - F(t)] \right| < k \right] \\ &= P\left[\sqrt{n}|\hat{F}_n(t) - F(t)| < k \text{ for all } t \right] \\ &= P\left[\hat{F}_n(t) - \frac{k}{\sqrt{n}} < F(t) < \hat{F}_n(t) + \frac{k}{\sqrt{n}} \ \text{ for all } \ t \right] . \end{aligned}$$

That is,

$$\hat{F}_n(\cdot) \pm \frac{k}{\sqrt{n}}$$

is an approximate 95% confidence band for the function $F(\cdot)$.

Recall that in Example 4.2 we had a process, $X_n(\cdot)$, that converged <u>pointwise</u> to the zero function, but where $\|X_n(\cdot)\|$ didn't converge to zero, and where $X_n(\cdot)$ didn't converge weakly (due to the lack of tightness of the sequence). We know from the continuous mapping theorem that if a process, $(X_n(\cdot))$, converges weakly, the corresponding sequence of random variables, $\|X_n(\cdot)\|$ converges in distribution. The converse of this last result need not hold.

Next, recall the delta method for sequences of random variables. A common version of this says that if

$$\sqrt{n}(Z_n - \theta) \xrightarrow{\mathcal{L}} Z \qquad \text{(for some random variable } Z\text{)} ,$$

and $g(\cdot)$ is some differentiable function, then

$$\sqrt{n}(g(Z_n) - g(\theta)) \xrightarrow{\mathcal{L}} g'(\theta)Z .$$

In fact, even

$$\sqrt{n}(g(Z_n) - g(\theta)) = g'(\theta)\sqrt{n}(Z_n - \theta) + o_p(1).$$

Now consider what might be an analogous result for stochastic processes, which is sometimes called a functional delta method. The following is a special case of a more general theorem proven in Andersen, Borgan, Gill and Keiding (1993). For any real-valued function $g(\cdot)$, and any stochastic process, say $X(\cdot)$, define the stochastic process $g(X(\cdot))$ to be the process whose value at time t is $g(X(t))$. That is, g is a pointwise transformation of $X(\cdot)$.

Theorem 4.5 (Functional Delta-method). Suppose that

$$\sqrt{n}(X_n(\cdot) - \mu(\cdot)) \xrightarrow{w} Z(\cdot)$$

as $n \to \infty$, Z has continuous sample paths, and $g(\cdot)$ is a continuously differentiable function. Then

$$\sqrt{n}\left(g(X_n(\cdot)) - g(\mu(\cdot))\right) \xrightarrow{w} g'(\mu(\cdot))Z(\cdot),$$

and, moreover,

$$\sqrt{n}\left(g(X_n(\cdot)) - g(\mu(\cdot))\right) = g'((\mu(\cdot))\sqrt{n}(X_n(\cdot) - \mu(\cdot)) + o_p(1).$$

Example 4.5: Consider again Example 4.4 where we examine the e.c.d.f. based on n i.i.d. nonnegative continuous random variables, such as the survival times for a group of patients. In such a setting, interest often focuses on the cumulative hazard function $\Lambda(\cdot)$, define as $\Lambda(t) = -ln(1 - F(t)) = \int_0^t \lambda(u)du$, where $\lambda(\cdot)$ is the hazard function. A natural choice as an estimator of $\Lambda(\cdot)$ is

$$\hat{\Lambda}_n(\cdot) = -ln(1 - \hat{F}_n(\cdot)).$$

To investigate the behavior of this estimator, let us apply the functional delta method with $g(u) = -ln(1 - u)$ to the result that

$$\sqrt{n}(\hat{F}_n(\cdot) - F(\cdot)) \stackrel{w}{\to} W_0(F(\cdot)).$$

This yields

$$\sqrt{n}\left(\hat{\Lambda}_n(\cdot) - \Lambda(\cdot)\right) \quad \stackrel{w}{\to} \quad \frac{W_0(F(\cdot))}{1 - F(\cdot)} \quad . \tag{4.9}$$

This limiting process is clearly a zero-mean Gaussian process, and has covariance function (for $s \le t$)

$$\begin{aligned} C(s,t) &= \text{Cov}\left(\frac{W_0(F(s))}{1 - F(s)}, \frac{W_0(F(t))}{1 - F(t)}\right) \\ &= \frac{F(s)(1 - F(t))}{(1 - F(s))(1 - F(t))} \quad = \quad \frac{F(s)}{1 - F(s)}. \end{aligned}$$

Thus, the limiting process is an independent increments process (Exercise 5). Equation (4.9) can be used, for example, to form an approximate confidence band for $\Lambda(\cdot)$ just as we used (4.8) to obtain a confidence band for $F(\cdot)$ (Exercise 7).

Alternatively, one could simply transform the bands we obtained previously for $F(\cdot)$.

Exercises

1. Consider the *ecdf* denoted $\hat{F}_n(\cdot)$, based on $T_1, T_2, \ldots, T_n$ (*i.i.d.* survival times $\sim F(\cdot)$ continuous).

   Let
   $$Z_n(t) \stackrel{\text{def}}{=} \sqrt{n}\left(\hat{F}_n(t) - F(t)\right).$$

   Then for any $k > 0$ and $0 < t_1 < t_2 < \cdots < t_k < \infty$, show that as $n \to \infty$
   $$(Z_n(t_1), \ldots, Z_n(t_k)) \xrightarrow{\mathcal{L}} (Z(t_1), \ldots, Z(t_k)),$$

   where $Z(\cdot)$ is a Gaussian process with mean 0 and covariance function
   $$\text{Cov}\ (Z(s), Z(t)) = F(s)\,(1 - F(t)) \qquad \text{for } s < t.$$

2. Suppose you wanted to "simulate" values from a Gaussian process $Z(t)$ for $0 < t < t^*$, where t^* is some fixed constant. Let $\mu(t)$ and $c(s,t)$ denote the mean and covariance functions of $Z(\cdot)$. Although $Z(\cdot)$ is defined for all t, it is sufficient to generate realizations from $Z(\epsilon), Z(2\epsilon), \ldots, Z(t^*)$, where $\epsilon > 0$ is small.

   How, in principle, could you do this?

3. Suppose that $X_n(\cdot) \xrightarrow{w} X(\cdot)$ and that $Y_n(\cdot) \xrightarrow{P} m(\cdot)$, where $m(\cdot)$ is some (nonrandom) uniformly continuous function, and (4.3) holds for X_n. Suppose further that $(X_n(\cdot))$ and $(Y_n(\cdot))$ are bounded. Prove that
   $$X_n(\cdot)Y_n(\cdot) \xrightarrow{w} m(\cdot)X(\cdot).$$

   Comment: This is another variation of Slutsky's theorem for stochastic processes. The proof is similar to that presented in the appendix of this unit.

4. Show that $C(s,t) = F(s)/(1 - F(s))$ below equation (4.9) implies that the limiting process has independent increments.

5. Suppose that $T_1, T_2, \ldots T_n$ are i.i.d. observations from some continuous survival distribution with c.d.f. $F(\cdot)$. Suppose that the T_i are observed (no censoring), that the density function corresponding to F is positive for all $t > 0$, and that you want to find a 95% confidence band for $F(\cdot)$ over the interval (t_1^*, t_2^*), for some specified values $0 < t_1^* < t_2^*$.

   (a) Using the asymptotic properties of the empirical c.d.f., find an expression for an approximate 95% confidence band (do not worry about whether the band could sometimes be above 1 or below 0).

   (b) Assuming your band involves some unknown constant needed to achieve the desired confidence level, clearly outline a procedure that would enable you to approximate this unknown quantity.

6. Form a confidence band for Λ based on (4.9). Also form a confidence band for Λ based on the confidence band for $F(\cdot)$ given in this unit.

Proof of Glivenko-Cantelli Theorem: Suppose that $r \geq 2$ is an integer and define

$$x_{r,k} = inf\{u : F(u) \geq k/r\} \qquad \text{for } k = 1, 2, \cdots, k-1$$

where we take $x_{r,0} = -\infty$ and $x_{r,r} = \infty$. Then if $t \in [x_{r,k}, x_{r,k+1})$,

$$\begin{aligned}
\hat{F}_n(t) - F(t) &\leq \hat{F}_n(x_{r,k+1}-) - F(t) \\
&= (\hat{F}_n(x_{r,k+1}-) - F(x_{r,k+1}-)) + (F(x_{r,k+1}-) - F(t)) \\
&\leq \hat{F}_n(x_{r,k+1}-) - F(x_{r,k+1}-) + \frac{1}{r} \ .
\end{aligned}$$

Similarly,

$$\begin{aligned}
\hat{F}_n(t) - F(t) &\geq \hat{F}_n(x_{r,k}) - F(t) \\
&= (\hat{F}_n(x_{r,k}) - F(x_{r,k})) - (F(t) - F(x_{r,k})) \\
&\geq \hat{F}_n(x_{r,k}) - F(x_{r,k}) - \frac{1}{r} \ .
\end{aligned}$$

Note that these inequalities hold for every $\omega \in \Omega$. Thus, for <u>any</u> t (not just $t \in [x_{r,k}, x_{r,k+1})$), we have that

$$\mid \hat{F}_n(t) - F(t) \mid$$

$$\leq max_{1 \leq k \leq r-1;\ 1 \leq j \leq r-1}\{\mid \hat{F}_n(x_{r,k}) - F(x_{r,k}\ ) \mid, \mid \hat{F}_n(x_{r,j}-) - F(x_{r,j}-) \mid\} + \frac{1}{r} \ ,$$

and so

$$sup_t \mid \hat{F}_n(t) - F(t) \mid \leq max\{\mid \hat{F}_n(x_{r,k} - F(x_{r,k}\ ) \mid, \mid \hat{F}_n(x_{r,j}-) - F(x_{r,j}-) \mid\} + \frac{1}{r} \ ,$$

where we have left the subscripts off of the "max" on the right hand side for notational simplicity. Taking the limsup of each side (as $n \to \infty$), we get

$$limsup_{n \to \infty} sup_t \mid \hat{F}_n(t) - F(t) \mid \leq \frac{1}{r} \ ,$$

where we have used the facts that $\hat{F}_n(u) \stackrel{a.s.}{\to} F(u)$ and $\hat{F}_n(u-) \stackrel{a.s.}{\to} F(u-)$ for any $u \in (-\infty, \infty)$. Since r is arbitrary, we have that

$$limsup_{n\to\infty} sup_t \mid \hat{F}_n(t) - F(t) \mid \leq 0 \qquad \text{with probability 1 },$$

and thus that

$$lim_{n\to\infty} sup_t \mid \hat{F}_n(t) - F(t) \mid = 0 \qquad \text{with probability 1 },$$

since $sup_t \mid \hat{F}_n(t) - F(t) \mid \geq 0$ with probability 1. That is,

$$\|\hat{F}_n(\cdot) - F(\cdot)\| \stackrel{a.s.}{\to} 0 \qquad \text{as } n \to \infty .$$

Note that since $(\hat{F}_n - F)(\cdot)$ is right-continuous, $\sup_t \mid \hat{F}_n(t;\omega) - F(t) \mid$ is unchanged if the set of t over which the supremum is taken is restricted to the rationals. Thus, $\|\hat{F}_n(\cdot) - F(\cdot)\|$ is a measurable random variable.

Proof of Theorem 4.3 (a): For arbitrary k and times $t_1, \cdots, t_k$, consider the k-vector $(X_n(t_1)+Y_n(t_1), \ \cdots, X_n(t_k)+Y_n(t_k))^T$. Because $X_n(\cdot)$ converges weakly to $X(\cdot)$, so do its finite dimensional distributions. Since $Y_n(\cdot) \stackrel{P}{\to} m(\cdot)$, the vector $(Y_n(t_1), \cdots, Y_n(t_k))^T$ converges in probability to $(m(t_1), \cdots, m(t_k))^T$. Thus, by Slutsky's theorem for vectors,

$$(X_n(t_1)+Y_n(t_1), \ \cdots, X_n(t_k)+Y_n(t_k))^T \ \stackrel{\mathcal{L}}{\to} \ (X(t_1)+m(t_1), \ \cdots, X(t_k)+m(t_k))^T$$

that is, the finite-dimensional distributions of $X_n(\cdot) + Y_n(\cdot)$ converge to those of $X(\cdot) + m(\cdot)$.

It remains to show that the sequence $(X_n(\cdot) + Y_n(\cdot))$ is tight. We show that (4.3) holds. Note that for any s and t,

$$(X_n(t)+Y_n(t))-(X_n(s)+Y_n(s)) \ = \ (X_n(t)-X_n(s)) + (Y_n(t)-m(t)) + (m(s)-Y_n(s)) + (m(t)-m(s)) \, .$$

Thus, for any $\delta > 0$,

$$sup_{|t-s|<\delta} \mid (X_n(t) + Y_n(t)) - (X_n(s) + Y_n(s)) \mid \ \leq \ sup_{|t-s|<\delta} |X_n(t) - X_n(s)|$$

$$+ \ sup_{|t-s|<\delta} |Y_n(t) - m(t)| \ + \ sup_{|t-s|<\delta} |m(s) - Y_n(s)| \ + \ sup_{|t-s|<\delta} |m(t) - m(s)|$$

$$\leq sup_{|t-s|<\delta} |X_n(t) - X_n(s)| + sup_t |Y_n(t) - m(t)| + sup_s |m(s) - Y_n(s)| + sup_{|t-s|<\delta} |m(t) - m(s)|$$

It follows that for any $\epsilon > 0$,

$$limsup_n P(sup_{|t-s|<\delta} |(X_n(t) + Y_n(t)) - (X_n(s) + Y_n(s))| > \epsilon)$$

$$\leq limsup_n P(sup_{|t-s|<\delta} |X_n(t) - X_n(s)| > \epsilon/4) + 2 limsup_n P(sup_t |Y_n(t) - m(t)| > \epsilon/4)$$

$$+ P(sup_{|t-s|<\delta} |m(t) - m(s)| > \epsilon/4) \ .$$

If we now take limits as $\delta \downarrow 0$, the first term on the right hand side of this inequality goes to zero because (4.3) holds for $(X_n(\cdot))$, the second term goes to zero because $Y_n(\cdot) \xrightarrow{P} m(\cdot)$, and the third term goes to zero because $m(\cdot)$ is uniformly continuous. Thus,

$$lim_{\delta\downarrow 0} limsup_n P[sup_{|t-s|<\delta} \mid [X_n(t) + Y_n(t)] - [X_n(s) + Y_n(s)] \mid \ > \ \epsilon] \ = \ 0 \ ;$$

this implies that the sequence $(X_n(\cdot) + Y_n(\cdot))$ is tight.

References

Andersen PK, Borgan, Gill R, and Keiding N. (1993). Statistical Models Based on Counting Processes, Springer-Verlag.

Billingsly P. (1968). Convergence of Probability Measures, John Wiley & Sons, New York.

Fleming TR and Harrington DP (1991). Counting Processes and Survival Analysis, Wiley, New York.

Pollard D (1990). Empirical Processes: Theory and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 2.

Royden R (1968). Real Analysis. Macmillen.

van der Vaart A & Wellner J (1996). Weak Convergence and Empirical Processes. Springer-Verlag