The Convergence of Random Variables and Functions

Seung Hyun Kim

August 27, 2024

Chapter 1

Basic Elements of Probability Theory

In this chapter we define the basic elements of probability theory, namely random variables, their distribution functions, expected values and the independence of random variables. Other concepts that are of secondary importance but which will be employed extensively later on, such as uniform integrability, are also discussed.

1.1 Pushforward Measures

Let $(E, \mathcal{E}, \mu)$ be a measure space, $(F, \mathcal{F})$ a measurable space, and $f : E \to F$ a function measurable relative to $\mathcal{E}$ and $\mathcal{F}$. Define the function $v : \mathcal{F} \to [0, +\infty]$ as

$$v(A) = \mu(f^{-1}(A))$$

for any $A \in \mathcal{F}$, which is well defined because $f^{-1}(A) \in \mathcal{E}$ by measurability, and takes values in $[0, +\infty]$ because μ does. We can show that v defines a measure on $(F, \mathcal{F})$:

- $v(\emptyset) = \mu(f^{-1}(\emptyset)) = 0$, and
- For any disjoint $\{A_n\}_{n \in N_+} \subset \mathcal{F}$, $\{f^{-1}(A_n)\}_{n \in N_+} \subset \mathcal{E}$ is disjoint (for any $n, m \in N_+$ such that $n \neq m$, $f^{-1}(A_n) \cap f^{-1}(A_m) = f^{-1}(A_n \cap A_m) = \emptyset$), and as such, denoting $A = \bigcup_n A_n$,

  $$v(A) = \mu(f^{-1}(A)) = \mu(\bigcup_n f^{-1}(A_n)) = \sum_{n=1}^{\infty} \mu(f^{-1}(A_n)) = \sum_{n=1}^{\infty} v(A_n)$$

  by the countable additivity of μ.

We call v the pushforward of μ with respect to f, and denote it by $v = \mu \circ f^{-1}$. The following theorem relates the integral of a measurable function under v with that under μ:

Theorem 1.1 Let $(E, \mathcal{E}, \mu)$ be a measure space, $(F, \mathcal{F})$ a measurable space, and $h :\in \mathcal{E}/\mathcal{F}$. Defining $v = \mu \circ h^{-1}$, for any $f \in \mathcal{F}_+$, we have

$$\int_F f dv = \int_E (f \circ h) d\mu.$$

If f is a $\mathcal{F}$-measurable complex function that is v-integrable, then $f \circ h$ is $\mathcal{E}$-measurable and μ-integrable with integral given by

$$\int_F f dv = \int_E (f \circ h) d\mu.$$

Proof) Suppose initially f is a $\mathcal{F}$-measurable simple function with canonical form $f = \sum_{i=1}^n \alpha_i \cdot I_{A_i}$ for $\alpha_1, \cdots, \alpha_n \in [0, +\infty)$ and $A_1, \cdots, A_n \in \mathcal{F}$. Then,

$$\begin{aligned} \int_F f dv = \sum_{i=1}^n \alpha_i \cdot v(A_i) &= \sum_{i=1}^n \alpha_i \cdot \mu(h^{-1}(A_i)) \\ &= \sum_{i=1}^n \alpha_i \cdot \int_E (I_{A_i} \circ h)\, d\mu \\ &= \int_E (f \circ h)\, d\mu. \qquad \text{(Linearity of Integration)} \end{aligned}$$

Now let $f \in \mathcal{F}_+$ in general. Then, there exists a sequence $\{f_n\}_{n \in N_+}$ of $\mathcal{F}$-measurable simple functions increasing to f; for any $n \in N_+$,

$$\int_F f_n dv = \int_E (f_n \circ h) d\mu$$

by our above result, and becasue $\{f_n \circ h\}_{n \in N_+}$ is a sequence of non-negative $\mathcal{E}$-measurable functions increasing to $f \circ h$, which is $\mathcal{E}$-measurable because measurability is preserved across compositions, by the MCT we have

$$\int_F f dv = \lim_{n \to \infty} \int_F f_n dv = \lim_{n \to \infty} \int_E (f_n \circ h) d\mu = \int_E (f \circ h) d\mu.$$

Finally, suppose that f is a $\mathcal{E}$-measurable complex function that is v-integrable. This indicates that

$$\int_F Re(f)^{\pm} dv, \int_F Im(f)^{\pm} dv < +\infty,$$

and as such that

$$\int_E (Re(f)^{\pm} \circ h) d\mu, \int_E (Im(f)^{\pm} \circ h) d\mu < +\infty$$

as well by the preceding result. Since

$$Re(f)^{\pm} \circ h = Re(f \circ h)^{\pm}$$

and a similar result holds for $Im(f \circ h)$, it follows that $f \circ h$ is $\mathcal{E}$-measurable and μ-integrable. By the linearity of integration for integrable complex functions, we also have

$$\int_E (f \circ h) d\mu = \int_F f dv.$$

Q.E.D.

The pushforward measure and its integral is integral (pun intended) to the study of the distribution of random variables, as well as ergodicity.

1.2 The Lebesgue-Stieltjes Integral on the Real Line

Here we construct the Lebesgue-Stieltjes measure in the same manner as we did the Lebesgue measure, namely through the Riesz representation theorem and the Riemann-Stieltjes integral as our positive linear functional.

Since we will at present mainly focus on euclidean space, we reintroduce some notations pertaining to euclidean spaces. Let the standard euclidean topology on the euclidean k-space $\mathbb{R}^k$ be denoted $\tau_{\mathbb{R}^k}$; recall that $\tau_{\mathbb{R}^k}$, the metric topology induced by the euclidean metric on $\mathbb{R}^k$, is equivalent to the product topology $\underbrace{\tau_{\mathbb{R}} \times \cdots \times \tau_{\mathbb{R}}}_{k}$ on $\mathbb{R}^k$, where $\tau_{\mathbb{R}}$ is the standard euclidean topology n $\mathbb{R}$, which is exactly the order topology on $\mathbb{R}$. It follows that the Borel σ-algebra $\mathcal{B}(\mathbb{R}^k)$ generated by $\tau_{\mathbb{R}^k}$ is the product σ-algebra $\underbrace{\mathcal{B}(\mathbb{R}) \times \cdots \times \mathcal{B}(\mathbb{R})}_{k}$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra generated by $\tau_{\mathbb{R}}$.
The Lebesgue measure on $\mathbb{R}^k$ will be denoted λ_k and the collection of all Lebesgue measurable sets by $\mathcal{L}_k$; the space $(\mathbb{R}^k, \mathcal{L}_k, \lambda_k)$ is the completion of the corresponding Borel space on $\mathbb{R}^k$. λ_k is also a regular Borel measure that is translation-invariant. If we write λ without a subscript, then it denotes the Lebesgue measure on the real line or some Borel-measurable subset of the real line.

Due to the complexity of constructing the Lebesgue-Stieltjes measure on an arbitrary k-dimensional space, we only construct the version of the measure on the real line. As in the case of the Lebesgue measure, we start by constructing the Riemann-Stieltjes Integral.

1.2.1 Properties of Increasing Right Continuous Functions

Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing and right-continuous function; it then follows that F has the following properties:

- **F has left limits**
  For any $x \in \mathbb{R}$, the set $A = \{F(y) \mid y < x\}$ is bounded above by $F(x) < +\infty$ due to the monotonicity of F, so by the least upper bound property of the real line, there exists an $\alpha \in \mathbb{R}$ such that

  $$\alpha = \sup A = \sup_{y<x} F(y).$$

  We claim that α is precisely the left limit of F at x: to see this, let $\{x_n\}_{n\in N_+}$ be any sequence in $\mathbb{R}$ that increases to x but does not include x. Since $\{F(x_n)\}_{n\in N_+} \subset A$ increasing sequence ($x_n \nearrow x$, it has α as an upper bound and thus converges to $\sup_{n\in N_+} F(x_n) \leq \alpha$. Suppose that $\sup_{n\in N_+} F(x_n) < \alpha$; then, by the definition of the supremum, there exists an $y < x$ such that $\sup_{n\in N_+} F(x_n) < F(y) \leq \alpha$, but letting $\epsilon = x - y > 0$, there exists an

$N \in N_+$ such that

$$x - x_n < \epsilon \quad \text{for any } n \geq N.$$

It follows that $y < x_N$ and thus $F(y) < F(x_N)$, a contradiction. Therefore, $\sup_{n \in N_+} F(x_n) = \alpha$. This holds for any sequence in $\mathbb{R}$ that increases to x but does not include x, so

$$\alpha = \lim_{y \uparrow x} F(y) = F(x-),$$

the left limit of F at x.
This makes F a cádlág function (or rcll, right continuous with left limits).

- **F has countably many discontinuities**
  To see this, let $D \subset \mathbb{R}$ be the set of all discontinuities of F, and define the function $r : D \to \mathbb{Q}$ so that, for any $x \in D$, $r(x)$ is some rational number in the interval $(F(x-), F(x))$. For any $x, y \in D$ such that $x \neq y$, assuming without loss of generality that $x < y$, we have

$$F(x-) < r(x) < F(x) \leq F(y-) < r(y) < F(y),$$

where we used the fact that $F(y-) = \sup_{t<y} F(t) \geq F(x)$. Therefore, $x \neq y$ implies that $r(x) \neq r(y)$, or that r is a one-to-one function. Since $\mathbb{Q}$ is countable, this means that D must also be countable.

1.2.2 The Riemann-Stieltjes Integral

We now construct the Riemann-Stieltjes Integral with respect to F: we proceed in steps.

Step 1: Integral for Step Functions

Let $f : [a, b] \to \mathbb{C}$ be any complex function defined on the interval $[a, b]$; then, for any $n \in N_+$, we define the functional Λ_n^F of f as

$$\Lambda_n^F f = \sum_{i=1}^{n} f(x_i) \cdot (F(x_i) - F(x_{i-1})),$$

where $\{x_0, \cdots, x_n\} \subset [a, b]$ is a partition of $[a, b]$ defined as $x_i = a + \frac{b-a}{n} i$ for $0 \leq i \leq n$. For notational brevity, we put $\Delta F(x_i) = F(x_i) - F(x_{i-1})$.
Note how Λ_n is a positive linear functional; for any $f, g : [a, b] \to \mathbb{R}$ and $z \in \mathbb{C}$,

$$\Lambda_n^F(zf + g) = z \cdot \sum_{i=1}^{n} f(x_i) \cdot \Delta F(x_i) + \sum_{i=1}^{n} g(x_i) \cdot \Delta F(x_i) = z \cdot \Lambda_n^F f + \Lambda_n^F g,$$

while if f is non-negative, then

$$\Lambda_n^F f = \sum_{i=1}^{n} f(x_i) \cdot \Delta F(x_i) \geq 0$$

because each $f(x_i)$ and $F(x_i)$ are non-negative (F is an increasing function). Consequently, Λ_n^F is also montonic, that is,

$$\Lambda_n^F f \leq \Lambda_n^F g$$

if $f \leq g$.

Step 2: Integral for Continuous Compactly Supported Functions

Now choose any real valued $f \in C_c(\mathbb{R})$, that is, let f be a continuous compactly supported real-valued function on $\mathbb{R}$. Letting the (compact) support of f be $K = \overline{\{f \neq 0\}} \subset \mathbb{R}$, there exists a closed interval $[a,b]$ containing K (due to the Heine-Borel theorem, K is bounded). f is uniformly continuous on the real line (for a proof, refer to the construction of the Lebesgue measure in the measure theory text), so for any $\epsilon > 0$ there exists a $\delta > 0$ such that

$$|f(x) - f(y)| < \epsilon \quad \text{for any } |x - y| < \delta.$$

Choose $N \in N_+$ such that $\frac{b-a}{N} < \delta$, and let $n \geq N$. Letting $\{x_0, \cdots, x_n\} \subset [a,b]$ be the corresponding partition of $[a,b]$, define $g_n, h_n : [a,b] \to \mathbb{C}$ as

$$g_n(x) = \min_{y \in [x_{i-1}, x_i]} f(x), \quad h_n(x) = \max_{y \in [x_{i-1}, x_i]} f(x)$$

for any $x \in [a,b]$, assuming that if $x \in [x_{i-1}, x_i]$ for some $0 \leq i \leq n$. g_n, h_n take values in $\mathbb{R}$ because each $[x_{i-1}, x_i]$ is compact and f is continuous (by an application of the extreme value theorem). Because $|x_i - x_{i-1}| = \frac{b-a}{n} < \delta$, it follows from the uniform continuity result that

$$|g_n(x) - h_n(x)| = \max_{x,y \in [x_{i-1}, x_i]} |f(x) - f(y)| < \epsilon$$

for any $x \in [a,b]$.

By another application of the extreme value theorem, we can see that there exist $x^*, x_* \in \mathbb{R}$ such that

$$f(x^*) = \max_{x \in \mathbb{R}} f(x) \quad \text{and} \quad f(x_*) = \min_{x \in \mathbb{R}} f(x).$$

Since $g_n(x) \leq f(x) \leq h_n(x)$ for any $x \in [a,b]$, we have

$$f(x_*) \cdot (F(b) - F(a)) \leq \Lambda_n^F g_n = \sum_{i=0}^{n} g_n(x) \cdot \Delta F(x_i) \leq \sum_{i=0}^{n} f(x) \cdot \Delta F(x_i) = \Lambda_n^F f$$

$$\leq \sum_{i=0}^{n} h_n(x) \cdot \Delta F(x_i) = \Lambda_n^F h_n \leq f(x^*) \cdot (F(b) - F(a))$$

and

$$\Lambda_n^F h_n - \Lambda_n^F g_n = \sum_{i=0}^{n} (h_n(x) - g_n(x)) \cdot \Delta F(x_i) < \epsilon \cdot \sum_{i=0}^{n} \Delta F(x_i) = \epsilon \cdot (F(b) - F(a)).$$

This holds for any $n \geq N$, so

$$\liminf_{n\to\infty} \Lambda_n^F g_n \leq \liminf_{n\to\infty} \Lambda_n^F f,$$
$$\limsup_{n\to\infty} \Lambda_n^F f \leq \limsup_{n\to\infty} \Lambda_n^F h_n,$$

where all limits are in $\mathbb{R}$ becauase the sequences $\{\Lambda_n^F g_n\}_{n\in N_+}$ and $\{\Lambda_n^F h_n\}_{n\in N_+}$ take values in $[f(x_*) \cdot (F(b) - F(a)), f(x^*) \cdot (F(b) - F(a))]$.
We can also see that

$$\limsup_{n\to\infty} \Lambda_n^F h_n - \liminf_{n\to\infty} \Lambda_n^F g_n \leq \limsup_{n\to\infty} \left(\Lambda_n^F h_n - \Lambda_n^F g_n\right) \leq \epsilon \cdot (F(b) - F(a)).$$

By implication,

$$0 \leq \limsup_{n\to\infty} \Lambda_n^F f - \liminf_{n\to\infty} \Lambda_n^F f \leq \epsilon \cdot (F(b) - F(a)).$$

Finally, this holds for any $\epsilon > 0$, so

$$\limsup_{n\to\infty} \Lambda_n^F f = \liminf_{n\to\infty} \Lambda_n^F f = \lim_{n\to\infty} \Lambda_n^F f.$$

The Riemann-Stieltjes integral of f with respect to F is defined as

$$\Lambda^F f = \lim_{n\to\infty} \Lambda_n^F f \in \mathbb{R}.$$

For an arbitrary complex valued $f \in C_c(\mathbb{R})$, since $Re(f), Im(f)$ are real-valued continuous compactly supported functions on $\mathbb{R}$, by the linearity of each Λ_n^F we can define

$$\begin{aligned}\Lambda^F f &:= \Lambda^F Re(f) + i \cdot \Lambda^F Im(f) \\ &= \lim_{n\to\infty} \Lambda_n^F Re(f) + i \cdot \left(\lim_{n\to\infty} \Lambda_n^F Im(f)\right) \\ &= \lim_{n\to\infty} \Lambda_n^F f \in \mathbb{C}.\end{aligned}$$

That $\Lambda^F : C_c(\mathbb{R}) \to \mathbb{C}$ is a positive linear functional follows from the fact that each Λ_n^F is: for any $f, g \in C_c(\mathbb{R})$ and $z \in \mathbb{C}$, $zf + g \in C_c(\mathbb{R})$, so we have

$$\Lambda^F(zf + g) = \lim_{n\to\infty} \Lambda_n^F(zf + g)$$

$$= z \cdot \left(\lim_{n\to\infty} \Lambda_n^F f \right) + \lim_{n\to\infty} \Lambda_n^F g$$
$$= z \cdot \Lambda^F f + \Lambda^F g,$$

and if f takes values in $[0, +\infty)$, then because each $\Lambda_n^F f \geq 0$, we have

$$\Lambda^F f = \lim_{n\to\infty} \Lambda_n^F f \geq 0$$

as well.

1.2.3 Application of the Riesz Representation Theorem

We can now finally construct the Lebesgue-Stieltjes measure using the positive linear functional that is the Riemann-Stieltjes integral with respect to F.

Theorem 1.2 (Construction of the Lebesgue-Stieltjes Measure on $\mathbb{R}^1$)
Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing, right continuous function. Then, there exists a σ-algebra $\mathcal{L}_F$ and a measure λ_F on $\mathbb{R}$ such that:

i) $\mathcal{L}_F$ contains every Borel set on $\mathbb{R}$, that is, $\mathcal{B}(\mathbb{R}) \subset \mathcal{L}_F$.

ii) **(Completeness)** $(\mathbb{R}, \mathcal{L}_F, \lambda_F)$ is a complete measure space.

iii) **(Regularity of the Measure)** λ_F is a regular Borel measure, that is, for any $A \in \mathcal{L}_F$,

$$\lambda_F(A) = \inf\{\lambda_F(V) \mid A \subset V, V \in \tau_{\mathbb{R}}\}$$
$$= \sup\{\lambda_F(K) \mid K \subset A, K \text{ is compact}\}$$

iv) **(The Approximation Property)** For any $A \in \mathcal{L}_F$ and $\epsilon > 0$, there exist open and closed sets V and K such tat $K \subset A \subset V$ and

$$\lambda_F(V \setminus K) < \epsilon.$$

v) For any half-open interval $(a, b]$ on $\mathbb{R}$ such that $a < b$,

$$\lambda_F((a, b]) = F(b) - F(a).$$

Proof) The topological space $(\mathbb{R}, \tau_{\mathbb{R}})$ is a locally compact Hausdorff space. We saw earlier that the Riemann-Stieltjes integral $\Lambda^F : C_c(\mathbb{R}) \to \mathbb{C}$ is a positive linear functional; by the Riesz representation theorem, it follows that there exists a σ-algebra $\mathcal{L}_F$ on $\mathbb{R}$ and a measure λ_F on $(\mathbb{R}, \mathcal{L}_F)$ such that:

- $\mathcal{B}(\mathbb{R}) \subset \mathcal{L}_F$,

– $(\mathbb{R}, \mathcal{L}_F, \lambda_F)$ is complete,
– $\lambda_F(K) < +\infty$ for any compact $K \subset \mathbb{R}$,
– For any $A \in \mathcal{L}_F$,

$$\lambda_F(A) = \inf\{\lambda_F(V) \mid A \subset V, V \in \tau_{\mathbb{R}}\},$$

– For any $A \in \mathcal{L}_F$ such that $\lambda_F(A) < +\infty$ or $A \in \tau_{\mathbb{R}}$,

$$\lambda_F(A) = \sup\{\lambda_F(K) \mid K \subset A, K \text{ is compact}\},$$

and
– For any $f \in C_c(\mathbb{R})$,

$$\Lambda^F f = \int_{\mathbb{R}} f d\lambda_F.$$

Becauase $\mathbb{R}$ is also σ-compact (it is the union of the sequence $\{[-n,n]\}_{n\in N_+}$ of compact sets in $\mathbb{R}$), by theorem 4.3 of the measure theory text it also follows that λ_F is a regular measure that possesses the approximation property.

It remains to be seen that λ_F is the Lebesgue-Stieltjes measure representing F, that is,

$$\lambda_F((a,b]) = F(b) - F(a).$$

To this end, choose a half open interval $(a,b] \subset \mathbb{R}$ such that $a < b$. For any $n \in N_+$, define

$$A_n = \left[a + \frac{b-a}{2n}, b - \frac{b-a}{2n}\right] \subset (a,b).$$

Because each A_n is a compact set contained in the open set (a,b), and $(\mathbb{R}, \tau_{\mathbb{R}})$ is a locally compact Hausdorff space, by Urysohn's lemma there exists a $f_n \in C_c(\mathbb{R})$ such that $A_n \prec f_n \prec (a,b)$, that is, f_n takes values in $[0,1]$, $f_n(x) = 1$ for any $x \in A_n$, and the support of f_n is contained in (a,b). For any $m \in N_+$, by the monotonicity of Λ^F_m and the fact that $I_{A_n} \leq f_n \leq I_{(a,b)}$, we have

$$\Lambda^F_m I_{A_n} \leq \Lambda^F_m f_n \leq \Lambda^F_m I_{(a,b)}.$$

By definition,

$$\Lambda^F_m I_{A_n} = F\left(b - \frac{b-a}{2n}\right) - F\left(a + \frac{b-a}{2n}\right)$$

and, because $I_{(a,b)}$ is 0 except on the closed interval $[a,b]$,

$$\Lambda^F_m I_{(a,b)} = F\left(a + \frac{b-a}{m}(m-1)\right) - F(a).$$

Therefore, taking $m \to \infty$ on both sides of the above inequality, by the existence of left limits for F we obtain

$$F\left(b - \frac{b-a}{2n}\right) - F\left(a + \frac{b-a}{2n}\right) \leq \Lambda^F f_n \leq F(b-) - F(a).$$

Taking $n \to \infty$ again, the right continuity and the existence of left limits for F once again tell us that

$$\lim_{n\to\infty} \Lambda^F f_n = F(b-) - F(a).$$

From the Riesz representation theorem we concluded that

$$\Lambda^F f_n = \int_{\mathbb{R}} f_n d\lambda_F$$

for any $n \in N_+$. Clearly, $\{f_n\}_{n\in N_+}$ is a sequence of non-negative functions that increases to $I_{(a,b)}$. In addition, because each f_n is continuous, it is Borel-measurable. It follows from the MCT that

$$\lim_{n\to\infty} \Lambda^F f_n = \lim_{n\to\infty} \int_{\mathbb{R}} f_n d\lambda_F = \int_{\mathbb{R}} I_{(a,b)} d\lambda_F = \lambda_F((a,b)).$$

Therefore,

$$\lambda_F((a,b)) = F(b-) - F(a) = \sup_{x<b} F(x) - F(a).$$

To extend this result to half-open intervals, note that

$$(a,b] = \bigcap_n \left(a, b + \frac{1}{n}\right),$$

where $\{\left(a, b+\frac{1}{n}\right)\}_{n\in N_+}$ is a decreasing sequence of sets in $\mathcal{L}_F$ with $\lambda_F((a,b+1)) = \sup_{x<b+1} F(x) - F(a) < +\infty$. By sequential continuity, we now have

$$\begin{aligned}\lambda_F((a,b]) = \lim_{n\to\infty} \lambda_F\left(\left(a, b+\frac{1}{n}\right)\right) &= \lim_{n\to\infty} \left(\sup_{x<b+1/n} F(x) - F(a)\right) \\ &= \inf_{n\in N_+} \sup_{x<b+1/n} F(x) - F(a).\end{aligned}$$

Note that $F(b) \leq \sup_{x<b+1/n} F(x)$ for any $n \in N_+$, so that $F(b) \leq \inf_{n\in N_+} \sup_{x<b+1/n} F(x) = \beta$. Suppose this holds as a strict inequality. Then, letting $\epsilon = \beta - F(b) > 0$, because $\{b + \frac{1}{n}\}_{n\in N_+}$ is a sequence in $\mathbb{R}$ strictly decreasing to b, by the right continuity of F there exists an $N \in N_+$ such that

$$F\left(b + \frac{1}{n}\right) - F(b) < \epsilon$$

for any $n \geq N$. This implies that

$$\sup_{x<b+1/N} F(x) \leq F(b+1/N) < \beta,$$

which contradicts the fact that β is the infimum of $\sup_{x<b+1/n} F(x)$ over $n \in N_+$. It must therefore be the case that

$$F(b) = \inf_{n \in N_+} \sup_{x<b+1/n} F(x),$$

and as such,

$$\lambda_F((a,b]) = F(b) - F(a).$$

Q.E.D.

We have constructed a measure λ_F defined on all Borel sets$\mathcal{B}(\mathbb{R})$ that assigns the mass

$$\lambda_F((a,b]) = F(b) - F(a) \quad \text{and} \quad \lambda_F((a,b)) = F(b-) - F(a)$$

to half-open intervals $(a,b]$ and open intervals (a,b) on the real line. If F were continuous on $\mathbb{R}$, then $F(b-) = F(b)$ and $\lambda_F((a,b]) = \lambda((a,b))$, meaning that $\lambda_F(\{b\}) = 0$ for any $b \in \mathbb{R}$. In fact, the Lebesgue measure on $\mathbb{R}$ is precisely the Lebesgue-Stieltjes measure on $\mathbb{R}$ with $F(x) = x$ for any $x \in \mathbb{R}$.
For the integral of some Borel measurable complex or non-negative function f with respect to λ_F, we often write

$$\int_{\mathbb{R}} f d\lambda_F = \int_{-\infty}^{\infty} f(x) dF(x),$$

given that the integral is well-defined.

1.3 Random Variables and their Distributions

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ an arbitrary measurable space. A random variable taking values in $(E, \mathcal{E})$ is a function $X : \Omega \to E$ that is measurable with respect to $\mathcal{H}$ and $\mathcal{E}$, that is, $X^{-1}(A) \in \mathcal{H}$ for any $A \in \mathcal{E}$. This indicates that the quantity

$$\mathbb{P}(X \in A) = \mathbb{P}\left(X^{-1}(A)\right),$$

or the probability that X takes values in A, is well-defined in $[0,1]$ for any $A \in \mathcal{E}$.
Below we define the distribution, density and distribution function of this arbitrary random variable X.

1.3.1 Distribution

The distribution μ of X is defined as the pushfoward of $\mathbb{P}$ with respect to X, that is, as the measure $\mu = \mathbb{P} \circ X^{-1}$ on $(E, \mathcal{E})$. As such, the value μ assigns to some measurable set $A \in \mathcal{E}$ is

$$\mu(A) = \mathbb{P}(X^{-1}(A)),$$

or precisely the probability that X takes values in A. Since $\mu(E) = \mathbb{P}(X^{-1}(E)) = 1$, μ is a probability measure.

1.3.2 Density

Suppose that v is a σ-finite measure on $(E, \mathcal{E})$ such that the distribution μ of X is absolutely continuous with respect to v, that is, $\mu << v$. By the Radon-Nikodym Theorem, because v is σ-finite and μ is finite, there exists a Radon-Nikodym derivative f of μ with respect to v that is unique a.e. $[v]$, that is, f is a non-negative $\mathcal{E}$-measurable function such that

$$\mu(A) = \int_A f dv$$

for any $A \in \mathcal{E}$, and if the above relationship holds for any other $\mathcal{E}$-measurable function g, then $f = g$ a.e. $[v]$. We call the equivalence class $[f]_v$ the density of X with respect to v; for notational simplicity, we refer to the equivalence class $[f]_v$ by its representative f.

1.3.3 Distribution Function

In the case $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, a related concept is the distribution function of X, which is the function $F : \mathbb{R} \to [0,1]$ defined as

$$F(x) = \mathbb{P}(X \leq x) = \mu((-\infty, x])$$

for any $x \in \mathbb{R}$ and the distribution μ of X. Distribution functions are useful because they determine the distribution of a random variable. To see this, note that $\{(-\infty, x] \mid x \in \mathbb{R}\}$ is a π-system that generates $\mathcal{B}(\mathbb{R})$; if two real random variables X, Y have the same distribution function, then their distributions agree on this π-system, and because they are finite, this also implies that they agree on the Borel σ-algebra $\mathcal{B}(\mathbb{R})$.

The following are some properties of distribution functions that are easily recovered:

- F **takes values in** $[0, 1]$
  This trivially follows from the fact that $\mathbb{P}$ takes values in $[0, 1]$.

- F **is increasing on** $\mathbb{R}$
  This too is trivial; for any $x, y \in \mathbb{R}$ such that $x \leq y$, we have the inclusion $(-\infty, x] \subset (-\infty, y]$, so that, by the monotonicity of measures,

$$F(x) = \mu((-\infty, x]) \leq \mu((-\infty, y]) = F(y).$$

- F **is right continuous on** $\mathbb{R}$
  For any $x \in \mathbb{R}$ and a sequence $\{x_n\}_{n \in N_+}$ that decreases to x, we can see that the sequence $\{(-\infty, x_n]\}_{n \in N_+}$ of sets is decreasing with intersection $(-\infty, x]$. Since μ, being a probability measure, satisfies $\mu((-\infty, x_1]) < +\infty$, by sequential continuity

$$F(x) = \mu((-\infty, x]) = \lim_{n \to \infty} \mu((-\infty, x_n]) = \lim_{n \to \infty} F(x_n),$$

  which shows that F is right continuous.
  Combined with the monotonoicity of F, we obtain the more precise result

$$F(x_n) \searrow F(x).$$

- $F(x) \to 0$ **as** $x \to -\infty$ **and** $F(x) \to 1$ **as** $x \to +\infty$
  Let $\{x_n\}_{n \in N_+}$ be a sequence decreasing to $-\infty$. Then, the sequence $\{(-\infty, x_n]\}_{n \in N_+}$ is a decreasing sequence of Borel measurable sets with intersection $\emptyset$. By sequential continuity again,

$$\lim_{n \to \infty} F(x_n) = \lim_{n \to \infty} \mu((-\infty, x_n]) = \mu(\emptyset) = 0.$$

  This holds for any sequence decreasing to $-\infty$, so

$$\lim_{x \searrow -\infty} F(x) = 0.$$

On the other hand, let $\{x_n\}_{n\in N_+}$ be a sequence increasing to $+\infty$. Then, the sequence $\{(-\infty, x_n]\}_{n\in N_+}$ is an increasing sequence of Borel measurable sets with union $\mathbb{R}$. By sequential continuity again,

$$\lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} \mu((-\infty, x_n]) = \mu(\mathbb{R}) = 1.$$

This holds for any sequence increasing to $+\infty$, so

$$\lim_{x\nearrow+\infty} F(x) = 1.$$

In general any function $F : \mathbb{R} \to \mathbb{R}$ that satisfies the four properties above is called a distribution function. The following is a neat corollary of theorem 1.2:

Corollary to Theorem 1.2 Let $F : \mathbb{R} \to \mathbb{R}$ be a distribution function. Then, there exists a random variable X taking values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that F is the distribution function of X. Another way to put it is that there exists a unique probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$F(x) = \mu((-\infty, x])$$

for any $x \in \mathbb{R}$.

Proof) Because F is increasing and right continuous, by theorem 1.2 there exists a measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that

$$\mu((a, b]) = F(b) - F(a)$$

for any half-open interval $(a, b] \subset \mathbb{R}$ such that $a < b$. For any $x \in \mathbb{R}$, because $\{(-n + x, x]\}_{n\in N_+}$ is an increasing sequence of Borel sets with union $(-\infty, x]$, by sequential continuity

$$\mu((-\infty, x]) = \lim_{n\to\infty} \mu((-n + x, x]) = F(x) - \lim_{n\to\infty} F(-n + x) = F(x),$$

where the fourth property of distribution functions justifies the last equality. Finally, by sequential continuity and the fourth property again,

$$\mu(\mathbb{R}) = \lim_{n\to\infty} F(n) = 1,$$

so μ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The uniqueness of μ follows from the fact that F determines μ, as stated earlier.
The existence of a random variable X with F as its distribution function is then trivial; we need only put the probability space $(\Omega, \mathcal{H}, \mathbb{P})$ as $\Omega = \mathbb{R}$, $\mathcal{H} = \mathcal{B}(\mathbb{R})$ and $\mathbb{P} = \mu$, and the random variable defined as $X(\omega) = \omega$ for any $\omega \in \Omega$ has F as its distribution function. Q.E.D.

Returning to our original setting, in which F is the distribution function of some real valued random variable X with distribution μ, note that, for any $x \in \mathbb{R}$,

$$\mu((-\infty, x)) = \lim_{n\to\infty} \mu\left(\left(-\infty, x - \frac{1}{n}\right]\right) = \sup_{n\to\infty} F\left(x - \frac{1}{n}\right) = F(x-)$$

by the sequential continuity of μ. This tells us that

$$\mu(\{x\}) = \mu((-\infty, x]) - \mu((-\infty, x)) = F(x) - F(x-).$$

We say that X is a continuous random variable if its distribution function F is continuous; by implication, $\mu(\{x\}) = \mathbb{P}(X = x) = 0$, that is, the probability that X equals a single value is 0. If, in addition, μ is absolutely continuous with respect to the Lebesgue measure λ, then the density f of X with respect to λ is called the probability density function (PDF) of X.

It is worth touching on the (infamous) relationship between the density of a random variable and its distribution function.

Lemma 1.3 Let X be a real-valued random variable with distribution μ, density f with respect to the Lebesgue measure, and distribution function F. If f is continuous at some $x \in \mathbb{R}$, then F is differentiable at x with derivative $F'(x) = f(x)$.

Proof) For any $x \in \mathbb{R}$,

$$F(x) = \mu((-\infty, x]) = \int_{-\infty}^{x} f(t)dt$$

by the definition of the distribution function and density with respect to λ. The result now follows directly from the first fundamental theorem of calculus, but for the sake of completeness we state the proof here.

For any non-zero $h \in \mathbb{R}$, the above relationship implies

$$\begin{aligned}\frac{F(x+h) - F(x)}{h} &= \frac{1}{h}\left[\mu((-\infty, x+h]) - \mu((-\infty, x])\right] \\ &= \frac{1}{h}\mu((x, x+h]) = \frac{1}{h}\int_{x}^{x+h} f(t)dt,\end{aligned}$$

so that

$$\left|\frac{F(x+h) - F(x)}{h} - f(x)\right| = \left|\frac{1}{h}\int_{x}^{x+h} (f(t) - f(x))\, dt\right| \leq \frac{1}{h}\int_{x}^{x+h} |f(t) - f(x)|dt.$$

By the continuity of f at x, for any $\epsilon > 0$ there exists a $\delta > 0$ such that

$$|f(x) - f(t)| < \epsilon \quad \text{for any } |x - t| < \delta.$$

As such, for any non-zero $|h| < \delta$, we have

$$\left| \frac{F(x+h) - F(x)}{h} - f(x) \right| \leq \frac{1}{h} \int_x^{x+h} |f(t) - f(x)| dt \leq \epsilon.$$

This holds for any $\epsilon > 0$, so by definition

$$\lim_{h \to 0} \left| \frac{F(x+h) - F(x)}{h} - f(x) \right| = 0,$$

and we have

$$F'(x) = f(x).$$

Q.E.D.

1.4 The Information Contained in a Random Variable

Here we introduce the idea of σ-algebras as representing the information contained in a certain random variable. Let $(E,\mathcal{E})$ be a measurable space and X a random variable taking values in $(E,\mathcal{E})$. Define the collection

$$\sigma X = \{X^{-1}(A) \mid A \in \mathcal{E}\}$$

of $\mathcal{H}$-measurable subsets of Ω (they are $\mathcal{H}$-measurable becauase X is a random variable). It is easily shown that σX is a σ-algebra on Ω:

- $\Omega = X^{-1}(E)$, so $\Omega \in \sigma X$.
- For any $H \in \sigma X$, there exists an $A \in \mathcal{E}$ such that $H = X^{-1}(A)$; then, $H^c = X^{-1}(A^c) \in \sigma X$ as well.
- For any countable collection $\{H_n\}_{n\in N_+} \subset \sigma X$, letting $H_n = X^{-1}(A_n)$ for some $A_n \in \mathcal{E}$ for all $n \in N_+$, define $A = \bigcup_n A_n \in \mathcal{E}$. It follows that

$$\bigcup_n H_n = \bigcup_n X^{-1}(A_n) = X^{-1}(A) \in \sigma X.$$

Since every set in σX is contained in $\mathcal{H}$, σX is a sub σ-algebra of $\mathcal{H}$. In fact, σX is the smallest (coarsest) σ-algebra on Ω relative to which X is measurable. For this reason, we call σX the σ-algebra generated by X.
We can furnish an intuitive generating set for the σ-algebra generated by a random variable taking values in $(E,\mathcal{E})$, given that we know of a generating set for $\mathcal{E}$.

Lemma 1.4 Let X be a random variable taking values in the measurable space $(E,\mathcal{E})$, and $\mathcal{E}_0$ a subset of E that generates $\mathcal{E}$. Then, the collection

$$\mathcal{H}_0 = \{X^{-1}(A) \mid A \in \mathcal{E}_0\}$$

of subsets of Ω generates σX. In addition, if $\mathcal{E}_0$ is a π-system, so is $\mathcal{H}_0$.

Proof) Clearly, because $\mathcal{H}_0 \subset \sigma X$, the σ-algebra generated by $\mathcal{H}_0$ is contained in σX.
To show the reverse inclusion, define

$$\mathcal{D} = \{A \subset E \mid X^{-1}(A) \in \sigma\mathcal{H}_0\}.$$

We can show that $\mathcal{D}$ is a σ-algebra on E:

i) $E \in \mathcal{D}$ because $X^{-1}(E) = \Omega$ is contained in any σ-algebra on Ω.

ii) For any $A \in \mathcal{D}$, by definition we have $X^{-1}(A) \in \sigma\mathcal{H}_0$, and because σ-algebras are closed under complements, $X^{-1}(A^c) = \left(X^{-1}(A)\right)^c \in \sigma\mathcal{H}_0$. This means that $A^c \in \mathcal{D}$ as well.

iii) For any $\{A_n\}_{n\in N_+}$ with union $A = \bigcup_n A_n$, because $X^{-1}(A_n) \in \sigma\mathcal{H}_0$ for any $n \in N_+$ and $\sigma\mathcal{H}_0$ is closed under countable unions,

$$X^{-1}(A) = \bigcup_n X^{-1}(A_n) \in \sigma\mathcal{H}_0,$$

which tells us that $A \in \mathcal{D}$.

Since $X^{-1}(A) \in \mathcal{H}_0 \subset \sigma\mathcal{H}_0$ for any $A \in \mathcal{E}_0$, $\mathcal{D}$ contains the generating set $\mathcal{E}_0$ of $\mathcal{E}$. Therefore, $\mathcal{D}$ also contains $\mathcal{E}$, which implies that $X^{-1}(A) \in \sigma\mathcal{H}_0$ for any $A \in \mathcal{E}$ and thus that $\sigma X \subset \sigma\mathcal{H}_0$. Putting the results together, we have $\sigma\mathcal{H}_0 = \sigma X$, so that $\mathcal{H}_0$ is a generating set for σX.

Finally, suppose that $\mathcal{E}_0$ is a π-system. Then, choosing any $H_1, H_2 \in \mathcal{H}_0$ and letting $A_i \in \mathcal{E}_0$ satisfy $H_i = X^{-1}(A_i)$ for $i = 1, 2$, we have

$$H_1 \cap H_2 = X^{-1}(A_1 \cap A_2) \in \mathcal{H}_0$$

because $A_1 \cap A_2 \in \mathcal{E}_0$. Therefore, $\mathcal{H}_0$ is also a π-system.
Q.E.D.

Heuristically, σX represents the amount of information contained in X, since σX is exactly the collection of all the events whose probabilities we would know if X were known. If $\sigma X \subset \sigma Y$ for some other random variable Y, then knowledge of Y allows us to find the probabilities of a greater number of events compared to knowledge of X; it is in this sense that Y contains more information than X. In fact, if $\sigma X \subset \sigma Y$, then knowledge of Y essentially implies knowledge of X. This means that X is included in the information set of Y, an idea that is mathematically articulated through the fact that X is σY-measurable in this case.
We can actually go one step further if X is non-negative or complex. An important related result states that, the claim that a non-negative or complex random variable X is included in the information set of a random variable Y taking values in an arbitrary space is equivalent to stating that X is actually a function of Y.

Theorem 1.5 (The Doob-Dynkin Lemma)
Let Y be a random variable taking values in the measurable space $(E, \mathcal{E})$. X is a σY-measurable non-negative or complex random variable if and only if there exists a non-negative or complex $\mathcal{E}$-measurable function f such that $X = f \circ Y$.

Proof) The sufficiency part is very easy to show. Let f be some non-negative or complex $\mathcal{E}$-measurable function; then, because $Y \in \mathcal{H}/\mathcal{E}$ by definition of a random variable and measurability is preserved across compositions, the composition $f \circ Y$ is $\mathcal{H}$-measurable

and thus a random variable. In addition, for any Borel-measurable A,

$$X^{-1}(A) = Y^{-1}(f^{-1}(A)) \in \sigma Y$$

because $f^{-1}(A) \in \mathcal{E}$. This shows us that X is σY-measurable as well.

To show necessity, we follow the usual construction.
Suppose initially that X is a σY-measurable simple function with canonical form $X = \sum_{i=1}^{n} \alpha_i \cdot I_{H_i}$ for $\alpha_1, \cdots, \alpha_n \in [0, +\infty)$ and $H_1, \cdots, H_n \in \sigma Y$. By definition, for any $1 \leq i \leq n$, there exists an $A_i \in \mathcal{E}$ such that $H_i = Y^{-1}(A_i)$; as such,

$$X = \sum_{i=1}^{n} \alpha_i \cdot I_{Y^{-1}(A_i)} = \left(\sum_{i=1}^{n} \alpha_i \cdot I_{A_i} \right) \circ Y.$$

Defining

$$f = \sum_{i=1}^{n} \alpha_i \cdot I_{A_i},$$

because $\alpha_1, \cdots, \alpha_n \in [0, +\infty)$ and $A_1, \cdots, A_n \in \mathcal{E}$, f is a $\mathcal{E}$-measurable simple function; we have thus shown that $X = f \circ Y$ for some $f \in \mathcal{E}_+$.

Now let X be an arbitrary non-negative σY-measurable function. Letting $\{X_n\}_{n \in N_+}$ be a sequence of σY-measurable simple functions increasing to X, for each $n \in N_+$ the preceding result tells us that there exists a $\mathcal{E}$-measurable simple function f_n such that $X_n = f_n \circ Y$. Defining $f = \sup_{n \in N_+} f_n$, f is a $\mathcal{E}$-measurable non-negative function. Furthermore, for any $\omega \in \Omega$, because $\{X_n(\omega) = f_n(Y(\omega))\}_{n \in N_+}$ is an increasing sequence, it converges to $\sup_{n \in N_+} f_n(Y(\omega)) = f(Y(\omega))$. Since $X_n(\omega) \nearrow X(\omega)$, by the uniqueness of the limit, we have $X(\omega) = f(Y(\omega))$. This holds for any $\omega \in \Omega$, so ultimately we have

$$X = f \circ Y$$

for some $f \in \mathcal{E}_+$.

Finally, let X be a complex-valued σY-measurable function. $Re(X)^{\pm}$ and $Im(X)^{\pm}$ are non-negative σY-measurable functions, so by the preceding result, there exist non-negative $\mathcal{E}$-measurable functions f_+, f_-, g_+, g_- such that

$$Re(X)^{\pm} = f_{\pm} \circ Y \quad \text{and} \quad Im(X)^{\pm} = g_{\pm} \circ Y.$$

Now define the $\mathcal{E}$-measurable function

$$\bar{f}_+ = f_+ \cdot I_{\{f_+ < +\infty\}},$$

and likewise for $\bar{f}_-, \bar{g}_+$ and $\bar{g}_-$. Since $Re(X)^{\pm}$ and $Im(X)^{\pm}$ are real-valued (X is complex valued),

$$Re(X)^+ = f_+ \circ Y = \bar{f}_+ \circ Y,$$

where the second equality follows because f_+ and $\bar{f}_+$ agree on the points at which f_+ is finite. Therefore, defining

$$f = \bar{f}_+ - \bar{f}_- + i\,(\bar{g}_+ - \bar{g}_-)\,,$$

which is well-defined because each function comprising the term on the right is real-valued, f is a $\mathcal{E}$-measurable complex function and

$$X = f \circ Y.$$

Q.E.D.

1.5 Expected Values

Let X once again be a random variable taking values in an arbitrary measurable space $(E,\mathcal{E})$, and $f : E \to \mathbb{C}$ a complex $\mathcal{E}$-measurable function. Then, $f \circ X$ is a complex random variable, and the expected value of $f \circ X$ is defined as the integral of $f \circ X$ with respect to $\mathbb{P}$, if it exists:

$$\mathbb{E}[f \circ X] := \int_\Omega (f \circ X) d\mathbb{P}.$$

Suppose μ is the distribution of X, g is its density with respect to some σ-finite measure v, and, in the case that $(E,\mathcal{E}) = (\mathbb{R},\mathcal{B}(\mathbb{R}))$, F is its distribution function. The following are equivalent representations of the expected value of $f \circ X$:

In Terms of the Distribution

Since $\mu = \mathbb{P} \circ X^{-1}$, by theorem 1.1, if f is μ-integrable or non-negative, then the expectation of $f \circ X$ is well-defined and given by

$$\mathbb{E}[f \circ X] = \int_\Omega (f \circ X) d\mathbb{P} = \int_E f d\mu.$$

In Terms of the Density

Since $g \in \mathcal{E}_+$ is the Radon-Nikodym derivative of μ with respect to v, if fg is v-integrable or non-negative, then the expectation of $f \circ X$ is well-defined and given by

$$\mathbb{E}[f \circ X] = \int_E f d\mu = \int_E f(x)g(x)dv(x).$$

In the special case that $(E,\mathcal{E}) = (\mathbb{R}^k,\mathcal{B}(\mathbb{R}^k))$ and $v = \lambda_k$, we can write

$$\mathbb{E}[f \circ X] = \int_{\mathbb{R}^k} f(x)g(x)dx$$

if fg is Lebesgue-integrable or non-negative.

In Terms of the Distribution Function

Suppose $(E,\mathcal{E}) = (\mathbb{R},\mathcal{B}(\mathbb{R}))$. From theorem 1.2 and the determining property of the distribution function, we can see that

$$\mu = \lambda_F$$

on $\mathcal{B}(\mathbb{R})$, where λ_F is the Lebesgue-Stieltjes measure with respect to F. It follows that, if f is λ_F-integrable or non-negative, then the expectation of $f \circ X$ is well-defined and given by

$$\mathbb{E}[f \circ X] = \int_{\mathbb{R}} f d\lambda_F = \int_{-\infty}^{\infty} f(x)dF(x).$$

1.5.1 Moments of a Random Variable

Let X be a random variable taking values in $(E,\mathcal{E})$. For any non-negative or real-valued $\mathcal{E}$-measurable function f and $k \in [1,+\infty)$, the expected value $\mathbb{E}\left[(f\circ X)^k\right]$ is called the kth moment of the random variable $f\circ X$, should the integral exist. For any real or complex valued random variable X, $\mathbb{E}\left[X^k\right]$ is the kth moment of X, and $\mathbb{E}\left[|X|^k\right]$ the kth absolute moment of X.
It is clear that X has finite kth absolute moments if and only if $X \in L^k(\mathcal{H},\mathbb{P})$. $X \in L^k(\mathcal{H},\mathbb{P})$ implies $X \in L^m(\mathcal{H},\mathbb{P})$ for any $m<k$: if $1 \le m < k < +\infty$, then because $\frac{m}{k} + \frac{k-m}{k} = 1$, by Hölder's inequality

$$\begin{aligned}\|X\|_m = \left(\int_\Omega |X|^m d\mathbb{P}\right)^{\frac{1}{m}} &\le \left(\int_\Omega |X|^k d\mathbb{P}\right)^{\frac{1}{k}} \mathbb{P}(\Omega)^{\frac{k-m}{km}} \\ &= \left(\int_\Omega |X|^k d\mathbb{P}\right)^{\frac{1}{k}} = \|X\|_k.\end{aligned}$$

Therefore, for any $1 \le m < k < +\infty$, if X has finite kth absolute moments, then it has finite mth absolute moments as well.

Some moments are encountered more often than others, and so have special designations:

- **Mean**
  Let X be a real or complex random variable with finite first absolute moment. Then, the mean of X is defined as $\mathbb{E}[X]$, which is well-defined (X is $\mathbb{P}$-integrable).

- **Variance**
  Let X be a real random variable with finite second absolute moment. Then, X also has a finite first absolute moment, so that its mean $\mu = \mathbb{E}[X]$ is well-defined. Then, the variance of X is defined as

  $$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2,$$

  where the second inequality follows from the $\mathbb{P}$-integrability of $X - \mathbb{E}[X]$ and the linearity of integration.
  We saw above that Hölder's inequality implies

  $$\mathbb{E}[|X|] = \|X\|_1 \le \|X\|_2 = \left(\mathbb{E}\left[X^2\right]\right)^{\frac{1}{2}}.$$

  it follows that

  $$\mathbb{E}[X]^2 \le \mathbb{E}[|X|]^2 \le \mathbb{E}\left[X^2\right] < +\infty,$$

  so $\mathrm{Var}[X] \ge 0$. That is, the variance of a random variable is always non-negative.

Finally, it is worth noting the implication of $\mathrm{Var}[X] = 0$. In this case,

$$\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \int_\Omega (X(\omega) - \mathbb{E}[X])^2 d\mathbb{P}(\omega) = 0,$$

and because $(X - \mathbb{E}[X])^2$ is a non-negative $\mathcal{H}$-measurable function, by the vanishing property for non-negative functions

$$(X(\omega) - \mathbb{E}[X])^2 = 0$$

for $\mathbb{P}$-almost every $\omega \in \Omega$, that is, $X = \mathbb{E}[X]$ almost surely.

- **Covariance**
  This moment concerns two random variables. Let X and Y be real random variables with finite second absolute moments. Then, $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are well-defined, and by Hölder's inequality,

$$\mathbb{E}[|XY|] \leq \|X\|_2 \|Y\|_2 < +\infty,$$

  meaning that $\mathbb{E}[XY]$ is well-defined as well. The covariance of X and Y is defined as

$$\mathrm{Cov}[X,Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

  If $\mathrm{Var}[X], \mathrm{Var}[Y] > 0$, then the correlation coefficient of X and Y is defined as

$$\mathrm{corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}.$$

  If $\mathrm{Cov}[X,Y] = 0$, or equivalently, $\mathrm{corr}[X,Y] = 0$, then we say X and Y are uncorrelated.

- **Skewness**
  Let X be a real random variable with finite third absolute moment and positive variance. The skewness of X is defined as

$$\mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}[X]}}\right)^3\right],$$

  which measures how skewed the distribution of X is to either side of its mean. The distribution of X is said to be symmetric around the mean if its skewness is 0.

- **Kurtosis**
  Let X be a real random variable with finite fourth absolute moment and positive variance.

The kurtosis of X is defined as

$$\mathbb{E}\left[\left(\frac{X-\mathbb{E}[X]}{\sqrt{\mathrm{Var}[X]}}\right)^4\right],$$

and measures the tail thickness of the distribution of X, that is, how fast the distribution of X peters out as it goes to $\pm\infty$.

Let X be a real random variable with distribution μ. Its moment-generating function (MGF) $m : \mathbb{R} \to [0, +\infty]$ is defined as

$$m(t) = \mathbb{E}[\exp(tX)] = \int_{-\infty}^{\infty} \exp(tx) d\mu(x)$$

for any $t \in \mathbb{R}$; the integral in question always exists in $[0, +\infty]$ because the mapping $x \mapsto \exp(tx)$ on $\mathbb{R}$ is non-negative and measurable for any $t \in \mathbb{R}$.
The function m is called the moment-generating function of X because its kth derivative yields the kth moment of X for any $k \in N_+$, if the moment exists in $\mathbb{R}$. It is also useful to work with MGFs because, if they are real valued on the real line, then they determine the distribution of the associated random variables; that is, if two random variables have the same (real-valued) MGFs, then they have the same distribution. Proofs for these results are omitted here, but can be found in most probability theory texts.

1.5.2 Moments of Random Vectors and Matrices

The discussion so far can be easily extended to random vectors and matrices. Suppose that $X = (X_1, \cdots, X_k)$ is a random vector taking values in $\mathbb{R}^k$. Then, for any $k \in [1, +\infty)$, the kth absolute moment of X is defined as $\mathbb{E}\left[|X|^k\right]$, whee $|\cdot|$ now denotes the euclidean norm on $\mathbb{R}^k$. Here, X has finite kth absolute moments if and only if $|X| \in L^k(\mathcal{H}, \mathbb{P})$, and as such, it still holds that X has finite absolute mth moments for any $m < k$ if X has finite absolute kth moments. Note also that X has finite kth moments if and only if $X_1, \cdots, X_k$ do: this follows from the inequality

$$|X_i| \leq |X| \leq \sum_{j=1}^{k} |X_j|$$

for any $1 \leq i \leq k$, which implies that

$$\|X_i\|_k \leq \||X|\|_k \leq \sum_{j=1}^{k} \|X_j\|_k$$

where the last inequality follows from Minkowski's inequality.

The special moments are defined analogously to the univariate case: throughout, let X be a random vector taking values in $\mathbb{R}^k$.

- **Mean**
  If X has finite first absolute moments, then $X_1, \cdots, X_k$ do as well. This means that the mean of X is well-defined as

  $$\mathbb{E}[X] = (\mathbb{E}[X_1], \cdots, \mathbb{E}[X_k]).$$

- **Variance**
  If X has finite second absolute moments, then so do $X_1, \cdots, X_k$, and the mean of X is well-defined as the vector collecting the means of $X_1, \cdots, X_k$. Thus, we can define the variance of X (also referred to as the variance-covariance matrix of X) as

  $$\begin{aligned}\mathrm{Var}[X] = \begin{pmatrix} \mathrm{Var}[X_1] & \cdots & \mathrm{Cov}[X_1, X_k] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_k, X_1] & \cdots & \mathrm{Var}[X_k] \end{pmatrix} &= \mathbb{E}\left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])'\right] \\ &= \mathbb{E}[XX'] - \mathbb{E}[X]\mathbb{E}[X]'.\end{aligned}$$

  Note that, because $\mathrm{Cov}[X_i, X_j] = \mathrm{Cov}[X_j, X_i]$ for any $1 \leq i, j \leq k$, $\mathrm{Var}[X]$ must be a

symmetric matrix. Furthermore, for any $\alpha \in \mathbb{R}^k$,

$$\alpha' \mathrm{Var}[X]\alpha = \mathbb{E}\left[\alpha'(X - \mathbb{E}[X])(X - \mathbb{E}[X])'\alpha\right] = \mathbb{E}\left[Y^2\right] \geq 0$$

for the real random variable $Y = \alpha'(X - \mathbb{E}[X])$. Therefore, $\mathrm{Var}[X]$ is a positive semidefinite matrix.

Positive definiteness in this case plays the same role as positivity in the univariate case: if $\mathrm{Var}[X]$ is positive semidefinite but not positive definite, then there exists a non-zero $\alpha \in \mathbb{R}^k$ such that

$$\alpha' \mathrm{Var}[X]\alpha = \mathbb{E}\left[\alpha'(X - \mathbb{E}[X])(X - \mathbb{E}[X])'\alpha\right] = 0.$$

This means that the variance of the random variable $Y = \alpha'(X - \mathbb{E}[X])$ is 0, which indicates that $Y = \mathbb{E}[Y] = \mathbf{0}$ almost surely. In other words, X is contained in the $k-1$-dimensional subspace $\{x \in \mathbb{R}^k \mid \alpha' x = \alpha' \mathbb{E}[X]\}$ almost surely.

- **Covariance**

  Let Y be another k-dimensional real random vector. Suppose X and Y have finite second absolute moments. Then, the covariance matrix of X and Y is defined as

$$\begin{aligned}\mathrm{Cov}[X,Y] = \begin{pmatrix} \mathrm{Cov}[X_1,Y_1] & \cdots & \mathrm{Cov}[X_1,Y_k] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_k,Y_1] & \cdots & \mathrm{Cov}[X_k,Y_k] \end{pmatrix} &= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])'\right] \\ &= \mathbb{E}\left[XY'\right] - \mathbb{E}[X]\mathbb{E}[Y]',\end{aligned}$$

  analogously to the univariate case. Note that $\mathrm{Cov}[X,Y]$ is not necessarily symmetric.

As with univariate random variables, we can also define MGFs for random vectors as well. For any random vector X in $\mathbb{R}^k$ with distribution μ, its MGF $m : \mathbb{R}^k \to [0, +\infty]$ is defined as

$$m(t) = \mathbb{E}\left[\exp(t'X)\right] = \int_{\mathbb{R}^k} \exp(t'x) d\mu(x)$$

for any $t \in \mathbb{R}$. It is clear that, for any $t \in \mathbb{R}^k$, $m(t)$ is the MGF of the random variable $t'X$ evaluated at 1.

Random matrices are defined similarly to random vectors. Let $\|\cdot\|$ be a matrix norm with the usual properties of a matrix norm (e.g. the operator norm, the trace norm) and $\mathcal{B}(\mathbb{R}^{m\times n})$ the Borel σ-algebra on $\mathbb{R}^{m\times n}$ generated by the metric topology induced by $\|\cdot\|$. Then, we can define a random matrix X taking values in $(\mathbb{R}^{m\times n}, \mathcal{B}(\mathbb{R}^{m\times n}))$.

As with random vectors, for any $k \in [1, +\infty)$, the kth absolute moment of X is defined as $\mathbb{E}\left[\|X\|^k\right]$. Again, X has finite kth absolute moments if and only if $\|X\| \in L^k(\mathcal{H}, \mathbb{P})$, and as such,

if X has finite kth absolute moments, then X has finite pth absolute moments for any $p < k$ as well.

Similarly to random vectors, X has finite kth absolute moments if and only if each entry X_{ij} has finite kth absolute moments; this follows from the inequality (which holds for both the operator and trace norms)

$$|X_{lp}| \leq \|X\| \leq \sum_{i=1}^{m} \sum_{j=1}^{n} |X_{ij}|$$

for any $1 \leq l \leq m$, $1 \leq p \leq n$.

In the case of random matrices, only the mean of X is well-defined. Specifically, if X has finite first absolute moments, then the mean of X is defined as

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_{11}] & \cdots & \mathbb{E}[X_{1n}] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[X_{m1}] & \cdots & \mathbb{E}[X_{mn}] \end{pmatrix}.$$

1.5.3 Some Important Identities and Inequalities

Here we present some useful identities and inequalities associated with expected values:

Theorem 1.6 The following hold true:

i) **(Integral Representation)** For any non-negative real random variable X,

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\, dt.$$

ii) **(Markov's Inequality)** Let X be a real valued random variable. For any $\epsilon > 0$ and increasing non-negative Borel-measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $f(\epsilon) > 0$,

$$\mathbb{P}(X > \epsilon) \leq \frac{1}{f(\epsilon)} \mathbb{E}[f \circ X].$$

iii) **(Jensen's Inequality)** Let D be a non-empty closed convex subset of $\mathbb{R}^k$ and X a k-dimensional random vector taking values in the interior of D. Suppose X has finite first moments, and that $f : D \to \mathbb{R}$ is a convex function such that $f \circ X$ is integrable. Then, $\mathbb{E}[X]$ is contained in D^o, and

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f \circ X].$$

Proof) i) Let X be a real non-negative random variable. For any $\omega \in \Omega$,

$$\begin{aligned} X(\omega) = \int_0^{X(\omega)} dt &= \int_{\mathbb{R}} I_{[0, X(\omega))}(t) dt \\ &= \int_0^\infty I_{(t, +\infty)}(X(\omega)) dt, \end{aligned}$$

so by Fubini's theorem for non-negative functions,

$$\begin{aligned} \mathbb{E}[X] &= \int_\Omega \int_0^\infty I_{(t,+\infty)}(X(\omega)) dt d\mathbb{P} \\ &= \int_0^\infty \mathbb{E}\left[I_{(t,+\infty)} \circ X\right] dt \\ &= \int_0^\infty \mathbb{P}(X > t)\, dt. \end{aligned}$$

ii) Let $f : \mathbb{R} \to \mathbb{R}$ and $\epsilon > 0$ be chosen as in the claim above. If $X > \epsilon$, then $f \circ X > f(\epsilon)$ because f is increasing, and f is non-negative, so we have the inequality

$$f(\epsilon) \cdot I_{\{X > \epsilon\}} \leq f \circ X.$$

Integrating both sides with respect to $\mathbb{P}$ yields

$$f(\epsilon) \cdot \mathbb{E}\left[I_{\{X>\epsilon\}}\right] \leq \mathbb{E}[f \circ X]$$

by the monotonicity and linearity of the integration of non-negative functions. Since $\mathbb{E}\left[I_{\{X>\epsilon\}}\right] = \mathbb{P}(X > \epsilon)$ and $f(\epsilon) > 0$, it follows that

$$\mathbb{P}(X > \epsilon) \leq \frac{1}{f(\epsilon)} \mathbb{E}[f \circ X],$$

as was claimed.

iii) We first show that $\mathbb{E}[X] \in D^o$. Suppose, for the sake of contradiction, that $\mathbb{E}[X] \notin D^o$. Then, since D is a closed convex set with non-empty interior, by a separation result (refer to the text on hemicontinuity and convex analysis) there exists a non-zero $v \in \mathbb{R}^k$ such that $v'\mathbb{E}[X] > v'y$ for any $y \in D^o$. Since X takes values in D^o, it follows that

$$v'(\mathbb{E}[X] - X) > 0$$

on Ω. The linearity of expectations, however, implies that

$$\mathbb{E}[v'(\mathbb{E}[X] - X)] = 0.$$

The integrand is a non-negative function, so by the vanishing property we must have $v'(\mathbb{E}[X] - X) = 0$ almost surely, which contradicts the inequality above. Therefore, $\mathbb{E}[X] \in D^o$.

Now we prove the main result. Let W_f be the set of all affine minorants of f, that is, the set of affine functions $h : \mathbb{R}^k \to \mathbb{R}$ such that $h(x) \leq f(x)$ for any $x \in D$. Convex functions can be characterized as the supremum of these affine minorants, that is,

$$f(x) = \sup_{h \in W_f} h(x) \quad \text{for any } x \in D^o.$$

Choose any $h \in W_f$; by the definition of affine functions, there exist $v \in \mathbb{R}^k$ and $c \in \mathbb{R}$ such that

$$h(x) = v'x + c \quad \text{for any } x \in \mathbb{R}^k.$$

By the linearity of integration, $h \circ X$ is integrable with integral

$$\mathbb{E}[h \circ X] = v' \mathbb{E}[X] + c = h(\mathbb{E}[X]).$$

Since $h(x) \leq f(x)$ for any $x \in D^o$, and X takes values in D^o, we have $h \circ X \leq f \circ X$ on Ω and, by the monotonicity of integration,

$$h(\mathbb{E}[X]) = \mathbb{E}[h \circ X] \leq \mathbb{E}[f \circ X].$$

This holds for any $h \in W_f$, so

$$\mathbb{E}[f \circ X] \geq \sup_{h \in W_f} h(\mathbb{E}[X]).$$

Finally, $\mathbb{E}[X] \in D^o$ implies that

$$\mathbb{E}[f \circ X] \geq \sup_{h \in W_f} h(\mathbb{E}[X]) = f(\mathbb{E}[X]).$$

Q.E.D.

1.6 Characteristic Functions

Let μ be a finite measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$. The Fourier transform of μ is defined as the function $\varphi : \mathbb{R}^k \to \mathbb{C}$ such that

$$\varphi(t) = \int_{\mathbb{R}^k} \exp(it'x) d\mu(x) = \int_{\mathbb{R}^k} \left(\prod_{j=1}^{k} \exp(it_j x_j) \right) d\mu(x)$$

for any $t \in \mathbb{R}^k$. Note that the integral always exists because $|\exp(it'x)| = 1$ for any $t, x \in \mathbb{R}^k$ and μ is a finite measure.
The following are some basic properties of characteristic functions:

Theorem 1.7 Let μ be a finite measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ with characteristic function $\varphi : \mathbb{R}^k \to \mathbb{C}$. Then, the following hold true:

i) **(Continuity of the Characteristic Function)** φ is uniformly continuous on $\mathbb{R}^k$.

ii) **(Riemann-Lebesgue Lemma)** If μ is absolutely continuous with respect to the Lebesgue measure, then φ vanishes at infinity, that is, $\varphi(t) \to 0$ as $|t| \to \infty$.

Proof) The uniform continuity result is easy to prove. For any $t, h \in \mathbb{R}^k$,

$$\begin{aligned} |\varphi(t+h) - \varphi(t)| &= \left| \int_{\mathbb{R}^k} \left(\exp(i(t+h)'x) - \exp(it'x) \right) d\mu(x) \right| \\ &\leq \int_{\mathbb{R}^k} |\exp(it'x)| |\exp(ih'x) - 1| d\mu(x) = \int_{\mathbb{R}^k} |\exp(ih'x) - 1| d\mu(x). \end{aligned}$$

The last term does not depend on t and goes to 0 as $h \to \mathbf{0}$ by the BCT, so the uniform continuity of φ follows.

We now move onto the Riemann-Lebesgue Lemma.
Suppose $\mu << \lambda_k$. Then, by the Radon-Nikodym theorem, there exists a density f of μ with respect to λ_k. Then, for any $t \in \mathbb{R}^k$,

$$\varphi(t) = \int_{\mathbb{R}^k} \exp(it'x) f(x) dx.$$

Because $\int_{\mathbb{R}^k} f(x) dx = 1 < +\infty$, f is in $L^1(\mathbb{R}^k)$; by the continuity property of L^1 functions, for any $\epsilon > 0$ there exists a continuous compactly supported function g on $\mathbb{R}^k$ such that

$$\int_{\mathbb{R}^k} |f(x) - g(x)| dx < \frac{\epsilon}{2}.$$

Letting K be the compact support of g, because K is bounded (Heine-Borel theorem), there exists a k-cell $[-T, T]^k$ such that $K \subset [-T, T]^k$.

It now follows that

$$\begin{aligned}|\varphi(t)| &= \left|\int_{\mathbb{R}^k} \exp(it'x) f(x) dx\right| \\ &\leq \int_{\mathbb{R}^k} |f(x) - g(x)| dx + \left|\int_{\mathbb{R}^k} \exp(it'x) g(x) dx\right| \\ &< \frac{\epsilon}{2} + |\varphi_g(t)|,\end{aligned}$$

where we define

$$\varphi_g(t) = \int_{\mathbb{R}^k} \exp(it'x) g(x) dx$$

for any $t \in \mathbb{R}^k$. For any $t \neq \mathbf{0}$, by a linear change of variables we have

$$\begin{aligned}\int_{\mathbb{R}^k} \exp(it'x) g(x) dx &= \int_{\mathbb{R}^k} \exp\left(it'\left(x + \frac{t}{|t|^2}\pi\right)\right) g\left(x + \frac{t}{|t|^2}\pi\right) dx \\ &= -\int_{\mathbb{R}^k} \exp(it'x) g\left(x + \frac{t}{|t|^2}\pi\right) dx\end{aligned}$$

because $\exp(i\pi) = -1$. As such,

$$\begin{aligned}|\varphi_g(t)| &= \frac{1}{2}\left|\int_{\mathbb{R}^k} \exp(it'x) g(x) dx - \int_{\mathbb{R}^k} \exp(it'x) g\left(x + \frac{t}{|t|^2}\pi\right) dx\right| \\ &\leq \frac{1}{2}\int_{\mathbb{R}^k} \left|g(x) - g\left(x + \frac{t}{|t|^2}\pi\right)\right| dx.\end{aligned}$$

By the uniform continuity of g on $\mathbb{R}^k$, there exists a $\delta \in (0,1)$ such that

$$|g(x) - g(y)| < \frac{\epsilon}{(2T+2)^k}$$

for any $x, y \in \mathbb{R}^k$ such that $|x - y| < \delta$; because

$$\left|\frac{t}{|t|^2}\pi\right| = \frac{\pi}{|t|} \to 0,$$

as $|t| \to \infty$, there exists an $M > 0$ such that $\frac{\pi}{|t|} < \delta$ for any $|t| > M$.
Suppose that $x \notin [-T-1, T+1]^k$. It is immediately evident that $g(x) = 0$. Furthermore, for any $t \in \mathbb{R}^k$ such that $|t| > M$, because $\left|\frac{t_j}{|t|^2}\pi\right| < \delta < 1$ for any $1 \leq j \leq k$ and there exists a $1 \leq j \leq k$ such that

$$x_j < -T-1 \quad \text{or} \quad x_j > T+1,$$

we can see that

$$x_j + \frac{t_j}{|t|^2}\pi < x_j + 1 < -T \quad \text{or} \quad x_j + \frac{t_j}{|t|^2}\pi > x_j - 1 > T.$$

This means that $x+\frac{t}{|t|^2}\pi \notin [-T,T]^k$ and therefore that $g\left(x+\frac{t}{|t|^2}\pi\right)=0$.
We have thus seen that, for any $t\in\mathbb{R}^k$ such that $|t|>M$,

$$\begin{aligned}\int_{\mathbb{R}^k}\left|g(x)-g\left(x+\frac{t}{|t|^2}\pi\right)\right|dx &= \int_{[-T-1,T+1]^k}\left|g(x)-g\left(x+\frac{t}{|t|^2}\pi\right)\right|dx \\ &\leq \frac{\epsilon}{(2T+2)^k}\cdot\lambda_k\left([-T-1,T+1]^k\right)=\epsilon\end{aligned}$$

where the inequality follows from the uniform continuity of g and the fact that $\frac{\pi}{|t|}<\delta$ if $|t|>M$.
Therefore, for any $t\in\mathbb{R}^k$ such that $|t|>M$,

$$|\varphi(t)|<\frac{\epsilon}{2}+\frac{1}{2}\int_{\mathbb{R}^k}\left|g(x)-g\left(x+\frac{t}{|t|^2}\pi\right)\right|dx<\epsilon.$$

Such an $M>0$ exists for any $\epsilon>0$, so by definition

$$\lim_{|t|\to\infty}\varphi(t)=0.$$

Q.E.D.

Starting with the Fourier transform for finite measures, we can also define the Fourier transforms of random variables and measurable functions.
The Fourier transform of a random variable X taking values in $\mathbb{R}^k$ with distribution μ is then defined as the Fourier transform of μ.
The Fourier transform of a Borel-measurable, non-negative and Lebesgue integrable function f on $\mathbb{R}^k$, meanwhile, is defined as the Fourier transform of the indefinite integral of f with respect to the Lebesgue measure on $\mathbb{R}^k$. Specifically, let μ be the measure on $(\mathbb{R}^k,\mathcal{B}(\mathbb{R}^k))$ be defined as

$$\mu(A)=\int_A f(x)dx$$

for any $A\in\mathcal{B}(\mathbb{R}^k)$. Because f is integrable,

$$\mu(\mathbb{R}^k)=\int_{\mathbb{R}^k}f(x)dx<+\infty,$$

making μ a finite measure; this makes f the Radon-Nikodym derivative of μ with respect to λ_k. Then, the Fourier transform $\varphi:\mathbb{R}^k\to\mathbb{C}$ of f is defined as

$$\varphi(t)=\int_{\mathbb{R}^k}\exp(it'x)d\mu(x)=\int_{\mathbb{R}^k}\exp(it'x)f(x)dx$$

for any $t\in\mathbb{R}^k$.
Now let f be an arbitrary complex and Lebesgue integrable function on $\mathbb{R}^k$. By the integrability of f, $Re(f)^{\pm}$ and $Im(f)^{\pm}$ are measurable, non-negative and integrable functions on $\mathbb{R}^k$. Denoting

their Fourier transforms by $\varphi_\pm, \phi_\pm : \mathbb{R}^k \to \mathbb{C}$, the Fourier transform $\varphi : \mathbb{R}^k \to \mathbb{C}$ of f is defined as

$$\begin{aligned}\varphi(t) &= \varphi_+(t) - \varphi_-(t) + i\left(\phi_+(t) - \phi_-(t)\right) \\ &= \int_{\mathbb{R}^k} \exp(it'x) f(x) dx\end{aligned}$$

for any $t \in \mathbb{R}^k$. This is often taken to be the definition of the Fourier transform of arbitrary complex measurable functions on $\mathbb{R}^k$.
We will later derive the usual inversion results for the Fourier transforms of functions, rather than finite measures.

The Fourier transform is of great significance in probability theory because it determines the distribution of a random variable. That is, two random variables taking values in $\mathbb{R}^k$ with the same Fourier transform has the same distribution. This property is used to, later on, determine whether a set of random variables are independent, and whether a sequence or array of random variables converges in distribution.

1.6.1 The Dirichlet Integral

We first introduce an important integral that will be used to prove the inversion result.
Define $S : [0, +\infty) \to \mathbb{R}$ as

$$S(T) = \begin{cases} \int_0^T \frac{\sin(x)}{x} dx & \text{if } T > 0 \\ 0 & \text{if } T = 0 \end{cases}.$$

The integral is well-defined and finite because the integrand $\frac{\sin(x)}{x} \cdot I_{(0,T]}(x)$ is Borel measurable and bounded above by $I_{(0,T]}(x)$ ($\frac{\sin(x)}{x}$ is bounded above by 1), whose Lebesgue integral is finite and equal to T.
Unlike in the case where $\frac{\sin(x)}{x}$ is integrated over a finite interval $[0, T]$, its integral over $[0, +\infty)$ does not exist, since the integral of $\left|\frac{\sin(x)}{x}\right|$ over $[0, +\infty)$ is infinite. It does not even exist in the extended sense, since the integrals of both the positive and negative parts of $\frac{\sin(x)}{x}$ are infinite. It is, however, possible to obtain the limit of $S(T)$ as $T \to \infty$; this is a quantity distinct from the integral of $\frac{\sin(x)}{x}$ over $[0, +\infty)$ (although in the case of other integrands the two do agree). This limit is referred to as the Dirichlet integral, and is often denoted $\int_0^\infty \frac{\sin(x)}{x} dx$ (which represents a clear abuse of notation).

Lemma 1.8 Let $S : [0, +\infty) \to \mathbb{R}$ be defined as above. Then,

$$\lim_{T \to \infty} S(T) = \frac{\pi}{2}.$$

Proof) For any $x > 0$,

$$\frac{1}{x} = \int_0^\infty \exp(-tx)dt,$$

where the integral is over the interval $[0, +\infty)$. Therefore, for any $T > 0$, $S(T)$ can be written as

$$S(T) = \int_0^T \frac{\sin(x)}{x}dx = \int_0^T \int_0^\infty \exp(-tx)\sin(x)dtdx$$

by the linearity of integration ($\exp(-tx)$ is integrable with respect to t for any fixed $x > 0$). Because the integrand $\exp(-tx)\sin(x) \cdot I_{(0,T]}(x)$ is integrable with respect to the Lebesgue measure on $\mathbb{R}^2$, by Fubini's theorem we can interchange the order of integration to obtain

$$S(T) = \int_0^\infty \int_0^T \exp(-tx)\sin(x)dxdt.$$

For any $t > 0$, integration by parts tells us that

$$t \cdot \int_0^T \exp(-tx)cos(x)dx + \int_0^T \exp(-tx)sin(x)dx = 1 - \exp(-tT)\cos(T)$$

and

$$t \cdot \int_0^T \exp(-tx)sin(x)dx - \int_0^T \exp(-tx)cos(x)dx = -\exp(-tT)sin(T),$$

so we have

$$\int_0^T \exp(-tx)\sin(x)dx = \frac{1 - \exp(-tT)(t\sin(T) + \cos(T))}{t^2 + 1},$$

and because $\{0\}$ has Lebesgue measure 0,

$$\begin{aligned} S(T) &= \int_0^\infty \frac{1 - \exp(-tT)(t\sin(T) + \cos(T))}{t^2 + 1}dt \\ &= \int_0^\infty \frac{1}{t^2 + 1}dt - \int_0^\infty \frac{\exp(-tT)}{t^2 + 1}(t\sin(T) + \cos(T))\,dt. \end{aligned}$$

The first term is equal to $\frac{\pi}{2}$.
As for the latter term, define

$$g(t, T) = \frac{\exp(-tT)}{t^2 + 1}(t\sin(T) + \cos(T))$$

for any $T > 0$ and $t > 0$. It is evident that

$$|g(t, T)| \leq \frac{t + 1}{t^2 + 1}\exp(-t)$$

for any $t > 0$ and $T \geq 1$, and because

$$\frac{\partial}{\partial t}\left(\frac{t+1}{t^2+1}\exp(-t)\right) = -\frac{t(t^2+2t+3)\exp(-t)}{(t^2+1)^2} < 0$$

for any $t > 0$ and $\frac{t+1}{t^2+1}\exp(-t) = 1$ if $t = 0$, we can see that

$$\frac{t+1}{t^2+1}\exp(-t) \leq 1$$

for any $t \geq 0$. It follows that

$$\begin{aligned}\int_0^\infty \frac{t+1}{t^2+1}\exp(-t)dt &\leq \int_0^1 \frac{t+1}{t^2+1}\exp(-t)dt + 2\int_1^\infty \exp(-t)dt \\ &\leq 1 + 2\int_1^\infty \exp(-t)dt = 1 + 2\exp(-1) < +\infty.\end{aligned}$$

Finally, we know that

$$g(\cdot, T) \to 0$$

pointwise as $T \to \infty$.
Therefore, by the DCT, we have

$$\lim_{T\to\infty}\int_0^\infty g(t,T)dt = \int_0^\infty \left(\lim_{T\to\infty} g(t,T)\right)dt = 0.$$

We can now conclude that

$$\lim_{T\to\infty} S(T) = \frac{\pi}{2} + \lim_{T\to\infty}\int_0^\infty g(t,T)dt = \frac{\pi}{2}.$$

Q.E.D.

The result above tells us that the function S must be bounded on $[0, +\infty)$ (otherwise, there exists a subsequence $\{t_n\}_{n\in N_+}$ of N_+ such that $S(t_n) \to +\infty$ as $n \to \infty$, which is a contradiction). Thus, there exists an $M > 0$ such that $|S(T)| \leq M$ for any $T \geq 0$.

1.6.2 k-Cells with Vertices of Measure 0

Another important component to the inversion formula is the structure of euclidean k-space, and how the set of all k-cells is a π-system that generates the Borel σ-algebra on $\mathbb{R}^k$. We now introduce an additional wrinkle; it is not only the collection of all k-cells that generates the $\mathcal{B}(\mathbb{R}^k)$, but also any collection of open k-cells whose boundaries have measure 0 under some finite measure μ. The inversion formula holds for precisely these k-cells, so the formula, combined with the fact that these cells generates the Borel σ-algebra on $\mathbb{R}^k$, allows us to make a claim about the distributions of random variables in $\mathbb{R}^k$.
We state below the salient result concerning k-cells on $\mathbb{R}^k$. The proof is very detailed so as

to present and draw attention to the most important properties of euclidean spaces and how powerful they can be.

Lemma 1.9 Let μ, v be finite measures on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$, and define

$$\mathcal{P} = \left\{ \prod_{j=1}^{k} (a_j, b_j) \middle| \forall j, a_j < b_j, \text{ and } \mu\left(\prod_{j=1}^{k} \{a_j, b_j\}\right) = v\left(\prod_{j=1}^{k} \{a_j, b_j\}\right) = 0 \right\} \cup \{\emptyset\},$$

that is, as the collection of all open k-cells on $\mathbb{R}^k$ whose endpoint have measure 0 under μ and v. Then, $\mathcal{D}$ is a π-system that generates $\mathcal{B}(\mathbb{R}^k)$.

Proof) We first show that there are at most countably many points $x \in \mathbb{R}^k$ such that $\mu(\{x\}) > 0$; we present the standard argument for this claim. Let $m = \lceil \mu(\mathbb{R}^k) \rceil \in N_+$, and A the set of all points on $\mathbb{R}^k$ such that $\mu(\{x\}) > 0$. A can then be expressed as the union

$$A = \bigcup_n \underbrace{\left\{ x \in \mathbb{R}^k \mid \mu(\{x\}) > \frac{1}{n} \right\}}_{A_n}.$$

For any $n \in N_+$, suppose $|A_n| > nm$. Then, there exists a finite subset J of A_n with cardinality exactly nm, so that, by finite additivity,

$$m \geq \mu(A_n) \geq \mu(J) = \sum_{x \in J} \mu(\{x\}) > |J| \frac{1}{n} = m,$$

which is a contradiction. It follows that A_n contains at most nm elements, and because this holds for any $n \in N_+$, A is at most a countably infinite set.
Likewise, there are at most countably many points $x \in \mathbb{R}^k$ such that $v(\{x\}) > 0$. This means that there are at most countably many values $a_1, \cdots, a_k, b_1, \cdots, b_k \in \mathbb{R}$ such that $a_i < b_i$ for $1 \leq i \leq k$ and

$$\mu\left(\prod_{j=1}^{k} \{a_j, b_j\}\right) > 0 \quad \text{or} \quad v\left(\prod_{j=1}^{k} \{a_j, b_j\}\right) > 0.$$

Using this property, we can define a countable set of points on $\mathbb{R}^k$, which we denote $\mathbb{Q}^{k\prime}$. Specifically, for any $q \in \mathbb{R}^k$ and $n \in N_+$, define

$$W(q, n) = \prod_{j=1}^{k} (q_j - 2^{-n}, q_j + 2^{-n}),$$

where $n \in N_+$ and $q \in \mathbb{Q}^k$. The collection of all such k-cells is countable because $\mathbb{Q}^k$ and N_+ are.
For any $q \in \mathbb{Q}^k$ and $n \in N_+$, we define $W(q, n)' = W(q, n)$ if the vertices of $W(q, n)$ have

measure 0 under both μ and v. On the other hand, suppose that at least one of the vertices of $W(q,n)$ has positive measure under μ or v, we choose a $\delta \in (0, 2^{-n-1})$ so that the vertices of

$$W(q,n)' = \prod_{j=1}^{k} (q_j - 2^{-n} + \delta, q_j + 2^{-n} - \delta)$$

have measure 0 under both μ and v. The key idea is that such a $\delta > 0$ exists because there exists at most countably many points x on $\mathbb{R}^k$ such that $\mu(\{x\}) > 0$ or $v(\{x\}) > 0$. Due to our choice of δ, this k-cell has diameter between 2^{-n+1} and 2^{-n}, thus retaining its unique position among other similar k-cells.
Ultimately, the collection of all cells $W(q,n)'$ constructed as such is countably infinite, and the vertices of each cell have measure 0 under both μ and v. Denote the collection of all such cells as $\mathcal{W}'$.

Now we show that $\mathcal{P}$ is a π-system that generates $\mathcal{B}(\mathbb{R}^k)$. To show that $\mathcal{P}$ is a π-system, choose any two cells $A = \prod_{j=1}^{k}(a_j, b_j)$ and $B = \prod_{j=1}^{k}(c_j, d_j)$ in $\mathcal{P}$. By definition,

$$\mu\left(\prod_{j=1}^{k}\{a_j, b_j, c_j, d_j\}\right) = v\left(\prod_{j=1}^{k}\{a_j, b_j, c_j, d_j\}\right) = 0$$

If $b_j \leq c_j$ or $d_j \leq a_j$ for at least one $1 \leq j \leq k$, then $A \cap B = \emptyset \in \mathcal{P}$. As such, assume that $(a_j, b_j) \cap (c_j, d_j) \neq \emptyset$ for all $1 \leq j \leq k$. Because $(a_j, b_j) \cap (c_j, d_j)$ is also an open interval in this case, $A \cap B$ is an open k-cell. Furthermore, the endpoints of each interval are chosen from the set $\{a_j, b_j, c_j, d_j\}$, and since any point on $\mathbb{R}^k$ with these endpoints has measure 0 under both μ and v, $A \cap B$ is contained in $\mathcal{P}$. It follows that $\mathcal{P}$ is a π-system.

Moving onto showing that $\mathcal{P}$ generates $\mathcal{B}(\mathbb{R}^k)$, it is initially clear that, because each k-cell in $\mathcal{P}$ is Borel measurable, the σ-algebra generated by $\mathcal{P}$ is contained in $\mathcal{B}(\mathbb{R}^k)$.
To show the reverse inclusion, choose any open subset V of $\mathbb{R}^k$, and a point $x \in V$. Because V is open, there exists an $\epsilon > 0$ such that $B(x, \epsilon)$, the ϵ-ball around x under the euclidean metric, is contained in V. Choose $N \in N_+$ so that $2^{-N} < \frac{1}{\sqrt{k}} \frac{\epsilon}{2}$. Because $\mathbb{Q}^k$ is dense in $\mathbb{R}^k$, there exists a $q \in \mathbb{Q}^k$ such that $|x - q| < 2^{-N-1}$, and by implication $|x_j - q_j| < 2^{-N-1}$ for any $1 \leq j \leq k$.
The k-cell

$$W(q,N) = \prod_{j=1}^{k} (q_j - 2^{-N}, q_j + 2^{-N})$$

clearly contains x (in fact, it is contained in the smaller k-cell $W(q, N+1)$). Further-

more, $W(q,N)$ is contained in $B(x,\epsilon)$ and by implication V, since, for any $z \in W(q,N)$,

$$|z-x| \leq |q-x| + |q-z| < 2^{-N-1} + \left(\sum_{j=1}^{k} (q_j - z_j)^2 \right)^{\frac{1}{2}}$$
$$< \frac{\epsilon}{2} + \sqrt{k} \cdot 2^{-N} < \epsilon.$$

The k-cell $W(q,N)'$ in $\mathcal{W}'$ corresponding to $W(q,N)$ is contained in $W(q,N)$, so we have $W(q,N)' \subset V$. Furthermore, $W(q,N)'$ contains $W(q,N+1)$, in which x is contained. Therefore, we have

$$x \in W(q,N)' \subset V.$$

This holds for any $x \in V$, so V is the union of k-cells in $\mathcal{W}'$; $\mathcal{W}'$ is a countable collection of sets, so this means that V is a countable union of sets in $\mathcal{W}'$. Finally, each set in $\mathcal{W}'$ is an element of $\mathcal{P}$, which implies that V is an element of the σ-algebra generated by $\mathcal{P}$.

We have shown that the euclidean topology $\tau_{\mathbb{R}^k}$ on $\mathbb{R}^k$ is contained in $\sigma\mathcal{P}$. Since $\mathcal{B}(\mathbb{R}^k)$ is generated by $\tau_{\mathbb{R}^k}$, we finally have

$$\mathcal{B}(\mathbb{R}^k) \subset \sigma\mathcal{P},$$

which, combined with our previous observations, means that $\mathcal{P}$ is a π-system that generates the Borel σ-algebra on $\mathbb{R}^k$.
Q.E.D.

1.6.3 The Inversion Formula

We now state the inversion formula and its corollary, which are the main results of this section:

Theorem 1.10 (Lévy's Inversion Formula)
Let μ be a finite measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ with characteristic function $\varphi : \mathbb{R}^k \to \mathbb{C}$. Then, for any open k-cell $W = \prod_{j=1}^{k} (a_1, b_1)$, it holds that

$$\mu(W) + \frac{1}{2}\mu\left(\prod_{j=1}^{k} \{a_j, b_j\} \right) = \lim_{T \to \infty} \frac{1}{(2\pi)^k} \int_{[-T,T]^k} \left(\prod_{j=1}^{k} \frac{\exp(-ia_j t_j) - \exp(-ib_j t_j)}{it_j} \right) \varphi(t) dt.$$

Proof) The characteristic function φ is given by

$$\varphi(t) = \int_{\mathbb{R}^k} \exp(it'x) d\mu(x) = \int_{\mathbb{R}^k} \prod_{j=1}^{k} \exp(it_j x_j) d\mu(x)$$

for any $t \in \mathbb{R}^k$, so

$$\left(\prod_{j=1}^{k} \frac{\exp(-ia_jt_j) - \exp(-ib_jt_j)}{it_j}\right)\varphi(t) = \int_{\mathbb{R}^k} \prod_{j=1}^{k} \frac{\exp(it_j(x_j - a_j)) - \exp(it_j(x_j - b_j))}{it_j} d\mu(x)$$

for each k-dimensional vector t (if $t_j = 0$ for some j, then the fraction above is taken to be equal to $b_j - a_j$, its limit at 0 by L'Hospital's Rule). For any $1 \leq j \leq k$, note that

$$|\exp(it_j(x_j - a_j)) - \exp(it_j(x_j - b_j))| \leq t_j(b_j - a_j)$$

by the mean value theorem (applied to the function $z \mapsto \exp(iz)$ on $\mathbb{R}$), so that

$$\left|\frac{\exp(it_j(x_j - a_j)) - \exp(it_j(x_j - b_j))}{it_j}\right| \leq b_j - a_j.$$

For any $T > 0$, define

$$g(t,x) = \prod_{j=1}^{k} \frac{\exp(it_j(x_j - a_j)) - \exp(it_j(x_j - b_j))}{it_j} \cdot I_{[-T,T]^k}$$

for any $t, x \in \mathbb{R}^k$. Because g is bounded by $\prod_{j=1}^{k}(b_j - a_j) < +\infty$, μ is a finite measure and the k-dimensional Lebesgue measure λ_k is bounded on the k-cell $[-T,T]^k$, g is integrable with respect to the product measure $\mu \times \lambda_k$, and Fubini's theorem tells us that

$$\begin{aligned}\int_{[-T,T]^k} \left(\prod_{j=1}^{k} \frac{\exp(-ia_jt_j) - \exp(-ib_jt_j)}{it_j}\right)\varphi(t)dt &= \int_{\mathbb{R}^k}\int_{\mathbb{R}^k} g(t,x)d\mu(x)dt \\ &= \int_{\mathbb{R}^k}\int_{\mathbb{R}^k} g(t,x)dtd\mu(x).\end{aligned}$$

For any $x \in \mathbb{R}^k$,

$$\begin{aligned}\int_{\mathbb{R}^k} g(t,x)dt &= \int_{[-T,T]^k} \prod_{j=1}^{k} \frac{\exp(it_j(x_j - a_j)) - \exp(it_j(x_j - b_j))}{it_j} dt \\ &= \prod_{j=1}^{k} \left(\int_{-T}^{T} \frac{\exp(it_j(x_j - a_j)) - \exp(it_j(x_j - b_j))}{it_j} dt_j\right)\end{aligned}$$

once again by Fubini's theorem (the integrand is bounded, and the Lebesgue measure is finite on a compact set).
Using Euler's formula to expand the complex exponentials, for each $1 \leq j \leq k$ we have

$$\begin{aligned}&\int_{-T}^{T} \frac{\exp(it_j(x_j - a_j)) - \exp(it_j(x_j - b_j))}{it_j} dt_j \\ &= \int_{-T}^{T} \frac{1}{it_j}\left(\cos(t_j(x_j - a_j)) + i\sin(t_j(x_j - a_j))\right)dt_j - \int_{-T}^{T} \frac{1}{it_j}\left(\cos(t_j(x_j - b_j)) + i\sin(t_j(x_j - b_j))\right)dt_j\end{aligned}$$

$$= 2\int_0^T \frac{\sin(t_j(x_j-a_j)) - \sin(t_j(x_j-b_j))}{t_j} dt_j,$$

where the last equality follows from the fact that $\cos(\cdot)$ is an even function. Since the first term can be expressed as

$$\int_0^T \frac{\sin(t_j(x_j-a_j))}{t_j} dt_j = sgn(x_j-a_j)\cdot \int_0^{T|x_j-a_j|} \frac{\sin(z)}{z} dz = sgn(x_j-a_j)\cdot S(T|x_j-a_j|)$$

and likewise for the second term, where $sgn(\cdot)$ is a function that yields the sign of the argument, the above integral can now be written In terms of the function $S(\cdot)$ as

$$\int_{-T}^T \frac{\exp(it_j(x_j-a_j)) - \exp(it_j(x_j-b_j))}{it_j} dt_j$$
$$= 2\left[sgn(x_j-a_j)\cdot S(T|x_j-a_j|) - sgn(x_j-b_j)\cdot S(T|x_j-b_j|)\right].$$

So far, it has been shown that

$$\frac{1}{(2\pi)^k}\int_{[-T,T]^k} \left(\prod_{j=1}^k \frac{\exp(-ia_jt_j)-\exp(-ib_jt_j)}{it_j}\right)\varphi(t)dt$$
$$= \int_{\mathbb{R}^k} \prod_{j=1}^k \underbrace{\left(sgn(x_j-a_j)\cdot\frac{S(T|x_j-a_j|)}{\pi} - sgn(x_j-b_j)\cdot\frac{S(T|x_j-b_j|)}{\pi}\right)}_{g_{T,j}(x_j)} d\mu(x).$$

Because $S(\cdot)$ is bounded above by $M\in(0,+\infty)$, the integrand in the integral on the right is bounded above by $\left(\frac{2M}{\pi}\right)^k < +\infty$. Furthermore, in light of the fact that $\lim_{T\to\infty} S(T) = \frac{\pi}{2}$, as $T\to\infty$, each $g_{T,j}$ converges pointwise to the function $g_j:\mathbb{R}\to\mathbb{R}$ defined as

$$g_j = \frac{1}{2}\cdot I_{\{a_j,b_j\}} + I_{(a_j,b_j)}.$$

Finally, μ is a finite measure, so by the BCT,

$$\lim_{T\to\infty}\frac{1}{(2\pi)^k}\int_{[-T,T]^k}\left(\prod_{j=1}^k \frac{\exp(-ia_jt_j)-\exp(-ib_jt_j)}{it_j}\right)\varphi(t)dt = \int_{\mathbb{R}^k}\left(\prod_{j=1}^k g_j(x_j)\right)d\mu(x)$$
$$= \frac{1}{2}\mu\left(\prod_{j=1}^k\{a_j,b_j\}\right) + \mu\left(\prod_{j=1}^k(a_j,b_j)\right) = \frac{1}{2}\mu\left(\prod_{j=1}^k\{a_j,b_j\}\right) + \mu(W).$$

Q.E.D.

The inversion formula, together with the π-system for $\mathcal{B}(\mathbb{R}^k)$ furnished in lemma 1.9, directly lead to the next result:

Corollary to Theorem 1.10 Let μ, v be two proability measures on $(\mathbb{R}^k,\mathcal{B}(\mathbb{R}^k))$ with characteristic functions $\varphi,\phi:\mathbb{R}^k\to\mathbb{C}$. If $\varphi(t)=\phi(t)$ for every $t\in\mathbb{R}^k$, then $\mu = v$ on $\mathcal{B}(\mathbb{R}^k)$.

Proof) Let $\mathcal{P}$ be defined as in lemma 1.9. Then, for any $W = \prod_{j=1}^{k}(a_j, b_j) \in \mathcal{P}$, because

$$\mu\left(\prod_{j=1}^{k}\{a_j, b_j\}\right) = v\left(\prod_{j=1}^{k}\{a_j, b_j\}\right) = 0,$$

by the inversion formula we have

$$\mu(W) = \lim_{T\to\infty}\frac{1}{(2\pi)^k}\int_{[-T,T]^k}\left(\prod_{j=1}^{k}\frac{\exp(-ia_jt_j)-\exp(-ib_jt_j)}{it_j}\right)\varphi(t)dt$$
$$v(W) = \lim_{T\to\infty}\frac{1}{(2\pi)^k}\int_{[-T,T]^k}\left(\prod_{j=1}^{k}\frac{\exp(-ia_jt_j)-\exp(-ib_jt_j)}{it_j}\right)\phi(t)dt.$$

Since $\varphi = \phi$ on $\mathbb{R}^k$, it follows that

$$\mu(W) = v(W).$$

Thus, μ and v are measures that agree on the π-system $\mathcal{P}$ generating $\mathcal{B}(\mathbb{R}^k)$, and $\mu(\mathbb{R}^k) = v(\mathbb{R}^k) = 1 < +\infty$. It follows that $\mu = v$ on $\mathcal{B}(\mathbb{R}^k)$.
Q.E.D.

1.6.4 The Cramer-Wold Device

The most useful and frequently utilized result that stems from the inversion formula is the Cramer-Wold device, which comes in two forms; we state here its first form, which tells us that the distribution of a random vector is determined by the linear combination of its components.

Theorem 1.11 (Cramer-Wold Device I)
Let X, Y be two random vectors taking values in $\mathbb{R}^k$ with distributions μ and v.
$X \sim Y$ if and only if $r'X \sim r'Y$ for any non-zero $r \in \mathbb{R}^k$.

Proof) We first show necessity. Suppose that $X \sim Y$, that is, $\mu = v$ on $\mathcal{B}(\mathbb{R}^k)$. Choose some non-zero $r \in \mathbb{R}^k$, and define $f : \mathbb{R}^k \to \mathbb{R}$ as $f(x) = r'x$ for any $x \in \mathbb{R}^k$. Then, $r'X = f \circ X$, $r'Y = f \circ Y$, so that the distributions of $r'X$ and $r'Y$ are given as $\mu \circ f^{-1}$ and $v \circ f^{-1}$. Since $\mu = v$, this means that $r'X$ and $r'Y$ both have distributions $\mu \circ f^{-1}$, and as such, that they are identically distributed.

To show sufficiency, let $\varphi, \phi : \mathbb{R}^k \to \mathbb{C}$ be the characteristic functions of X and Y; by definition,

$$\varphi(t) = \int_{\mathbb{R}^k}\exp(-it'x)d\mu(x) \quad \text{and} \quad \phi(t) = \int_{\mathbb{R}^k}\exp(-it'x)dv(x)$$

for any $t \in \mathbb{R}^k$.
Suppose that $r'X \sim r'Y$ for any non-zero $r \in \mathbb{R}^k$. Choosing some non-zero $r \in \mathbb{R}^k$, define

f as above. Then, $f \circ X \sim f \circ Y$, which tells us that the characteristic functions of $f \circ X$ and $f \circ Y$ are the same. Denoting these functions by φ_r and ϕ_r, and using the fact that the distributions of $f \circ X$ and $f \circ Y$ are the pushforward measures $\mu \circ f^{-1}$ and $v \circ f^{-1}$, by theorem 1.1 we have

$$\begin{aligned}\varphi_r(t) = \int_{\mathbb{R}} \exp(itz) d(\mu \circ f^{-1})(z) &= \int_{\mathbb{R}^k} \exp(itf(x)) d\mu(x) \\ &= \int_{\mathbb{R}^k} \exp(itr'x) d\mu(x) = \varphi(tr)\end{aligned}$$

for any $t \in \mathbb{R}$. Likewise, $\phi_r(t) = \phi(tr)$ for any $t \in \mathbb{R}$. Putting $t = 1$ and making use of the fact that $\varphi_r = \phi_r$ on $\mathbb{R}$ yields

$$\varphi(r) = \varphi_r(1) = \phi_r(1) = \phi(r).$$

This holds for any non-zero $r \in \mathbb{R}^k$, and $\varphi(\mathbf{0}) = 1 = \phi(\mathbf{0})$, so by the inversion formula and its corollary, $\mu = v$ on $\mathcal{B}(\mathbb{R}^k)$, that is, X and Y are identically distributed.
Q.E.D.

The Cramer-Wold device thus allows us to make claims about the distribution of a random vector using only its projections onto a one-dimensional subspace. This property often comes in handy, as will be illustrated in the section on the multivariate normal distribution.

1.6.5 The Fourier Transform of Functions

As stated at the beginning of this section, we can opt to start from the Fourier transform of integrable functions rather than finite measures. We derive here the inversion formula for the Fourier transform of functions, which expresses them as functions of their characteristic functions.

Theorem 1.12 (Inverse Fourier Transform)
Let μ be a finite measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ with characteristic function $\varphi : \mathbb{R}^k \to \mathbb{C}$.
If φ is Lebesgue-integrable, then the unique continuous density f of μ with respect to the Lebesgue measure is defined as

$$f(x) = \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \exp(-it'x)\varphi(t) dt$$

for any $x \in \mathbb{R}^k$.

Proof) Suppose that φ is Lebesgue-integrable, that is,

$$\int_{\mathbb{R}^k} |\varphi(t)| dt < +\infty.$$

Choose any open k-cell $\prod_{j=1}^{k}(a_j, b_j)$ on $\mathbb{R}^k$, and define $g : \mathbb{R} \to \mathbb{C}$ as

$$g(t) = \left(\prod_{j=1}^{k} \frac{\exp(-ia_j t_j) - \exp(-ib_j t_j)}{it_j} \right) \varphi(t)$$

for any $t \in \mathbb{R}^k$. The sequence $\{g \cdot I_{[-T,T]^k}\}_{T>0}$ is bounded above by $\left(\prod_{j=1}^{k}(b_j - a_j)\right)|\varphi|$, which is Lebesgue integrable. Since $g \cdot I_{[-T,T]^k}$ converges pointwise to g, by the DCT g is Lebesgue integrable and

$$\lim_{T\to\infty} \int_{[-T,T]^k} g(t)dt = \int_{\mathbb{R}^k} g(t)dt.$$

By the inversion formula, we now have

$$\mu\left(\prod_{j=1}^{k}(a_j, b_j) \right) + \frac{1}{2}\mu\left(\prod_{j=1}^{k}\{a_j, b_j\} \right) = \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} g(t)dt.$$

Since $|g| \leq \left(\prod_{j=1}^{k}(b_j - a_j)\right)|\varphi|$, by the monotonicity of integration,

$$\begin{aligned} \frac{1}{2}\mu\left(\prod_{j=1}^{k}\{a_j, b_j\} \right) &\leq \mu\left(\prod_{j=1}^{k}(a_j, b_j) \right) + \frac{1}{2}\mu\left(\prod_{j=1}^{k}\{a_j, b_j\} \right) \\ &= \left| \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} g(t)dt \right| \leq \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} |g(t)|dt \\ &\leq \frac{1}{(2\pi)^k} \left(\prod_{j=1}^{k}(b_j - a_j) \right) \int_{\mathbb{R}^k} |\varphi(t)|dt. \end{aligned}$$

Sending $b \searrow a$ on both sides, by sequential continuity $\mu\left(\prod_{j=1}^{k}\{a_j, b_j\}\right) \to \mu(\{a\})$, while the right hand side converges to 0. This implies that $\mu(\{a\}) = 0$; because $a \in \mathbb{R}$ was chosen arbitrarily, every singleton has measure 0 under μ.

Let $F : \mathbb{R}^k \to [0, 1]$ be defined as

$$F(x) = \mu\left(\prod_{j=1}^{k}(-\infty, x_j) \right)$$

for any $x \in \mathbb{R}^k$. For purposes of illustration, here we assume that $k = 2$. In this case, for any $x \in \mathbb{R}^k$ and non-zero $h \in \mathbb{R}^k$, we have

$$\mu((x_1, x_1 + h_1) \times (x_2, x_2 + h_2)) = F(x_1 + h_1, x_2 + h_2) - F(x_1, x_2 + h_2) - F(x_1 + h_1, x_2) + F(x_1, x_2)$$

in the case that $h_1, h_2 > 0$. By the inversion formula, we have

$$\frac{F(x_1 + h_1, x_2 + h_2) - F(x_1, x_2 + h_2) - F(x_1 + h_1, x_2) + F(x_1, x_2)}{h_1 h_2}$$

$$= \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \left(\prod_{j=1}^{k} \frac{\exp(-it_j x_j) - \exp(-it_j(x_j + h_j))}{it_j h_j} \right) \varphi(t) dt,$$

and it can be shown that the above equation holds even when $h_1 < 0$ or $h_2 < 0$. The integrand in the integral on the right is bounded above as

$$\left| \left(\prod_{j=1}^{k} \frac{\exp(-it_j x_j) - \exp(-it_j(x_j + h_j))}{it_j h_j} \right) \varphi(t) \right| \leq |\varphi(t)|,$$

and

$$\lim_{h_1 \to 0} \lim_{h_2 \to 0} \left(\prod_{j=1}^{k} \frac{\exp(-it_j x_j) - \exp(-it_j(x_j + h_j))}{it_j h_j} \right) \varphi(t)$$
$$= -\prod_{j=1}^{k} \left[\frac{1}{it_j} \left(\frac{\partial}{\partial x_j} \exp(-it_j x_j) \right) \right] \varphi(t) = \exp(-it'x)\varphi(t)$$

for any $t \in \mathbb{R}^k$, so by repeated applications of the DCT,

$$\frac{\partial^2 F(x)}{\partial x_1 \partial x_2} = \lim_{h_1 \to 0} \lim_{h_2 \to 0} \frac{F(x_1 + h_1, x_2 + h_2) - F(x_1, x_2 + h_2) - F(x_1 + h_1, x_2) + F(x_1, x_2)}{h_1 h_2}$$
$$= \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \exp(-it'x)\varphi(t) dt.$$

This result is easily generalized to the case of $k > 2$.

Define $f : \mathbb{R}^k \to \mathbb{R}$ as

$$f(x) = \frac{\partial^k F(x)}{\partial x_1 \cdots \partial x_k}$$

for any $x \in \mathbb{R}^k$. Because F is increasing in each coordinate of x (by the monotonicity of measures), f is a non-negative function. It is also continuous on $\mathbb{R}^k$: to see this, choose any $x \in \mathbb{R}^k$ and a sequence $\{x_n\}_{n \in N_+} \subset \mathbb{R}^k$ converging to x. For any $n \in N_+$,

$$f(x_n) = \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \exp(-it'x_n)\varphi(t) dt.$$

Since $|\exp(-it'x_n)\varphi(t)| \leq |\varphi(t)|$ for any $t \in \mathbb{R}$ and $n \in N_+$, and $\exp(-it'x_n)\varphi(t) \to \exp(-it'x)\varphi(t)$ as $n \to \infty$, by the DCT we have

$$\lim_{n \to \infty} f(x_n) = \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \exp(-it'x)\varphi(t) dt = f(x);$$

it follows that f is continuous at x, and, indeed, on $\mathbb{R}^k$.
Define μ_0 as the indefinite integral of f with respect to the Lebesgue measure (which exists because f is a non-negative Borel measurable function). Then, for any open k-cell

$W = \prod_{j=1}^{k}(a_j, b_j)$, we have

$$\begin{aligned}
\mu(W) &= F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2) \\
&= \int_{a_2}^{b_2} \frac{\partial F(b_1, x_2)}{\partial x_2} dx_2 - \int_{a_2}^{b_2} \frac{\partial F(a_1, x_2)}{\partial x_2} dx_2 \\
&= \int_{a_2}^{b_2} \left(\frac{\partial F(b_1 x_2)}{\partial x_2} - \frac{\partial F(a_1, x_2)}{\partial x_2} \right) dx_2 \\
&= \int_{a_2}^{b_2} \int_{a_1}^{b_1} \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} dx_1 dx_2 \\
&= \int_W f(x) dx = \mu_0(W)
\end{aligned}$$

by the fundamental theorem of calculus (we put $k = 2$ for the purposes of illustration) and Fubini's theorem, which can be applied here thanks to the non-negativity of f. Choosing $W_n = [-n, n]^k$ for any $n \in N_+$, this result tells us that $\mu(W_n) = \mu_0(W_n)$ for any $n \in N_+$, and because $W_n \nearrow \mathbb{R}^k$ as $n \to \infty$, by sequential continuity

$$\mu(\mathbb{R}^k) = \mu_0(\mathbb{R}^k) = \int_{\mathbb{R}^k} f(x) dx < +\infty$$

by the finiteness of μ. Since the collection of open k-cells forms a π-system generating $\mathcal{B}(\mathbb{R}^k)$, we have $\mu = \mu_0$ on $\mathcal{B}(\mathbb{R}^k)$, that is, f is the Radon-Nikodym derivative of μ with respect to the Lebesgue measure, and in fact the only continuous version of the Radon-Nikodym derivative.
Q.E.D.

The above construction yields an alternative approach to the Fourier transform and its inverse on $\mathbb{R}^k$. Instead of starting with a finite measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$, we can start from a non-negative continuous Lebesgue integrable function f on $\mathbb{R}^k$. To illustrate how the two approaches are in fact equivalent, we first state the properties of the Fourier transform we have derived so far when starting from finite measures.
Throughout, let μ be a finite measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ and $\varphi : \mathbb{R}^k \to \mathbb{C}$ its Fourier transform.

- **Definition**
  For any $t \in \mathbb{R}^k$,

$$\varphi(t) = \int_{\mathbb{R}^k} \exp(it'x) d\mu(x).$$

- **Continuity**
  φ is uniformly continuous on $\mathbb{R}^k$.

- **The Riemann-Lebesgue Lemma**
  If μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^k$, then φ vanishes at infinity:

$$\varphi(t) \to 0 \quad \text{as} \quad |t| \to \infty.$$

- **The Inversion Formula**
  For any open k-cell $W = \prod_{j=1}^k (a_j, b_j)$,

$$\mu(W) + \frac{1}{2}\mu\left(\prod_{j=1}^k \{a_j, b_j\}\right) = \lim_{T\to\infty} \frac{1}{(2\pi)^k} \int_{[-T,T]^k} \left(\prod_{j=1}^k \frac{\exp(-ia_jt_j) - \exp(-ib_jt_j)}{it_j}\right) \varphi(t) dt.$$

  If φ is Lebesgue integrable, then the formula reduces to

$$\mu(W) = \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \left(\prod_{j=1}^k \frac{\exp(-ia_jt_j) - \exp(-ib_jt_j)}{it_j}\right) \varphi(t) dt.$$

We now state the properties of the Fourier transform of continuous and integrable functions, and how they can be deduced from that of the transform for finite measures.
Throughout, let $f : \mathbb{R}^k \to \mathbb{R}$ be a continuous non-negative Lebesgue integrable function on $\mathbb{R}^k$ with Fourier transform $\varphi : \mathbb{R}^k \to \mathbb{C}$.

- **Definition**
  For any $t \in \mathbb{R}^k$,

  $$\varphi(t) = \int_{\mathbb{R}^k} \exp(it'x) f(x) dx.$$

  Letting μ be the indefinite integral of f with respect to the Lebesgue measure, φ is the Fourier transform of μ, and can be written as $\varphi(T) = \int_{\mathbb{R}^k} \exp(it'x) d\mu(x)$ for any $t \in \mathbb{R}^k$.

- **Continuity**
  Because φ is the Fourier transform of the finite measure μ, it is uniformly continuous on $\mathbb{R}^k$.

- **Riemann-Lebesgue Lemma**
  Because μ has Radon-Nikodym derivative f with respect to the Lebesgue measure, μ is absolutely continuous with respect to the Lebesgue measure. By implication, its Fourier transform φ vanishes at infinity:

  $$\varphi(t) \to 0 \quad \text{as} \quad |t| \to \infty.$$

- **Inverse Fourier Transform**
  If $\int_{\mathbb{R}^k} |\varphi(t)| d_t < +\infty$, then by theorem 1.12, the unique continuous density f of μ with respect to the Lebesgue measure can be expressed as

  $$f(x) = \frac{1}{(2\pi)^k} \int_{\mathbb{R}^k} \exp(-it'x) \varphi(t) dt$$

  for any $x \in \mathbb{R}^k$.

The case for general complex-valued f can be easily deduced by making use of the construction of the Fourier transform of f from those of $Re(f)^{\pm}$ and $Im(f)^{\pm}$.

1.7 Multidimensional Distributions and Independence

We now generalize some of the concepts introduced in the previous section to random variables taking values in product spaces (these are often called random vectors, processes, or functions, depending on the cardinality of the product spaces in question).

Let T be an arbitrary index set and $\{(E_t, \mathcal{E}_t) \mid t \in T\}$ a collection of measurable spaces. Denote the product of these spaces by $(E, \mathcal{E}) = \bigotimes_{t \in T}(E_t, \mathcal{E}_t)$, where $\mathcal{E}$ is the σ-algebra generated by the π-system of measurable rectangles

$$\left\{ \prod_{t \in T} A_t \mid \forall t \in T, A_t \in \mathcal{E}_t, \text{ and only finitely many } A_t \text{ are not } E_t \right\}$$

on $E = \prod_{t \in T} E_i$. Let X be a random variable taking values in $(E, \mathcal{E})$. Letting X_t be the coordinate of X corresponding to $(E_t, \mathcal{E}_t)$, we can also write $X = (X_t)_{t \in T}$, that is, as a (possibly uncountable) sequence of random variables. It is clear that X is a random variable taking values in $(E, \mathcal{E})$ if and only if each X_t is a random variable in $(E_t, \mathcal{E}_t)$.

1.7.1 Joint and Marginal Distributions

Letting μ be the distribution of X, for any $t \in T$ the distribution μ_t of X_t satisfies

$$\mu_t(A_t) = \mathbb{P}(X_t \in A_t) = \mathbb{P}(X \in A_t \times E_{-t}) = \mu(A_t \times E_{-t}),$$

where $A_t \times E_{-t} = \prod_{s \in T} A_s$ for $A_s = E_s$ for any $s \in T$ such that $s \neq t$.
μ is referred to as the joint distribution of X, and each μ_t as the marginal distribution of X_t.

1.7.2 Joint and Marginal Densities

Suppose $T = \{1, \cdots, n\}$ for some $n \in N_+$ and that μ is absolutely continuous with respect to the product measure $v = v_1 \times \cdots \times v_n$ on $(E, \mathcal{E})$, where each v_i is a σ-finite measure on $(E_i, \mathcal{E}_i)$. Letting f be the density of $X = (X_1, \cdots, X_n)'$ with respect to v, we can see that, for any $A_i \in \mathcal{E}_i$,

$$\mu_i(A_i) = \int_{A_i} \left(\int_{E_{-i}} f(x_i, x_{-i}) dv_{-i}(x_{-i}) \right) dv_i(x_i)$$

by Fubini's theorem (f is a non-negative function). Since the mapping $x_i \mapsto \int_{E_{-i}} f(x_i, x_{-i}) dv_{-i}(x_{-i})$ defines a $\mathcal{E}_i$-measurable non-negative function, by definition

$$f_i = \int_{E_{-i}} f(\cdot, x_{-i}) dv_{-i}(x_{-i})$$

is the density of X_i with respect to v_i.
f is referred to as the joint density of X with respect to v, and each f_i as the marginal density

of X_i with respect to v_i.

1.7.3 The Information Contained in a Multidimensional Random Variable

Returning to the general case, the information contained in X is once again represented by the σ-algebra σX generated by $X = (X_t)_{t\in T}$. Once again, we can view σX as the amount of information contained in X. Due to the unique property of measurable rectangles, namely that they are essentially finite rectangles trivially extended to an infinite product space if need be, we have the following generating set for σX:

Lemma 1.13 Let T be an arbitrary index set, $\{(E_t, \mathcal{E}_t) \mid t \in T\}$ a collection of measurable spaces with product $(E, \mathcal{E})$, and $X = (X_t)_{t\in T}$ a random variable taking values in $(E, \mathcal{E})$. Define

$$\mathcal{H}_0 = \Big\{ \bigcap_{t\in J} X_t^{-1}(A_t) \mid J \subset T \text{ is finite}, \forall t \in J, A_t \in \mathcal{E}_t \Big\}.$$

Then, $\mathcal{H}_0$ is a π-system on Ω that generates σX.

Proof) For any finite subset $J \subset T$ and $A_t \in \mathcal{E}_t$ for any $t \in J$, defining $A = \prod_{t\in T} A_t$, in which $A_t = E_t$ for any $t \in T$ such that $t \notin J$,

$$X^{-1}(A) = \bigcap_{t\in J} X_t^{-1}(A_t) \in \sigma X,$$

so that $\mathcal{H}_0 \subset \sigma X$.

To show that the reverse inclusion holds, define $\mathcal{E}_0$ as the set of all measurable rectangles on E, that is, as

$$\mathcal{E}_0 = \Big\{ \prod_{t\in T} A_t \mid \exists J \subset T \text{ s.t. } J \text{ is finite}, A_t \in \mathcal{E}_t \text{ if } t \in J, A_t = E_t \text{ if } t \notin J \Big\}.$$

Since $\mathcal{E}_0$ is a π-system generating $\mathcal{E}$, by lemma 1.5 the set

$$\mathcal{D} = \{X^{-1}(A) \mid A \in \mathcal{E}_0\}$$

is a π-system on Ω generating σX.
Note that, for any measurable rectangle $A = \prod_{t\in T} A_t \in \mathcal{E}_0$ such that $A_t \neq E_t$ only for t in a finite set $J \subset T$,

$$X^{-1}(A) = \bigcap_{t\in J} X_t^{-1}(A_t) \in \mathcal{H}_0.$$

Conversely, for any $\bigcap_{t\in J} X_t^{-1}(A_t) \in \mathcal{H}_0$, where each $A_t \in \mathcal{E}_t$ and $J \subset T$ is finite, defining

$A_t = E_t$ for any $t \in T \setminus J$ and $A = \prod_{t\in T} A_t$, we have $A \in \mathcal{E}_0$ and thus

$$\bigcap_{t\in J} X_t^{-1}(A_t) = X^{-1}(A) \in \mathcal{D}.$$

It follows that $\mathcal{D} = \mathcal{H}_0$, and as such that $\mathcal{H}_0$ is a π-system on Ω that generates σX. Q.E.D.

The above lemma allows us to obtain a very intuitive characterization of the information contained in $X = (X_t)_{t\in T}$ as the totality of the information contained in each X_t. Recall that, given a collection of σ-algebras $\{\mathcal{F}_t \mid t \in T\}$ on Ω, the σ-algebra generated by their union $\bigcup_t mathcal F_t$ is denoted $\bigvee_t \mathcal{F}_t$.
A generating π-system for $\bigvee_t \mathcal{F}_t$ is given by

$$\mathcal{G}_0 = \Big\{ \bigcap_{t\in J} A_t \mid A_t \in \mathcal{F}_t, J \subset T \text{ is finite } \}.$$

For any finite intersection $\bigcap_{t\in J} A_t$, each $A_t \in \mathcal{F}_t \subset \bigvee_s \mathcal{F}_s$, and because σ-algebras are closed under intersections, $\bigcap_{t\in J} A_t \in \bigvee_t \mathcal{F}_t$; this implies that $\mathcal{G}_0 \subset \bigvee_t \mathcal{F}_t$ and thus that $\sigma\mathcal{G}_0 \subset \bigvee_t \mathcal{F}_t$.
On the other hand, any $A_t \in \mathcal{F}_t$ for some $t \in T$ is contained in $\mathcal{G}_0$ and thus $\sigma\mathcal{G}_0$ (put $J = \{t\}$), which implies that $\bigcup_t \mathcal{F}_t \subset \sigma\mathcal{G}_0$. Since $\bigvee_t \mathcal{F}_t$ is the σ-algebra generated by $\bigcup_t \mathcal{F}_t$, this means that $\bigvee_t \mathcal{F}_t \subset \sigma\mathcal{G}_0$, and as such that $\mathcal{G}_0$ generates $\bigvee_t \mathcal{F}_t$.
In light of this operation, the information contained in $X = (X_t)_{t\in T}$ is the totality of the information contained in each X_t in the sense that σX is precisely $\bigvee_t \sigma X_t$. This is formally shown below:

Theorem 1.14 Let T be an arbitrary index set, $\{(E_t, \mathcal{E}_t) \mid t \in T\}$ a collection of measurable spaces with product $(E, \mathcal{E})$, and $X = (X_t)_{t\in T}$ a random variable taking values in $(E, \mathcal{E})$. Then,

$$\sigma X = \bigvee_{t\in T} \sigma X_t.$$

Proof) From lemma 1.13, we can see that the set

$$\mathcal{H}_0 = \Big\{ \bigcap_{t\in J} X_t^{-1}(A_t) \mid A_t \in \mathcal{E}_t, J \subset T \text{ is finite} \Big\}$$

is a π-system generating σX, and we just showed above that

$$\mathcal{G}_0 = \Big\{ \bigcap_{t\in J} H_t \mid H_t \in \sigma X_t, J \subset T \text{ is finite } \}$$

is a π-system generating $\bigvee_{t\in T} \sigma X_t$. It is apparent, however, that $\mathcal{H}_0 = \mathcal{G}_0$, since any $H_t \in \sigma X_t$ can be written as $X_t^{-1}(A_t)$ for some $A_t \in \mathcal{E}_t$, and conversely, any $X_t^{-1}(A_t)$ is contained in σX_t if $A_t \in \mathcal{E}_t$. Therefore, $\mathcal{H}_0$ generates both σX and $\bigvee_{t\in T} \sigma X_t$, which means that they are the same σ-algebra on Ω.

Q.E.D.

The σ-algebra $\bigvee_t \sigma X_t$ is also sometimes denoted by $\sigma\{X_t \mid t \in T\}$. The claim of the above theorem can then be succinctly written as

$$\sigma X = \sigma\{X_t \mid t \in T\} \quad \text{for } X = (X_t)_{t\in T}.$$

Intuitively, if $S \subset T$, then the information contained in $\bar{X} = (X_t)_{t\in S}$ must also be contained in $X = (X_t)_{t\in T}$. This intuition is confirmed in the following result:

Lemma 1.15 Let T be an arbitrary index set, $\{(E_t,\mathcal{E}_t) \mid t \in T\}$ a collection of measurable spaces with product $(E,\mathcal{E})$, and $X = (X_t)_{t\in T}$ a random variable taking values in $(E,\mathcal{E})$.
Let $S \subset T$ be an index set nested in T and $\bar{X} = (X_t)_{t\in S}$. Then, X contains the information on $\bar{X}$, that is, $\sigma\bar{X} \subset \sigma X$.

Proof) Define $(F,\mathcal{F}) = \bigotimes_{t\in S}(E_t,\mathcal{E}_t)$, and $\mathcal{F}_0$ the set of all measurable rectangles on F. By lemma 1.5, the set

$$\mathcal{H}_0 = \{\bar{X}^{-1}(A) \mid A \in \mathcal{F}_0\}$$

generates $\sigma\bar{X}$. Choose any measurable rectangle $A = \prod_{t\in S} A_t \in \mathcal{F}_0$, where there exists a finite set $J \subset S$ such that $A_t \neq E_t$ only for $t \in J$. It follows that

$$\bar{X}^{-1}(A) = \bigcap_{t\in J} X_t^{-1}(A_t) \in \sigma X,$$

from the previous lemma, and as such, $\mathcal{H}_0 \subset \sigma X$. σX is a σ-algebra on $\mathcal{H}$, so $\sigma\mathcal{H}_0 = \sigma\bar{X} \subset \sigma X$.
Q.E.D.

The fact that $\mathcal{E}$ is generated by the collection of finite measurable rectangles allows for an intuitive and simple extension of the Doob-Dynkin lemma for multidimensional random variables:

Theorem 1.16 (The Multidimensional Doob-Dynkin Lemma)
Let T be an arbitrary index set, $\{(E_t,\mathcal{E}_t) \mid t \in T\}$ a collection of measurable spaces with product $(E,\mathcal{E})$, and $Y = (Y_t)_{t\in T}$ a random variable taking values in $(E,\mathcal{E})$.
X is a σY-measurable non-negative or complex random variable if and only if there exists a sequence $\{t_n\}_{n\in K} \subset T$, where $K = \{1,\cdots,N\}$ for N possibly infinity, and a non-negative or

complex function f defined on $\prod_n E_{t_n}$ and measurable relative to $\bigotimes_n \mathcal{E}_{t_n}$ such that

$$X = f \circ (Y_{t_1}, Y_{t_2}, \cdots).$$

Proof) Again, sufficiency is easily shown. Suppose that there exists a sequence $\{t_n\}_{n\in K} \subset T$, where $K = \{1, \cdots, N\}$ for N possibly infinity, and a non-negative or complex function f defined on $\prod_n E_{t_n}$ and measurable relative to $\bigotimes_n \mathcal{E}_{t_n}$ such that

$$X = f \circ (Y_{t_1}, Y_{t_2}, \cdots).$$

Defining $\bar{Y} = (Y_{t_n})_{n\in K}$, which is a random variable taking values in $(F, \mathcal{F}) = \prod_n (E_{t_n}, \mathcal{E}_{t_n})$ because each Y_{t_n} is a random variable in $(E_{t_n}, \mathcal{E}_{t_n})$, we can write $X = f \circ \bar{Y}$. By the Doob-Dynkin lemma X is $\sigma\bar{Y}$-measurable. Because $\{t_n\}_{n\in K} \subset T$, it follows from the previous result that $\sigma\bar{Y} \subset \sigma Y$ and thus that X is σY-measurable as well.

For necessity, we make use of the monotone class theorem for functions. Define $\mathcal{M}$ as the colleciton of all numerical functions defined on Ω such that there exists a $\{t_n\}_{n\in K} \subset T$, where $K = \{1, \cdots, N\}$ with N possibly infinity, such that $X = f \circ (Y_{t_1}, Y_{t_2}, \cdots)$ for a $\bigotimes_n \mathcal{E}_{t_n}$-measurable function f. We will show that $\mathcal{M}$ is a monotone class of functions on Ω:

i) $I_\Omega \in \mathcal{M}$ because $I_\Omega = I_{E_t} \circ Y_t$ for any $t \in T$; in this case, we take $N = 1$ and $\{t_n\}_{n\in K} = \{t\}$.

ii) For any $a, b \in \mathbb{R}$ and $X_1, X_2 \in \mathcal{M}_+$, there exists a sequence $\{t_n\}_{n\in K} \subset T$ such that

$$X_i = f_i \circ (Y_{t_1}, Y_{t_2}, \cdots)$$

for some measurable bounded real function f_i and $i = 1, 2$ (in the case that the sequences corresponding to each random variable is different, we need only define $\{t_n\}_{n\in K}$ as their union and trivially extend f_1, f_2 to functions on $\prod_n E_{t_n}$). Defining $f = af_1 + bf_2$, f is also a measurable bounded real function, and we have

$$aX_1 + bX_2 = f \circ (Y_{t_1}, Y_{t_2}, \cdots).$$

By definition, $aX_1 + bX_2 \in \mathcal{M}_b$.

iii) Let $\{X_n\}_{n\in N_+}$ be a sequence in $\mathcal{M}_+$ that increases to X. As before, there exists a sequence $\{t_k\}_{k\in K} \subset T$ such that, for each $n \in N_+$,

$$X_n = f_n \circ (Y_{t_1}, Y_{t_2}, \cdots)$$

for some non-negative measurable function f_n (again, if the sequences corresponding to each X_n is different, we let $\{t_k\}_{k\in K}$ be the union of those sequences and f_n the trivial extension to $\prod_k E_{t_k}$). Defining

$$f = \sup_{n\in N_+} f_n,$$

f is a $\bigotimes_k \mathcal{E}_{t_k}$-measurable non-negative function such that

$$X_n \nearrow f \circ (Y_{t_1}, Y_{t_2}, \cdots)$$

pointwise as $n \to \infty$. By the uniqueness of limits, it follows that

$$X = f \circ (Y_{t_1}, Y_{t_2}, \cdots)$$

and thus that $X \in \mathcal{M}_+$.

Choose any $H \in \sigma Y$; by definition, there exists a measurable rectangle $A = \prod_{t\in T} A_t$ on E such that $H = Y^{-1}(A)$. Because of the way measurable rectangles are defined, there exists a finite $J \subset T$ such that $A_t = E_t$ if $t \notin J$. Denoting $J = \{t_1, \cdots, t_N\}$ for $N \in N_+$, we have

$$I_H = I_{Y^{-1}(A)} = I_{A_{t_1}\times\cdots\times A_{t_N}} \circ (Y_{t_1}, \cdots, Y_{t_N}).$$

Since $I_{A_{t_1}\times\cdots\times A_{t_N}}$ is a non-negative $\bigotimes_n \mathcal{E}_{t_n}$-measurable function, it follows that $I_H \in \mathcal{M}$. Therefore, $\mathcal{M}$ contains all indicator functions I_H where $H \in \sigma Y$, and becauase σY is a π-system generating itself, by the monotone class theorem for functions, $\mathcal{M}$ contains every bounded or non-negative real-valued σY-measurable function. In particular, for any σY-measurable non-negative random variable X, there exists a sequence $\{t_n\}_{n\in K}$, where $K = \{1, \cdots, N\}$ for N possibly infinity, and a non-negative and $\bigotimes_n \mathcal{E}_{t_n}$-measurable function f such that

$$X = f \circ (Y_{t_1}, Y_{t_2}, \cdots).$$

We can now extend this result to arbitrary complex valued functions as in lemma 1.5. Q.E.D.

Of note is that, even though the indicator of each set in σY is the function of a finite number of coordinates of Y, because we take limits when extending this result to arbitrary non-negative σY-measurable random variables, such random variables are possibly functions of a countable many coordinates of Y. The important part is that none of them are functions of uncountably many coordinates of Y, even when the underlying index set T is uncountable.

1.8 Independence

Let T be an arbitrary index set, and $\{\mathcal{F}_t \mid t \in T\}$ a collection of sub σ-algebras of $\mathcal{H}$. Then, we say that $\{\mathcal{F}_t \mid t \in T\}$ is an independency, or that its elements are mutually independent, if

$$\mathbb{P}\left(\bigcap_{t\in J} A_t\right) = \prod_{t\in J} \mathbb{P}(A_t)$$

for any finite $J \subset T$ and $A_t \in \mathcal{F}_t$ for each $t \in J$. Clearly, any subcollection of $\{\mathcal{F}_t \mid t \in T\}$ is also an independency.
The following elementary result shows us that the partition of an independency is also an independency.

Theorem 1.17 Let T be an arbitrary index set, and $\{\mathcal{F}_t \mid t \in T\}$ a collection of sub σ-algebras of $\mathcal{H}$. Let $\{T_1, \cdots, T_k\}$ be a partition of T, and define $\mathcal{G}_i = \bigvee_{t\in T_i} \mathcal{F}_i$ for $1 \le i \le k$. Then, the collection $\{\mathcal{G}_1, \cdots, \mathcal{G}_k\}$ is also an independency.

Proof) We proceed by induction on k, the number of partitions of T. Suppose first that $k = 2$, so that $\{T_1, T_2\}$ is a partition of T. Defining $\mathcal{G}_1$ and $\mathcal{G}_2$ as above, recally that, for each $i = 1, 2$, the set

$$\Pi_i = \Big\{ \bigcap_{t\in J} A_t \mid A_t \in \mathcal{F}_t, J \subset T_i \text{ is finite} \Big\}$$

is a π-system generating $\mathcal{G}_i$. Choose any $B = \bigcap_{t\in J_2} B_t \in \Pi_2$ for some finite $J_2 \subset T_2$, and define

$$\mathcal{D}_1 = \{A \in \mathcal{H} \mid \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\}.$$

We can show that $\mathcal{D}_1$ is a λ-system on Ω:

i) $\Omega \in \mathcal{D}_1$ because $\mathbb{P}(\Omega \cap B) = \mathbb{P}(B) = \mathbb{P}(B)\mathbb{P}(\Omega)$.

ii) For any $A_1, A_2 \in \mathcal{D}_1$ such that $A_1 \subset A_2$, $A_2 \setminus A_1 \in \mathcal{H}$ and

$$\begin{aligned}
\mathbb{P}((A_2 \setminus A_1) \cap B) &= \mathbb{P}((A_2 \cap B) \setminus (A_1 \cap B)) \\
&= \mathbb{P}(A_2 \cap B) - \mathbb{P}(A_1 \cap B) && \text{(Finite additivity)} \\
&= (\mathbb{P}(A_2) - \mathbb{P}(A_1)) \cdot \mathbb{P}(B) \\
&= \mathbb{P}(A)\mathbb{P}(B). && \text{(Finite additivity)}
\end{aligned}$$

By definition, $A_2 \setminus A_1 \in \mathcal{D}_1$.

iii) For any $\{A_n\}_{n\in N_+} \subset \mathcal{D}_1$ with $A = \bigcup_n A_n$, $A \in \mathcal{H}$ and, by sequential continuity,

$$\mathbb{P}(A\cap B) = \lim_{n\to\infty} \mathbb{P}(A_n \cap B) = \Big(\lim_{n\to\infty} \mathbb{P}(A_n)\Big)\mathbb{P}(B) = \mathbb{P}(A)\mathbb{P}(B).$$

By definition, $A \in \mathcal{D}_1$.

We can also see that $\Pi_1 \subset \mathcal{D}_1$, since for any $A = \bigcap_{t\in J_1} A_t \in \Pi_1$ for some finite set $J_1 \subset T_1$,

$$\mathbb{P}(A\cap B) = \mathbb{P}\left(\left(\bigcap_{t\in J_1} A_t\right) \cap \left(\bigcap_{t\in J_2} B_t\right)\right) = \left(\prod_{t\in J_1} \mathbb{P}(A_t)\right)\left(\prod_{t\in J_2} B_t\right) = \mathbb{P}(A)\mathbb{P}(B)$$

because the collection of σ-algebras $\{\mathcal{F}_t \mid t \in J_1 \cup J_2\}$ is an independency. By the $\pi-\lambda$ theorem, it now follows that $\mathcal{G}_1 \subset \mathcal{D}_1$, that is,

$$\mathbb{P}(A\cup B) = \mathbb{P}(A)\mathbb{P}(B)$$

for any $A \in \mathcal{G}_1$. This holds for any $B \in \Pi_2$, so we can see that $\mathbb{P}(A\cup B) = \mathbb{P}(A)\mathbb{P}(B)$ for any $A \in \mathcal{G}_1$ and $B \in \Pi_2$.

Now choose any $A \in \mathcal{G}_1$ and define

$$\mathcal{D}_2 = \{B \in \mathcal{D}_2 \mid \mathbb{P}(A\cap B) = \mathbb{P}(A)\mathbb{P}(B)\}.$$

By almost the same process as above, we can show that $\mathcal{D}_2$ is a λ-system on Ω. In addition, $\Pi_2 \subset \mathcal{D}_2$ by the result shown above. By the $\pi-\lambda$ theorem, it now follows that $\mathcal{G}_2 \subset \mathcal{D}_2$, that is,

$$\mathbb{P}(A\cup B) = \mathbb{P}(A)\mathbb{P}(B)$$

for any $B \in \mathcal{G}_2$. This holds for any $A \in \mathcal{G}_1$, so we can see that $\mathbb{P}(A\cup B) = \mathbb{P}(A)\mathbb{P}(B)$ for any $A \in \mathcal{G}_1$ and $B \in \mathcal{G}_2$. By definition, $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent.

Suppose now that the claim of the theorem holds for some $k \geq 2$. Let $\{T_1, \cdots, T_{k+1}\}$ be a partition of T, and let $\mathcal{G}_i = \bigvee_{t\in T_i} \mathcal{F}_t$ for $1 \leq i \leq k+1$. Defining $T_0 = \{T_1, \cdots, T_k\}$, because the collection $\{\mathcal{F}_t \mid t \in T_0\}$, being a subcollection of $\{\mathcal{F}_t \mid t \in T\}$, is an independency. As such, by the inductive hypothesis $\mathcal{G}_1, \cdots, \mathcal{G}_k$ are independent, that is,

$$\mathbb{P}\left(\bigcap_{i=1}^{k} A_i\right) = \prod_{i=1}^{k} \mathbb{P}(A_i)$$

for any $A_i \in \mathcal{G}_i$ for $1 \leq i \leq k$.
Since $\{T_0, T_{k+1}\}$ is a partition of T, by the result for $k = 2$ we can see that $\bigvee_{i=1}^{k} \mathcal{G}_i$ and $\mathcal{G}_{k+1}$ are independent. Therefore, for any $A_1, \cdots, A_{k+1}$ such that $A_i \in \mathcal{G}_i$ for any

$1 \leq i \leq k+1$, since $\bigcap_{i=1}^{k} A_i \in \bigvee_{i=1}^{k} \mathcal{G}_i$,

$$\mathbb{P}\left(\bigcap_{i=1}^{k+1} A_i\right) = \mathbb{P}\left(\bigcap_{i=1}^{k} A_i\right)\mathbb{P}(A_{k+1}) = \prod_{i=1}^{k+1}\mathbb{P}(A_i)\,.$$

By definition, $\mathcal{G}_1, \cdots, \mathcal{G}_{k+1}$ is an independency, and the claim of the theorem follows by induction.
Q.E.D.

1.8.1 Characterizations of Independence

An equivalent formulation of independence in terms of integrals can now be easily given:

Theorem 1.18 (Characterization of Independence with Expected Values)
Let T be an arbitrary index set, and $\{\mathcal{F}_t \mid t \in T\}$ a collection of sub σ-algebras of $\mathcal{H}$. Then, $\{\mathcal{F}_t \mid t \in T\}$ is an independency if and only if, for any finite subset $\{t_1, \cdots, t_k\} \subset T$ and non-negative random variables $V_1, \cdots, V_k$ measurable relative to $\mathcal{F}_{t_1}, \cdots, \mathcal{F}_{t_k}$ respectively,

$$\mathbb{E}\left[\prod_{i=1}^{k} V_i\right] = \prod_{i=1}^{k}\mathbb{E}[V_i]\,.$$

Proof) Sufficiency follows immediately, since for any finite set $\{t_1, \cdots, t_k\} \subset T$, we can take $V_i = I_{A_i}$ for some $A_i \in \mathcal{F}_{t_i}$ for $1 \leq i \leq k$ and independence follows from the definition.

To show necessity, we proceed by induction on k. First choose any $\{t_1, t_2\} \subset T$ and denote $\mathcal{G}_1 = \mathcal{F}_{t_1}$, $\mathcal{G}_2 = \mathcal{F}_{t_2}$. Then, for any simple functions $V = \sum_{i=1}^{n} \alpha_i \cdot I_{A_i}$ and $U = \sum_{i=1}^{k} \beta_i \cdot I_{B_i}$ that are $\mathcal{G}_1$ and $\mathcal{G}_2$ measurable, respectively, we have

$$\begin{aligned}\mathbb{E}[VU] = \sum_{i=1}^{n}\sum_{j=1}^{k} \alpha_i\beta_j \cdot \mathbb{P}(A_i \cap B_j) &= \sum_{i=1}^{n}\sum_{j=1}^{k} \alpha_i\beta_j \cdot \mathbb{P}(A_i)\,\mathbb{P}(B_j) \\ &= \left(\sum_{i=1}^{n} \alpha_i \cdot \mathbb{P}(A_i)\right)\left(\sum_{i=1}^{k} \beta_i \cdot \mathbb{P}(B_i)\right) = \mathbb{E}[V]\,\mathbb{E}[U]\,,\end{aligned}$$

where the second equality follows from the independence of $\mathcal{G}_1$ and $\mathcal{G}_2$.
Now let $V \in \mathcal{G}_{1,+}$ and $U \in \mathcal{G}_{2,+}$. Letting $\{V_n\}_{n\in N_+}$ and $\{U_n\}_{n\in N_+}$ be sequences of $\mathcal{G}_1$ and $\mathcal{G}_2$-measurable simple functions increasing to V and U, by the MCT

$$\begin{aligned}\mathbb{E}[VU] &= \lim_{n\to\infty}\lim_{m\to\infty}\mathbb{E}[V_nU_m] \\ &= \left(\lim_{n\to\infty}\mathbb{E}[V_n]\right)\left(\lim_{m\to\infty}\mathbb{E}[U_m]\right) = \mathbb{E}[V]\,\mathbb{E}[U]\,.\end{aligned}$$

Now suppose that necessity holds for some $k \geq 2$. Choose a subset $T_0 = \{t_1, \cdots, t_{k+1}\} \subset T$ and $V_i \in \mathcal{F}_{t_i}$ for $1 \leq i \leq k$. Defining $\mathcal{G}_1 = \bigvee_{i=1}^{k} \mathcal{F}_{t_i}$ and $\mathcal{G}_2 = \mathcal{F}_{t_{k+1}}$, because $\{\{t_1, \cdots, t_k\}, \{t_{k+1}\}\}$ is a partition of T_0, by the previous theorem $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent. Because $\prod_{i=1}^{k} V_i$ is a $\mathcal{G}_1$-measurable non-negative function, by the result for $k = 2$ we have

$$\mathbb{E}\left[\prod_{i=1}^{k+1} V_i\right] = \mathbb{E}\left[\prod_{i=1}^{k} V_i\right] \mathbb{E}\left[V_{k+1}\right].$$

Finally, by the inductive hypothesis,

$$\mathbb{E}\left[\prod_{i=1}^{k+1} V_i\right] = \prod_{i=1}^{k} \mathbb{E}\left[V_i\right].$$

Therefore, we have

$$\mathbb{E}\left[\prod_{i=1}^{k+1} V_i\right] = \prod_{i=1}^{k+1} \mathbb{E}\left[V_i\right],$$

and necessity holds by induction.
Q.E.D.

Now we move on from the independence of σ-algebras to that of random variables. Let T be an arbitrary index set, and $\{(E_t, \mathcal{E}_t) \mid t \in T\}$ a collection of measurable spaces with product $(E, \mathcal{E})$. For any collection of random variables $\{X_t \mid t \in T\}$ such that X_t takes values in $(E_t, \mathcal{E}_t)$ for all $t \in T$, we say $\{X_t \mid t \in T\}$ is an independency, or that its elements are mutually independent, if the corresponding collection of σ-algebras $\{\sigma X_t \mid t \in T\}$ is an independency.
By the definition of independence stated above, we can equivalently say that $\{X_t \mid t \in T\}$ is an independency if

$$\mathbb{P}\left(\bigcap_{t \in J} \{X_t \in A_t\}\right) = \prod_{t \in J} \mathbb{P}\left(X_t \in A_t\right)$$

for any $A_t \in \mathcal{E}_t$ and a finite subset $J \subset T$.
Furthermore, theorem 1.18, together with the Doob-Dynkin lemma, shows us that $\{X_t \mid t \in T\}$ is an independency if and only if, for any finite subset $J \subset T$ and non-negative functions $\{f_t \mid t \in J\}$ such that each $f_t \in \mathcal{E}_{t,+}$, we have

$$\mathbb{E}\left[\prod_{t \in J} f_t \circ X_t\right] = \prod_{t \in J} \mathbb{E}\left[f_t \circ X_t\right].$$

Below we formulate, in terms of the distributions, a necessary and sufficient condition for a finite set of random variables to be independent:

Theorem 1.19 (Characterization of Independence with Distributions)
Let $\{X_1, \cdots, X_k\}$ be a collection of random variables such that each X_i takes values in the measurable space $(E_i, \mathcal{E}_i)$ and has distribution μ_i. $X_1, \cdots, X_k$ are independent if and only if their joint distribution π is the product of their individual distributions, that is,

$$\pi = \mu_1 \times \cdots \times \mu_k.$$

Proof) Throughout, let $X = (X_1, \cdots, X_k)$, so that X is a random variable taking values in $(E, \mathcal{E}) = \bigotimes_{i=1}^k (E_i, \mathcal{E}_i)$. We can then interpret π as the distribution $\mathbb{P} \circ X^{-1}$ of X.
Suppose that $X_1, \cdots, X_k$ are independent. Then, for any measurable rectangle $A_1 \times \cdots \times A_k$ on E, we have

$$\begin{aligned}\pi(A_1 \times \cdots \times A_k) &= \mathbb{P}(X \in A_1 \times \cdots \times A_k) = \mathbb{P}\left(\bigcap_{i=1}^k X_i^{-1}(A_i)\right) \\ &= \prod_{i=1}^k \mathbb{P}\left(X_i^{-1}(A_i)\right) = \prod_{i=1}^k \mu_i(A_i).\end{aligned}$$

Thus, $\pi = \mu_1 \times \cdots \times \mu_k$ on the set of all measurable rectangles, and because these sets form a π-system generating $\mathcal{E}$, and

$$\pi(E) = \prod_{i=1}^k \mu_i(E_i) = 1 < +\infty,$$

it follows that $\pi = \mu_1 \times \cdots \times \mu_k$ on $\mathcal{E}$.

Conversely, suppose that $\pi = \mu_1 \times \cdots \times \mu_k$ on $\mathcal{E}$. Then, for any $A_1, \cdots, A_k$ such that $A_i \in \mathcal{E}_i$ for $1 \leq i \leq k$,

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^k \{X_i \in A_i\}\right) &= \mathbb{P}(X \in A_1 \times \cdots \times A_k) \\ &= \pi(A_1 \times \cdots \times A_k) = \prod_{i=1}^k \mu_i(A_i) = \prod_{i=1}^k \mathbb{P}(X_i \in A_i).\end{aligned}$$

By definition, $X_1, \cdots, X_k$ are independent.
Q.E.D.

If each random variable has a density with respect to some σ-finite measure, then we can obtain a sharper characterization for independence:

Theorem 1.20 (Characterization of Independence with Densities)
Let $\{X_1, \cdots, X_k\}$ be a collection of random variables such that each X_i takes values in the measurable space $(E_i, \mathcal{E}_i)$ and has distribution μ_i, and let π be the joint distribution of $X_1, \cdots, X_k$. Suppose that, for each $1 \leq i \leq k$, there exists a σ-finite measure v_i on $(E_i, \mathcal{E}_i)$ with respect to which μ_i has a density f_i.
In this case, $X_1, \cdots, X_k$ are independent if and only if the product of their densities is a density of π with respect to the product measure $v_1 \times \cdots \times v_k$, that is, $f : \mathbb{E} \to [0, +\infty]$ defined as

$$f(x) = \prod_{i=1}^{k} f_i(x_i)$$

for any $x \in \mathbb{E}$ is a member of $\frac{\partial \pi}{\partial v_1 \times \cdots \times v_k}$.

Proof) Throughout, let $X = (X_1, \cdots, X_k)$, so that X is a random variable taking values in $(E, \mathcal{E}) = \bigotimes_{i=1}^{k} (E_i, \mathcal{E}_i)$. We can then interpret π as the distribution $\mathbb{P} \circ X^{-1}$ of X.
Suppose that $X_1, \cdots, X_k$ are independent. Then, by the previous theorem, π is the product of the individual distributions $\mu_1, \cdots, \mu_k$. Define $f : \mathbb{E} \to [0, +\infty]$ as above; then, f is $\mathcal{E}$-measurable and therefore the indefinite integral π_0 of f with respect to the product measure $v_1 \times \cdots \times v_k$ is well-defined.
For any measurable rectangle $A_1 \times \cdots \times A_k$,

$$\begin{aligned}\pi(A_1 \times \cdots \times A_k) &= \prod_{i=1}^{k} \mu_i(A_i) = \prod_{i=1}^{k} \int_{A_i} f_i(x_i) dv_i(x_i) = \int_{A_1} \cdots \int_{A_k} f(x) dv_k(x_k) \cdots dv_1(x_1) \\ &= \int_{A_1 \times \cdots \times A_k} f(x) d(v_1 \times \cdots \times v_k)(x) = \pi_0(A_1 \times \cdots \times A_k),\end{aligned}$$

where the second to last equality follows from Fubini's theorem. π and π_0 are thus probability measures that agree on the π-system of measurable rectangles, so that $\pi = \pi_0$ on $\mathcal{E}$. By implication, f is the density of π with respect to $v_1 \times \cdots \times v_k$.

Conversely, suppose that the $f : \mathbb{E} \to [0, +\infty]$ defined above is the density of π with respect to $v_1 \times \cdots \times v_k$. Then, for any measurable rectangle $A_1 \times \cdots \times A_k$,

$$\begin{aligned}\pi(A_1 \times \cdots \times A_k) &= \int_{A_1 \times \cdots \times A_k} f(x) d(v_1 \times \cdots \times v_k)(x) \\ &= \prod_{i=1}^{k} \int_{A_i} f_i(x_i) dv_i(x_i) = \prod_{i=1}^{k} \mu_i(A_i)\end{aligned}$$

by Fubini's theorem. Thus, π and $\mu_1 \times \cdots \times \mu_k$ are probability measures that agree on the set of all measurable rectangles, so $\pi = \mu_1 \times \cdots \times \mu_k$ on $\mathcal{E}$. By the previous theorem, this implies that $X_1, \cdots, X_k$ are independent.
Q.E.D.

Finally, we can also characterize independence by exploiting the fact that characteristic functions determine the distribution of a random vector. Below we illustrate this approach:

Theorem 1.21 (Characterization of Independence with Charcteristic Functions) Let X and Y be k-and m-dimensional real random vectors with characteristic functions φ_x and φ_y. Defining the $k+m$-dimensional real random vector $Z=(X',Y')'$ and letting its characteristic function be φ, X and Y are independent if and only if

$$\varphi((t',r')') = \varphi_x(t)\varphi_y(r)$$

for any $t \in \mathbb{R}^k$ and $r \in \mathbb{R}^m$.

Proof) Let X and Y have distributions μ_x and μ_y, and let the joint distribution of X and Y be π.
Suppose X and Y are independent. Then, $\pi = \mu_x \times \mu_y$, so that for any $t \in \mathbb{R}^k$, $r \in \mathbb{R}^m$ and $s=(t',r')' \in \mathbb{R}^{k+m}$, we have

$$\begin{aligned}
\varphi(s) &= \int_{\mathbb{R}^{k+m}} \exp(is'z)d\pi(z) \\
&= \int_{\mathbb{R}^{k+m}} \exp(it'x)\exp(ir'y)d(\mu_x \times \mu_y)(x,y) \\
&= \left(\int_{\mathbb{R}^k} \exp(it'x)d\mu_x(x)\right)\left(\int_{\mathbb{R}^m} \exp(ir'y)d\mu_y(y)\right) \qquad \text{(Fubini's Theorem)} \\
&= \varphi_x(t)\varphi_y(r).
\end{aligned}$$

Conversely, suppose that

$$\varphi((t',r')') = \varphi_x(t)\varphi_y(r)$$

for any $t \in \mathbb{R}^k$ and $r \in \mathbb{R}^m$. Letting $\pi_0 = \mu_x \times \mu_y$, the characteristic function ϕ of π_0 is defined as

$$\phi(s) = \varphi_x(t)\varphi_y(r)$$

for any $t \in \mathbb{R}^k$, $r \in \mathbb{R}^m$ and $s=(t',r')' \in \mathbb{R}^{k+m}$, as showed above. This means that $\varphi = \phi$ on $\mathbb{R}^{k+m}$, and because the characteristic function of a distribution determines the distribution itself, $\pi = \pi_0$ on $\mathcal{B}(\mathbb{R}^{k+m})$. As such, X and Y are independent.
Q.E.D.

1.8.2 Construction of Sequences of Independent Random Variables

Having defined the independence of random variables, the next question is naturally whether there actually exists a probability space on which independent random variables with specific distributions exist. This is a simple matter for a finite collection of random variables.
To illustrate this, let $(E_1,\mathcal{E}_1),\cdots,(E_n,\mathcal{E}_n)$ be measurable spaces and $\mu_1,\cdots,\mu_n$ probability measures on those spaces. Define $(\Omega,\mathcal{H})=\bigotimes_{i=1}^n(E_i,\mathcal{E}_i)$, $\mathbb{P}=\mu_1\times\cdots\times\mu_n$ and $X_i:\Omega\to E_i$ as

$$X_i(\omega_1,\cdots,\omega_n)=\omega_i$$

for any $(\omega_1,\cdots,\omega_n)\in\Omega$ and $1\leq i\leq n$. Then, $X_1,\cdots,X_n$ are independent random variables taking values in $(E_1,\mathcal{E}_1),\cdots,(E_n,\mathcal{E}_n)$ with distributions $\mu_1,\cdots,\mu_n$ and underlying probability space $(\Omega,\mathcal{H},\mathbb{P})$. We verify the claims one by one:

- $(\Omega,\mathcal{H},\mathbb{P})$ **is a probability space**
  This is trivial, since $\mu_1\times\cdots\times\mu_n$ is a product measure on $(\Omega,\mathcal{H})$ with total mass 1 because each μ_i is a probability measure.

- $X_1,\cdots,X_n$ **are random variables**
  For any $1\leq i\leq n$ and $A_i\in\mathcal{E}_i$,

  $$X_i^{-1}(A_i)=A_i\times E_{-i}\in\bigotimes_{i=1}^n\mathcal{E}_i=\mathcal{H},$$

  so X_i is a random variable taking values in $(E_i,\mathcal{E}_i)$.

- **For any** $1\leq i\leq n$**,** X_i **has distribution** μ_i
  Choosing any $A_i\in\mathcal{E}_i$,

  $$\mathbb{P}(X_i\in A_i)=(\mu_1\times\cdots\times\mu_n)(A_i\times E_{-i})=\mu_i(A_i).$$

- $X_1,\cdots,X_n$ **are independent**
  For any $A_1,\cdots,A_n$ such that $A_i\in\mathcal{E}_i$ for $1\leq i\leq n$,

  $$\mathbb{P}\left(\bigcap_{i=1}^n\{X_i\in A_i\}\right)=\mathbb{P}(A_1\times\cdots\times A_n)=\prod_{i=1}^n\mu_i(A_i)=\prod_{i=1}^n\mathbb{P}(X_i\in A_i),$$

  so by definition, $X_1,\cdots,X_n$ are independent.

However, the issue is more complicated when it comes to constructing a sequence of independent random variables with specific distributions because we cannot define the product of an infinte number of measures. Fortunately, there is a way to preserve the intuition of the construction

above when constructing a probability space underlying a sequence of independent random variables. Namely, we extend a pre-measure that yields the desired finite dimensional distribution, constructed exactly as above, to a probability measure using Caratheodory's extension theorem.

Cylinder Sets and Finite Dimensional Distributions

Formally, let $\{(E_i, \mathcal{E}_i, \mu_i) \mid i \in N_+\}$ be a collection of probability spaces, define for any $n \in N_+$ the probability space

$$(E^n, \mathcal{E}^n, \pi_n) = \bigotimes_{i=1}^{n} (E_i, \mathcal{E}_i, \mu_i),$$

and let $(E, \mathcal{E})$ be the product of $\{(E_i, \mathcal{E}_i) \mid i \in N_+\}$.
A set $A \in \mathcal{E}$ is said to be a cylinder set with base B if $B \in \mathcal{E}^n$ for some $n \in N_+$ and

$$A = B \times \left(\prod_{i=n+1}^{\infty} E_i \right).$$

n is then said to be the dimension of A, and we denote $A = [B]$. $\mathcal{F}_n$ denotes the set of all n-dimensional cylinder sets, and $\mathcal{F} = \bigcup_n \mathcal{F}_n$ is the set of all cylinder sets.
We can immediately tell that, for any $n, m \in N_+$ such that $n < m$, $\mathcal{F}_n \subset \mathcal{F}_m$. To see this, let $A \in \mathcal{F}_n$; by definition, there exists a $B \in \mathcal{E}^n$ such that $A = B \times E_{n+1} \times \cdots$. Then, for any $m > n$, we can write

$$A = [B] = [(B \times E_{n+1} \times \cdots \times E_m)],$$

where $B \times E_{n+1} \times \cdots \times E_m \in \mathcal{E}^m$ by the associativity of the product of σ-algebras. Therefore, $A \in \mathcal{F}_m$ as well.

Now we can define a function on $\mathcal{F}$ that yields all the desired finite dimensional distributions. Define $\mathbb{P}_0 : \mathcal{F} \to [0, 1]$ as

$$\mathbb{P}_0([B]) = \pi_n(B)$$

for any $[B] \in \mathcal{F}_n$. For any $n \in N_+$, the value of $\mathbb{P}_0$ for any n-dimensional cylinder set is equivalent to the value of its base under the product measure $\mu_1 \times \cdots \times \mu_n$. This is meant, in analogy with the finite case, to represent the fact that the first n random variables in the sequence have distributions $\mu_1, \cdots, \mu_n$ and are mutually independent.
Any cylinder set A can be a member of multiple $\mathcal{F}_n$, as seen above. Fortunately, it is easily seen that $\mathbb{P}_0$ assigns the same value to A regardless of which $\mathcal{F}_n$ it belongs to, so that the definition above is consistent: this is shown below.
Let $A \in \mathcal{F}_n$ for some $n \in N_+$, and choose any $m > n$. Then, letting the base of A be $B \in \mathcal{E}^n$

under $\mathcal{F}_n$, its base is $B \times E_{n+1} \times \cdots \times E_m$ under $\mathcal{F}_m$. We now have

$$\mathbb{P}_0([B]) = \pi_n(B) = \pi_m(B \times E_{n+1} \times \cdots \times E_m) = \mathbb{P}_0([B \times E_{n+1} \times \cdots \times E_m]),$$

so that $\mathbb{P}_0(A)$ is consistently defined.

Construction of the Probability Space

It remains to see that the $\mathcal{F}$ and $\mathbb{P}_0$ defined above can be extended to an actual probability space $(\Omega, \mathcal{H}, \mathbb{P})$, where $\mathcal{H}$ contains $\mathcal{F}$ and $\mathbb{P}$ agrees with $\mathbb{P}_0$ on $\mathcal{F}$. Then, letting $X_1, X_2, \cdots$ be defined as in the finite case, it can be easily seen that $\{X_n\}_{n \in N_+}$ is a sequence of independent random variables having the distributions $\{\mu_n\}_{n \in N_+}$.

Before diving into the proof of the desired result, we define some mathematical objects that assist with the proof. Retaining the setup established in the previous section, for any n-dimensional cylinder set $[B] \in \mathcal{F}_n$, define the sequence $\{Q_m(\cdot; [B])\}_{m \in N_+}$ of functions as follows: for any $m \in N_+$,

$$Q_m(x_1, \cdots, x_m; [B]) = \begin{cases} \int_{E_{m+1} \times \cdots \times E_n} I_B(x_1, \cdots, x_m, y) d(\mu_{m+1} \times \cdots \times \mu_n)(y) & \text{if } m < n \\ I_B(x_1, \cdots, x_n) & \text{if } m \geq n \end{cases}$$

for any $(x_1, \cdots, x_m) \in E^m$. Clearly, each $Q_m(\cdot; [B])$ is a $\mathcal{E}^m$-measurable function.
For any $m \in N_+$, if $m+1 < n$, then

$$\begin{aligned} Q_m(x_1, \cdots, x_m; [B]) &= \int_{E_{m+1}} I_B(x_1, \cdots, x_m, y) d(\mu_{m+1} \times \cdots \times \mu_n)(y) \\ &= \int_{E_{m+1}} \left[\int_{E_{m+2}} I_B(x_1, \cdots, x_m, x_{m+1}, y) d(\mu_{m+2} \times \cdots \times \mu_n)(y) \right] d\mu_{m+1}(x_{m+1}) \\ &= \int_{E_{m+1}} Q_{m+1}(x_1, \cdots, x_{m+1}; [B]) d\mu_{m+1}(x_{m+1}) \end{aligned}$$

by Fubini's theoerem, while if $m+1 \geq n$, then

$$Q_m(x_1, \cdots, x_m; [B]) = \int_{E_{m+1}} Q_{m+1}(x_1, \cdots, x_{m+1}; [B]) d\mu_{m+1}(x_{m+1})$$

trivially. Therefore, the above relationship holds for any $m \in N_+$ and $(x_1, \cdots, x_m) \in E^m$.
Finally, we can see that, for any n-dimensional cylinder set $[B] \in \mathcal{F}_n$,

$$\begin{aligned} \mathbb{P}_0([B]) = \pi_n(B) &= \int_{E_1 \times E_n} I_B(x) d(\mu_1 \times \cdots \times \mu_n)(x) \\ &= \int_{E_1} \left[\int_{E_2 \times \cdots \times E_n} I_B(x_1, y) d(\mu_2 \times \cdots \mu_n)(y) \right] d\mu_1(x_1) \\ &= \int_{E_1} Q_1(x_1; [B]) d\mu_1(x_1). \end{aligned}$$

The formal construction of the desired probability space is given below; it is a special case of Ionescu-Tulcea's theorem for the construction of sequences of random variables, and the proof remains virtually the same.

Theorem 1.22 (Ionescu-Tulcea's Theorem for Independent Random Variables) Let $\{(E_i, \mathcal{E}_i, \mu_i) \mid i \in N_+\}$ be a collection of probability spaces. Then, there exists a probability space $(\Omega, \mathcal{H}, \mathbb{P})$ and a sequence $\{X_n\}_{n \in N_+}$ of random variables with underlying probability space $(\Omega, \mathcal{H}, \mathbb{P})$ such that:

i) $\{X_n\}_{n \in N_+}$ is an independency.

ii) Each X_i takes values in $(E_i, \mathcal{E}_i)$ and has distribution μ_i.

Proof) Let the notations and definitions follow the setup given in the earlier sections. Our first goal is to show that $\mathcal{F}$, the set of all cylinder sets, is an algebra on E, and that $\mathbb{P}_0$ is a σ-additive pre-measure on $\mathcal{F}$. The probability space $(\Omega, \mathcal{H}, \mathbb{P})$ will then be defined as the extension of $\mathcal{F}$ and $\mathbb{P}_0$ obtained via Caratheodory's extension theorem.

Step 1: $\mathcal{F}$ is an algebra on E
Clearly, $E = [E_1] \in \mathcal{F}_1$.
For any $A \in \mathcal{F}$, letting $A = [B] \in \mathcal{F}_n$ for some $n \in N_+$ and $B \in \mathcal{E}^n$,

$$A^c = B^c \times E_{n+1} \times \cdots = [B^c] \in \mathcal{F}_n,$$

so that $\mathcal{F}$ is closed under complements.
Finally, for any $\{A_1, \cdots, A_n\} \subset \mathcal{F}$ with union A, let $m \in N_+$ be the highest dimension among $A_1, \cdots, A_n$. It follows that $A_1, \cdots, A_n \in \mathcal{F}_m$, so that there exist $B_1, \cdots, B_n \in \mathcal{E}^m$ such that $A_i = [B_i]$ for $1 \leq i \leq n$. Therefore,

$$A = \bigcup_{i=1}^{n} A_i = \left(\bigcup_{i=1}^{n} B_i \right) \times E_{m+1} \times \cdots = \left[\bigcup_{i=1}^{n} B_i \right] \in \mathcal{F}_m,$$

so that $A \in \mathcal{F}$ and $\mathcal{F}$ is closed under finite unions.
By definition, $\mathcal{F}$ is an algebra on E.

Step 2: $\mathbb{P}_0$ is a pre-measure on $\mathcal{F}$
Initially, we have

$$\mathbb{P}_0(\emptyset) = \mathbb{P}_0(\emptyset \times E_2 \times \cdots) = \mu_1(\emptyset) = 0.$$

For any disjiont $\{A_1, \cdots, A_n\} \subset \mathcal{F}$ with union A, letting m be the maximum dimension among $A_1, \cdots, A_n$ as before, there exist $B_1, \cdots, B_n \in \mathcal{E}^m$ such that $A_i = [B_i]$ for $1 \leq i \leq n$. Furthermore, for any $1 \leq i \neq j \leq n$,

$$A_i \cap A_j = [B_i \cap B_j] = \emptyset,$$

which means that $B_i \cap B_j = \emptyset$; $\{B_1, \cdots, B_n\}$ is a disjiont collection of sets in $\mathcal{E}^m$. As such, by the finite additivity of π_m,

$$\mathbb{P}_0(A) = \pi_m\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} \pi_m(B_i) = \sum_{i=1}^{n} \mathbb{P}_0(A_i).$$

It follows that $\mathbb{P}_0$ is finitely additive on $\mathcal{F}$, making it a pre-measure on that algebra.

Step 3: $\mathbb{P}_0$ is σ-additive

We first show that $\mathbb{P}_0$ is continuous sequences of sets in $\mathcal{F}$ that decrease to the empty set.

Let $\{A_n\}_{n \in N_+}$ be a sequence of cylinder sets such that $A_{n+1} \subset A_n$ for any $n \in N_+$ and $A = \bigcap_n A_n = \emptyset$. By the monotonicity of pre-measures, the sequence $\{\mathbb{P}_0(A_n)\}_{n \in N_+}$ in $[0, 1]$ is decreasing; therefore, its limit is well-defined as its infimum. We want to show that the limit is equal to 0, so assume otherwise, that is, suppose that

$$\lim_{n \to \infty} \mathbb{P}_0(A_n) > 0.$$

We now show that this leads to a contradiction via induction.

For any $n \in N_+$, we have

$$\mathbb{P}_0(A_n) = \int_{E_1} Q_1(x_1; A_n) d\mu_1(x_1).$$

Since $Q_1(x_1; A_n)$ is an integral of the x_1-section of the base of A_n, the sequence $\{Q_1(x_1; A_n)\}_{n \in N_+}$ takes values in $[0, 1]$ and decreases for any $x_1 \in E_1$. This means that the pointwise limit of $Q_1(\cdot; A_n)$ as $n \to \infty$ is well-defined as its infimum, and by the bounded convergence theorem,

$$\lim_{n \to \infty} \mathbb{P}_0(A_n) = \int_{E_1} \left(\inf_{n \in N_+} Q_1(x_1; A_n)\right) d\mu_1(x_1).$$

Because this integral is positive, there must exist an $x_1^* \in E_1$ such that

$$\inf_{n \in N_+} Q_1(x_1^*; A_n) > 0.$$

Now suppose, for some $k \geq 1$, that there exist $x_1^* \in E_1, \cdots, x_k^* \in E_k$ such that

$$\inf_{n \in N_+} Q_k(x_1^*, \cdots, x_k^*; A_n) > 0.$$

For any $n \in N_+$, we have

$$Q_k(x_1^*, \cdots, x_k^*; A_n) = \int_{E_{k+1}} Q_{k+1}(x_1^*, \cdots, x_k^*, y; A_n) d\mu_{k+1}(y),$$

and the sequence $\{Q_{k+1}(x_1^*, \cdots, x_k^*, \cdot; A_n)\}_{n \in N_+}$ of functions on E_{k+1} takes values in $[0, 1]$ and is decreasing pointwise by the same line of reasoning as above. Therefore, by the BCT,

$$\begin{aligned} \inf_{n \in N_+} Q_k(x_1^*, \cdots, x_k^*; A_n) &= \lim_{n \to \infty} Q_k(x_1^*, \cdots, x_k^*; A_n) \\ &= \int_{E_{k+1}} \left(\inf_{n \in N_+} Q_{k+1}(x_1^*, \cdots, x_k^*, y; A_n) \right) d\mu_{k+1}(y), \end{aligned}$$

and because this integral must be positive by the inductive hypothesis, there exists an $x_{k+1}^* \in E_{k+1}$ such that

$$\inf_{n \in N_+} Q_{k+1}(x_1^*, \cdots, x_k^*, x_{k+1}^*; A_n) > 0.$$

Therefore, by induction, there exists a sequence $x^* = \{x_i^*\}_{i \in N_+} \in E$ such that

$$Q_k(x_1^*, \cdots, x_k^*; A_n) > 0$$

for any $k \in N_+$ and $n \in N_+$. This means that, for any $n \in N_+$, letting $A_n = [B] \in \mathcal{F}_m$ for some $m \in N_+$,

$$Q_m(x_1^*, \cdots, x_m^*; [B]) = I_B(x_1^*, \cdots, x_m^*) = 1,$$

or that $(x_1^*, \cdots, x_m^*) \in B$. As such,

$$x^* \in B \times E_{m+1} \times \cdots = A_n,$$

and because this holds for any $n \in N_+$, x^* is contained in $A = \bigcap_n A_n$. This contradicts the assumption that $A = \emptyset$, so it must be the case that

$$\lim_{n \to \infty} \mathbb{P}_0(A_n) = 0.$$

It is easy now to show that this implies σ-additivity. For any disjiont sequence $\{A_n\}_{n \in N_+} \subset$

$\mathcal{F}$ such that $A = \bigcup_n A_n \in \mathcal{F}$. Define $\{H_n\}_{n\in N_+}$ as

$$H_n = A \setminus \left(\bigcup_{i=1}^{n} A_i \right)$$

for any $n \in N_+$. Becuase algebras are closed under finite unions and set differences, $\{H_n\}_{n\in N_+}$ is a sequence of sets in $\mathcal{F}$ that decrease to $\emptyset$. By the finite additivity of $\mathbb{P}_0$ on $\mathcal{F}$,

$$\begin{aligned}\mathbb{P}_0(A) &= \mathbb{P}_0\left(A \setminus \left(\bigcup_{i=1}^{n} A_i \right) \right) + \mathbb{P}_0\left(\bigcup_{i=1}^{n} A_i \right) \\ &= \mathbb{P}_0(H_n) + \sum_{i=1}^{n} \mathbb{P}_0(A_i).\end{aligned}$$

By the preceding result, the first term goes to 0 as $n \to \infty$, so it follows that

$$\mathbb{P}_0(A) = \sum_{n=1}^{\infty} \mathbb{P}_0(A_n),$$

which proves σ-additivity.

Step 4: Construction of the Probability Space and Random Variables
Because $\mathcal{F}$ is an algebra on E and $\mathbb{P}_0$ a σ-additive pre-measure on $\mathcal{F}$, by Caratheodory's extension theorem there exists a complete measure space $(E, \mathcal{M}, \mu)$ such that $\mathcal{F} \subset \mathcal{M}$ and $\mu(A) = \mathbb{P}_0(A)$ for any $A \in \mathcal{F}$. In particular, because $E \in \mathcal{F}$, $\mu(E) = \mathbb{P}_0(E) = 1$; defining $\Omega = E$, $\mathcal{H} = \mathcal{M}$ and $\mathbb{P} = \mu$, the triple $(\Omega, \mathcal{H}, \mathbb{P})$ is a probability space.

Now define the sequence $\{X_n\}_{n\in N_+}$ of functions on Ω as follows: for any $n \in N_+$,

$$X_n(\omega) = \omega_n$$

for any $\omega = \{\omega_i\}_{i\in N_+} \in \Omega$. Then, X_n is a function taking values in $(E_n, \mathcal{E}_n)$, and because

$$X_n^{-1}(A_n) = [E_1 \times \cdots \times E_{n-1} \times A_n] \in \mathcal{F}_n \subset \mathcal{H}$$

for any $A_n \in \mathcal{E}_n$, X_n is a random variable.
Furthermore, for any $A_n \in \mathcal{E}_n$, because $X_n^{-1}(A_n)$ is a cylinder set,

$$\begin{aligned}\mathbb{P}\left(X_n^{-1}(A_n) \right) &= \mathbb{P}_0(X_n^{-1}(A_n)) \\ &= \mathbb{P}_0([E_1 \times \cdots \times E_{n-1} \times A_n]) = \pi_n(E_1 \times \cdots \times E_{n-1} \times A_n) = \mu_n(A_n).\end{aligned}$$

Therefore, X_n has distribution μ_n.

Finally, choose any finite set $J = \{t_1, \cdots, t_n\} \subset N_+$ and $A_1 \in \mathcal{E}_{t_1}, \cdots, A_n \in \mathcal{E}_{t_n}$. Letting $t_1 < \cdots < t_n = m$ without loss of generality,

$$\bigcap_{i=1}^{n} X_{t_i}^{-1}(A_i) = \left[\prod_{j=1}^{m} B_j\right],$$

where $B_{t_i} = A_{t_i}$ for any $1 \leq i`k$ and $B_j = E_j$ if $j \notin J$. Therefore,

$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^{n} X_{t_i}^{-1}(A_i)\right) &= \mathbb{P}_0\left(\left[\prod_{n=1}^{m} B_i\right]\right) \\ &= \pi_m\left(\prod_{n=1}^{m} B_i\right) = \prod_{i=1}^{n} \mu_i(A_{t_i}) = \prod_{i=1}^{n} \mathbb{P}\left(X_{t_i} \in A_i\right).\end{aligned}$$

By definition, $\{X_n\}_{n \in N_+}$ is an independency.
Q.E.D.

1.9 Normal Distributions

Normal distributions, also called Gaussian distributions, form the crux of many theorems in probability theory. They also have a rich structure of their own, so we take a moment to study the normal distribution and its multivariate generalizations.

1.9.1 The Standard Univariate Normal Distribution

A real random variable X is said to have the standard normal distribution if its density $f : \mathbb{R} \to [0, +\infty]$ with respect to the Lebesgue measure is defined as

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

for any $x \in \mathbb{R}$; this relationship is succinctly stated as $X \sim \mathcal{N}[0, 1]$.
f is continuous on $\mathbb{R}$, so by lemma 1.3, the distribution function F of X is differentiable on $\mathbb{R}$ with derivative f. This implies that F is continuous, and as such that X is a continuous random variable.

If $X \sim \mathcal{N}[0, 1]$, X has finite kth moments for every $k \in N_+$:

$$\begin{aligned}
\mathbb{E}\left[|X|^k\right] &= \int_{-\infty}^{\infty} |x|^k f(x) dx \\
&= 2\left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \int_0^{\infty} x^k \exp\left(-\frac{1}{2}x^2\right) dx \\
&= 2^{\frac{k}{2}+1}\left(\frac{1}{2\pi}\right)^{\frac{1}{2}} \int_0^{\infty} z^{\frac{k}{2}} \exp(-z)(2z)^{-\frac{1}{2}} dz \qquad \text{(Substitution with } \sqrt{2z} = x\text{)} \\
&= 2^{\frac{k}{2}} \frac{1}{\sqrt{\pi}} \int_0^{\infty} z^{\frac{k+1}{2}-1} \exp(-z) dz \\
&= 2^{\frac{k}{2}} \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{k+1}{2}\right),
\end{aligned}$$

where $\Gamma : \mathbb{R} \to \mathbb{R}$ denotes the gamma function. We are thus able to obtain some elementary absolute moments:

$$\begin{aligned}
\mathbb{E}[|X|] &= \sqrt{\frac{2}{\pi}}\Gamma(1) = \sqrt{\frac{2}{\pi}} \\
\mathbb{E}\left[X^2\right] &= \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) = \frac{2}{\sqrt{\pi}}\frac{1}{2} \cdot \Gamma\left(\frac{1}{2}\right) = 1 \\
\mathbb{E}\left[|X|^3\right] &= 2^{\frac{3}{2}} \frac{1}{\sqrt{\pi}}\Gamma(2) = 2\sqrt{\frac{2}{\pi}} \\
\mathbb{E}\left[X^4\right] &= \frac{4}{\sqrt{\pi}}\Gamma\left(\frac{5}{2}\right) = \frac{4}{\sqrt{\pi}}\frac{3}{2}\Gamma\left(\frac{3}{2}\right) = 3.
\end{aligned}$$

In addition, because $x \mapsto x^k \exp\left(-\frac{1}{2}x^2\right)$ is an odd function on $\mathbb{R}$ for any odd $k \in N_+$, the odd

moments of X are all 0, that is,

$$\mathbb{E}\left[X^k\right] = 0 \quad \text{if } k \text{ is odd.}$$

These yield the familiar results concerning the standard normal distribution; if $X \sim \mathcal{N}[0,1]$, then:

- The mean of X is $\mu = \mathbb{E}[X] = 0$
- The variance of X is $\sigma^2 = \mathbb{E}[X^2] - \mu^2 = 1$
- The skewness of X is $\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \mathbb{E}[X^3] = 0$
- The kurtosis of X is $\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \mathbb{E}[X^4] = 3$.

For $X \sim \mathcal{N}[0,1]$ with density $\phi : \mathbb{R} \to [0,+\infty)$, the distribution function $\Phi : \mathbb{R} \to [0,1]$ of X is defined as

$$\Phi(x) = \int_{-\infty}^{x} \phi(z)dz = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)dz$$

for any $x \in \mathbb{R}$. As stated above, $\Phi'(x) = \phi(x)$ for any $x \in \mathbb{R}$.

$X \sim \mathcal{N}[0,1]$ also has simple and tractable moment generating and characteristic functions.

Lemma 1.22 (MGF and Characteristic Function of Normal Distribution)
Let $X \sim \mathcal{N}[0,1]$ have moment generating function and characteristic function $m : \mathbb{R} \to [0,+\infty]$ and $\varphi : \mathbb{R} \to \mathbb{C}$, respectively. Then,

$$m(t) = \exp\left(\frac{1}{2}t^2\right) \quad \text{and} \quad \varphi(t) = \exp\left(-\frac{1}{2}t^2\right)$$

for any $t \in \mathbb{R}$.

Proof) We start with the moment generating function. For any $t \in \mathbb{R}$,

$$\begin{aligned}
m(t) = \mathbb{E}[\exp(tX)] &= \int_{-\infty}^{\infty} \exp(tx)\phi(x)dx \\
&= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x^2+2tx)\right)dx \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}t^2\right) \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x+t)^2\right)dx \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}t^2\right) \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}z^2\right)dz \\
&\qquad\qquad \text{(Linear change of variables)} \\
&= \exp\left(\frac{1}{2}t^2\right) \cdot \int_{-\infty}^{\infty} \phi(z)dz \\
&= \exp\left(\frac{1}{2}t^2\right).
\end{aligned}$$

Moving onto the characteristic function, note that, for any non-zero $t \in \mathbb{R}$,

$$\begin{aligned}
\varphi(t) &= \int_{-\infty}^{\infty} \exp(itx)\phi(x)dx \\
&= \int_{-\infty}^{\infty} \cos(tx)\phi(x)dx + i \cdot \int_{-\infty}^{\infty} \sin(tx)\phi(x)dx \qquad \text{(Euler's formula)} \\
&= 2\int_{0}^{\infty} \cos(tx)\phi(x)dx \qquad (\cos(tx)\phi(x) \text{ is even, while } \sin(tx)\phi(x) \text{ is odd}) \\
&= \sqrt{\frac{2}{\pi}} \cdot \int_{0}^{\infty} \cos(tx)\exp\left(-\frac{1}{2}x^2\right)dx.
\end{aligned}$$

This immediately tells us that $\varphi(t)$ is real valued.
Integration by parts similar to that used in the derivation of the Dirichlet integral reveals that

$$\int_{0}^{\infty} \cos(tx)\exp\left(-\frac{1}{2}x^2\right)dx - \frac{1}{t}\int_{0}^{\infty} x\sin(tx)\exp\left(-\frac{1}{2}x^2\right)dx = 0,$$

so that

$$\varphi(t) = \frac{1}{t} \cdot \left[\sqrt{\frac{2}{\pi}}\int_{0}^{\infty} x\sin(tx)\exp\left(-\frac{1}{2}x^2\right)dx\right].$$

Now we turn away from φ for a moment to obtain the derivative of φ at t. For any non-zero $h \in \mathbb{R}$,

$$\frac{\varphi(t+h)-\varphi(t)}{h} = \sqrt{\frac{2}{\pi}} \cdot \int_{0}^{\infty} \frac{\cos((t+h)x)-\cos(tx)}{h}\exp\left(-\frac{1}{2}x^2\right)dx.$$

Since

$$\left|\frac{\cos((t+h)x)-\cos(tx)}{h}\exp\left(-\frac{1}{2}x^2\right)\right| \leq x\exp\left(-\frac{1}{2}x^2\right)$$

for any $h \neq 0$ and $x \geq 0$, where

$$\int_{0}^{\infty} x\exp\left(-\frac{1}{2}x^2\right)dx = 1 < +\infty,$$

and

$$\lim_{h\to 0}\frac{\cos((t+h)x)-\cos(tx)}{h} = \frac{\partial\cos(tx)}{\partial t} = -x\sin(tx)$$

for any $x \geq 0$, so by the DCT,

$$\varphi'(t) = \lim_{h\to 0}\frac{\varphi(t+h)-\varphi(t)}{h} = -\sqrt{\frac{2}{\pi}} \cdot \int_{0}^{\infty} x\sin(tx)\exp\left(-\frac{1}{2}x^2\right)dx.$$

From the result derived earlier from integration by parts, we can now see that

$$\varphi'(t) = -t \cdot \varphi(t).$$

Suppose that $\varphi(t) > 0$, so that

$$-t = \frac{\varphi'(t)}{\varphi(t)} = \frac{d}{dt}\log(\varphi(t))$$

(the derivative on the right is well-defined because the continuity of φ ensures that φ is positive on a neighborhood of 0). It follows that

$$\log(\varphi(t)) = -\frac{1}{2}t^2 + c,$$

or equivalently,

$$\varphi(t) = C \cdot \exp\left(-\frac{1}{2}t^2\right)$$

for some $c \in \mathbb{R}$ and $C > 0$ ($C = \exp(c)$ in this case). Since φ is continuous on $\mathbb{R}$ and $\varphi(0) = 1$, we must have

$$1 = \lim_{t \to 0} \varphi(t) = C,$$

and therefore,

$$\varphi(t) = \exp\left(-\frac{1}{2}t^2\right)$$

for any $t \in \mathbb{R}$.
Q.E.D.

1.9.2 General Univariate Normal Distributions

Normal distributions are easy to work with because all normally distributed real random variables can be expressed as an affine function of a standard normally distributed random variable. Let $Z \sim \mathcal{N}[0,1]$. For any $\mu \in \mathbb{R}$ and $\sigma^2 \geq 0$, the random variable defined as

$$X = \sigma \cdot Z + \mu$$

would have mean μ and variance σ^2. We say that X has the normal distribution with mean μ and variance σ^2, and denote this by $X \sim \mathcal{N}[\mu, \sigma^2]$.
Suppose $\sigma^2 > 0$. Letting ϕ be the standard normal density and v the distribution of X, note that, for any Borel set $A \subset \mathbb{R}$,

$$v(A) = \mathbb{E}[I_A \circ X] = \mathbb{E}[I_A \circ (\sigma Z + \mu)]$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} I_A(\sigma \cdot z + \mu)\, \phi(z) dz \\
&= \frac{1}{\sigma} \cdot \int_{-\infty}^{\infty} I_A(x) \phi\left(\frac{x-\mu}{\sigma}\right) dx \qquad \text{(Substitution with } \tfrac{x-\mu}{\sigma} = z\text{)} \\
&= \int_A \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx.
\end{aligned}
$$

As such,

$$
\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)
$$

is precisely the density of $X \sim \mathcal{N}[\mu, \sigma^2]$ if $\sigma^2 > 0$.
If $\sigma^2 = 0$, then because $X = \mu$ almost surely, the distribution of X is not absolutely continuous with respect to the Lebesgue measure and thus X does not admit a density.
Given $X \sim \mathcal{N}[\mu, \sigma^2]$ for some $\sigma^2 > 0$, we can conversely define $Z = \frac{X-\mu}{\sigma}$ and find that $Z \sim \mathcal{N}[0,1]$.
The operation $x \mapsto \frac{x-\mu}{\sigma}$ is referred to as the normalization of the random variable X.

The moment generating and characteristic functions of $X \sim \mathcal{N}[\mu, \sigma^2]$ are now easily found. Letting $Z = \frac{X-\mu}{\sigma}$, $Z \sim \mathcal{N}[0,1]$ and $X = \sigma Z + \mu$, so the moment generating function $m : \mathbb{R} \to [0, +\infty]$ and the characteristic function $\varphi : \mathbb{R} \to \mathbb{C}$ of X are defined as

$$
\begin{aligned}
m(t) = \mathbb{E}[\exp(tX)] &= \exp(\mu t) \cdot \mathbb{E}[\exp(t\sigma Z)] \\
&= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \\
\varphi(t) = \mathbb{E}[\exp(itX)] &= \exp(i\mu t) \cdot \mathbb{E}[\exp(it\sigma Z)] \\
&= \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right)
\end{aligned}
$$

for any $t \in \mathbb{R}$, where we used the formulas for the mgf and characteristic function of the standard normal distribution.

An important characteristic of the normal distribution is that the linear combination of any independent normally distributed random variables is also normally distributed.

Theorem 1.23 (Preservation of Normality under Linear Combinations)
Let X, Y be two independent real random variables such that $X \sim \mathcal{N}[\mu_x, \sigma_x^2]$ and $Y \sim \mathcal{N}\left[\mu_y, \sigma_y^2\right]$. Then,

$$
\alpha X + Y \sim \mathcal{N}\left[\alpha\mu_x + \mu_y, \alpha^2\sigma_x^2 + \sigma_y^2\right].
$$

Proof) If $\alpha = 0$, then $\alpha X + Y = Y$, and the result follows trivially.
Suppose that $\alpha \neq 0$. Then, letting φ be the characteristic function of the real random

variable $\alpha X + Y$ and φ_x, φ_y the characteristic functions of X and Y, we have

$$\begin{aligned}\varphi(t) = \mathbb{E}\left[\exp(it'Z)\right] &= \mathbb{E}\left[\exp(it\alpha X)\exp(itY)\right] \\ &= \mathbb{E}\left[\exp(it\alpha X)\right]\mathbb{E}\left[\exp(itY)\right] = \varphi_x(t\alpha)\varphi_y(t)\end{aligned}$$

for any $t \in \mathbb{R}$, where the third equality follows because X and Y are independent. Since X and Y are normally distributed, their characteristic functions are given by

$$\begin{aligned}\varphi_x(t) &= \exp\left(i\mu_x - \frac{1}{2}\sigma_x^2 t^2\right) \\ \varphi_y(t) &= \exp\left(i\mu_y - \frac{1}{2}\sigma_y^2 t^2\right)\end{aligned}$$

for any $t \in \mathbb{R}$. It follows that

$$\varphi(t) = \exp\left(i(\alpha\mu_x + \mu_y)t - \frac{1}{2}(\alpha^2\sigma_x^2 + \sigma_y^2)t^2\right)$$

for any $t \in \mathbb{R}$, which is exactly the characteristic function of a normal distrubtion with mean $\alpha\mu_x + \mu_y$ and variance $\alpha^2\sigma_x^2 + \sigma_y^2$. Therefore,

$$\alpha X + Y \sim \mathcal{N}\left[\alpha\mu_x + \mu_y, \alpha^2\sigma_x^2 + \sigma_y^2\right].$$

Q.E.D.

This easily generalizes to finite linear combinations, so that, for any collection of normally distributed random variables $X_1, \cdots, X_n$ such that $X_i \sim \mathcal{N}\left[\mu_i, \sigma_i^2\right]$ for $1 \leq i \leq n$ and real numbers $\alpha_1, \cdots, \alpha_n$, we have

$$\sum_{i=1}^{n} \alpha_i X_i \sim \mathcal{N}\left[\sum_{i=1}^{n} \alpha_i \mu_i, \sum_{i=1}^{n} \alpha_i \cdot \sigma_i^2\right].$$

In algebraic terms, we can say that the collection

$$\mathcal{N} = \{X \in \mathcal{H}/\mathcal{B}(\mathbb{R}) \mid X \sim \mathcal{N}\left[\mu, \sigma^2\right] \text{ for some } \mu \in \mathbb{R}, \sigma^2 \geq 0\}$$

forms a linear subspace of the set of all real random variables.

1.9.3 The Multivariate Normal Distribution

We now study the multivariate generalization of the normal distribution. A k-dimensional real random vector X is said to be normally distributed if, for any $t \in \mathbb{R}^k$, $t'X$ is a (possibly degenerate) normally distributed random variable. Taking t to be equal to each standard basis vector on $\mathbb{R}^k$, it follows that each coordinate of X is normally distributed. Thus, each coordinate of X and by extension X itself has finite absolute moments of all integer orders, so that we can define the mean and variance of X; denote them by $\mu \in \mathbb{R}^k$ and $V \in \mathbb{R}^{k\times k}$ (in particular, recall that V is positive definite). We can then denote $X \sim \mathcal{N}[\mu, V]$, in analogy with the univariate case.
If X is a normally distributed k-dimensional random vector with mean $\mathbf{0}$ and variance I_k, then we say that X has the standard normal distribution. The coordiantes of X are all standard normally distributed real random variables, and they are pairwise uncorrelated.

We can easily obtain the MGF and characteristic function of X. For any $t \in \mathbb{R}^k$, $t'X$ is a normally distributed random variable such that

$$\mathbb{E}[t'X] = t'\mu \quad \text{and} \quad \operatorname{Var}[t'X] = t'\operatorname{Var}[X]\,t = t'Vt,$$

so letting $m : \mathbb{R}^k \to [0, +\infty]$ and $\varphi : \mathbb{R}^k \to \mathbb{C}$ be the MGF and characteristic functions of X, we have

$$m(t) = \mathbb{E}[\exp(t'X)] = \exp\left(t'\mu + \frac{1}{2}t'Vt\right)$$
$$\varphi(t) = \mathbb{E}[\exp(it'X)] = \exp\left(it'\mu - \frac{1}{2}t'Vt\right).$$

for any $t \in \mathbb{R}^k$. These expressions are easily seen as multivariate generalizations of the MGF and characteristic function of a univariate normally distributed random variable.
The formula for the characteristic function of normally distributed random vectors derived above yield the following independence result, which is of great importance:

Theorem 1.24 (Independence and Uncorrelatedness for Normal Vectors)
Suppose X and Y are k-and m-dimensional random vectors, and that the $k+m$-dimensional random vector $Z = (X', Y')'$ is normally distributed. Then, X and Y are independent if and only if they are uncorrelated.

Proof) Necessity follows immediately. To see sufficiency, suppose that X and Y are uncorrelated. For any $t \in \mathbb{R}^k$ and $r \in \mathbb{R}^m$,

$$t'X = \begin{pmatrix} t' & \mathbf{0}' \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \quad \text{and} \quad r'Y = \begin{pmatrix} \mathbf{0}' & r' \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix},$$

and because $Z = (X', Y')'$ is normally distributed, $t'X$ and $r'Y$ are normally distributed real random variables. By definition, X and Y are normally distributed, and let μ_x, μ_y be the means of X, Y and V_x, V_y their variances, so that Z has mean $(\mu_x', \mu_y')'$ and

variance $V = \begin{pmatrix} V_x & O \\ O & V_y \end{pmatrix}$. Let φ_x and φ_y be the characteristic functions of X and Y, and φ the characteristic function of the $k+m$-dimensional random vector $Z = (X,Y)$. Then, for any $t \in \mathbb{R}^k$ and $r \in \mathbb{R}^m$, we have

$$\begin{aligned}\varphi\left(\begin{pmatrix} t \\ r \end{pmatrix}\right) &= \exp\left(i\begin{pmatrix} t' & r' \end{pmatrix}\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} - \frac{1}{2}\begin{pmatrix} t' & r' \end{pmatrix}\begin{pmatrix} V_x & O \\ O & V_y \end{pmatrix}\begin{pmatrix} t \\ r \end{pmatrix}\right) \\ &= \exp\left(it'\mu_x - \frac{1}{2}t'V_x t\right)\exp\left(ir'\mu_y - \frac{1}{2}r'V_y r\right) = \varphi_x(t)\varphi_y(r).\end{aligned}$$

Therefore, X and Y are independent.
Q.E.D.

The above theorem implies that, if $Z \sim \mathcal{N}[\mathbf{0}, I_k]$, then the coordinates $Z_1, \cdots, Z_k$ of Z are independent standard normally distributed normal random variables. Letting ϕ be the density of a standard normally distributed real random variable, the independence of $Z_1, \cdots, Z_k$ implies that Z is absolutely continuous with respect to the Lebesgue measure with unique continuous density $f : \mathbb{R}^k \to [0, +\infty)$ defined as

$$f(x) = \prod_{i=1}^{k} \phi(x_i) = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} \exp\left(-\frac{1}{2}x'x\right)$$

for any $x \in \mathbb{R}^k$.

As in the univariate case, affine functions of normally distributed random vectors is also a random vector. To see this, let X be a k-dimensional normally distributed random vector, and $A \in \mathbb{R}^m$ and $C \in \mathbb{R}^{m \times k}$ for some $m \in N_+$. Define $Y = A + CX$; then, for any $t \in \mathbb{R}^m$,

$$t'Y = t'A + (t'C)X,$$

and because $(t'C)X$ is a normally distributed real random variable, so is $t'Y$. By definition, Y is a normally distributed m-dimensional random vector.
This invariance under affine transformations leads to the following characterization for normally distributed random vectors:

Theorem 1.25 (Characterization of Multivariate Normality)
Let X be a k-dimensional real random vector. Then, X is normally distributed if and only if there exists an $A \in \mathbb{R}^k$, a matrix $C \in \mathbb{R}^{k \times m}$ of full rank m, where $m \leq k$, and an m-dimensional standard normally distributed random vector Z such that $X = A + CZ$. In particular, if the variance of X is positive definite, then we can take $m = k$.

Proof) Suppose that there exists an $A \in \mathbb{R}^k$, a matrix $C \in \mathbb{R}^{k \times m}$, where $m \leq k$, and an m-

dimensional standard normally distributed random vector Z such that $X = A + CZ$. It immediately follows that X is normally distributed because normality is preserved under affine transformations.

Conversely, suppose that X is normally distributed with mean $\mu \in \mathbb{R}^k$ and variance $V \in \mathbb{R}^{k \times k}$. Since V is positive semidefinite, it admits an eigendecomposition $V = PDP'$ for orthogonal matrix $P \in \mathbb{R}^{k \times k}$ and diagonal matrix D whose diagonal elements equal the eigenvalues of V. Letting $m \leq k$ be the rank of V, D consists of exactly m diagonal elements and $k - m$ diagonal entries equal to 0. Assume that the non-zero elements of D are ordered first; letting these entries be $\lambda_1 \geq \cdots \geq \lambda_m > 0$, define the $k \times m$ matrix

$$\tilde{D} = \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_m} \end{pmatrix},$$

and let $\tilde{P} \in \mathbb{R}^{k \times m}$ collect the first m eigenvectors comprising P. Then, $\tilde{D}$ and $\tilde{P}$ have full rank m, and it follows that

$$V = PDP' = \tilde{P}\tilde{D}^2\tilde{P}' = CC',$$

where $C = \tilde{P}\tilde{D} \in \mathbb{R}^{k \times m}$ also has full rank m. Now define

$$Z = \tilde{D}^{-1}\tilde{P}'(X - \mu).$$

It follows that Z is a normally distributed random vector with mean $\mathbf{0}$ and variance

$$\mathrm{Var}[Z] = \tilde{D}^{-1}\tilde{P}'\mathrm{Var}[X]\,\tilde{P}\tilde{D}^{-1} = \tilde{D}^{-1}\tilde{P}'\tilde{P}\tilde{D}^2\tilde{P}'\tilde{P}\tilde{D}^{-1} = I_m,$$

where we used the fact that $\tilde{P}'\tilde{P} = I_m$ by the orthonormality of the eigenvectors comprising $\tilde{P}$. It follows that Z is an m-dimensional standard normally distributed random vector, and

$$\mu + CZ = \mu + \tilde{P}\tilde{P}'(X - \mu) = X.$$

If V is positive definite, then because all the eigenvalues of V are positive, we can take $m = k$ above. In fact, in the construction above C ends up being the Cholesky factor of V.
Q.E.D.

This characterization is actually our definition of normally distributed real random variables, and thus reveals that our definition for multivariate normally distributed random vectors is reasonable.

As in the univariate case, the density of a normally distributed random vector exists if and only if its variance is non-singular; in light of the positive semidefiniteness of the variance of a random vector, this is equivalent to the positive definiteness of the variance.

Theorem 1.26 (Density of the Multivariate Normal Distribution)
Let X be a k-dimensional normally distributed real random vector with mean $\mu \in \mathbb{R}^k$ and variance $V \in \mathbb{R}^{k\times k}$. Then, X has a density with respect to the Lebesgue measure if and only if V is positive definite, and in this case, the density $f : \mathbb{R}^k \to [0, +\infty)$ of X is given by

$$f(x) = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} |\det(V)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)\right)$$

for any $x \in \mathbb{R}^k$.

Proof) Suppose that X has a density f with respect to the Lebesgue measure, which implies that the distribution v of X is absolutely continuous with respect to the Lebesgue measure. Assume that V is not positive definite, or that it is singular. Then, there exists a non-zero $\alpha \in \mathbb{R}^k$ such that $\alpha' V \alpha = 0$. Defining $\alpha' X = Z$, Z is normally distributed with mean $\alpha'\mu$ and variance $\alpha' V \alpha = 0$, which tells us that $Z = \alpha'\mu$ almost surely. In other words, X takes values in the hyperspace $H = \{x \in \mathbb{R}^k \mid \alpha' x = \alpha'\mu\}$ almost surely;

$$v(H) = \mathbb{P}(X^{-1}(H)) = 1.$$

H is a $k-1$-dimensional subspace of $\mathbb{R}^k$, so letting $\{u_1, \cdots, u_{k-1}\} \subset \mathbb{R}^k$ be a basis generating H and defining the $k \times k$ matrix P as

$$P = \begin{pmatrix} u_1 & \cdots & u_{k-1} & \mathbf{0} \end{pmatrix},$$

we can express $H = P\mathbb{R}^k$. It follows that

$$\lambda_k(H) = |\det(P)|\lambda_k(\mathbb{R}^k) = 0.$$

By the absolute continuity of v with respect to λ_k, we must have $v(H) = 0$, a contradiction. Therefore, V must be positive definite.

Conversely, suppose that V is positive definite. Then, there exists a k-dimensional standard normally distributed random vector Z such that

$$X = \mu + CZ,$$

where C is the Cholesky factor of V. Letting v be the distribution of X and $A \in \mathcal{B}(\mathbb{R}^k)$,

by a linear change of variables we have

$$\begin{aligned} v(A) = \mathbb{E}[I_A \circ X] &= \mathbb{E}[I_A \circ (\mu + CZ)] \\ &= \int_{\mathbb{R}^k} I_A(\mu + cz) \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} \exp\left(-\frac{1}{2} z'z\right) dz \\ &= \int_{\mathbb{R}^k} I_A(x) \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} \exp\left(-\frac{1}{2}(x-\mu)'C^{-1\prime}C^{-1}(x-\mu)\right) \left|\det\left(C^{-1}\right)\right| dx \\ &= \int_A \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} \left|\det\left(C^{-1}\right)\right| \exp\left(-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)\right) dx. \end{aligned}$$

This implies that v is absolutely continuous with respect to the Lebesgue measure, and that

$$\left(\frac{1}{2\pi}\right)^{\frac{k}{2}} \left|\det\left(C^{-1}\right)\right| \exp\left(-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)\right)$$

is precisely the density of v with respect to the Lebesgue measure.
Finally, we can see that

$$\det(V) = \det(CC') = \det(C)^2 = \left(\det\left(C^{-1}\right)\right)^{-2},$$

so the density f can be written as

$$f(x) = \left(\frac{1}{2\pi}\right)^{\frac{k}{2}} |\det(V)|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)\right)$$

for any $x \in \mathbb{R}^k$.
Q.E.D.

1.10 Uniform Integrability

Here we introduce a concept crucial to the theory of martingales and convergence in L^p. First, we present a characterization of integrability.

Lemma 1.27 (Characterization of Integrability)
Let X be a non-negative random variable. X is integrable if and only if

$$\lim_{n\to\infty} \mathbb{E}\left[X \cdot I_{\{X>n\}}\right] = 0.$$

Proof) Suppose that X is integrable, that is, $\mathbb{E}[X] < +\infty$. Defining the sequence $\{X_n\}_{n\in N_+}$ of non-negative random variables defined as

$$X_n = X \cdot I_{\{X>n\}}$$

for any $n \in N_+$, note that $|X_n| \leq X$ for any $n \in N_+$ and $X_n \to 0$ pointwise. By the DCT, it now follows that

$$\lim_{n\to\infty} \mathbb{E}\left[X \cdot I_{\{X>n\}}\right] = \mathbb{E}\left[\lim_{n\to\infty} X_n\right] = 0.$$

Conversely, suppose that

$$\lim_{n\to\infty} \mathbb{E}\left[X \cdot I_{\{X>n\}}\right] = 0.$$

There then exists an $N \in N_+$ such that

$$\mathbb{E}\left[X \cdot I_{\{X>n\}}\right] < 1$$

for any $n \geq N$. Since

$$X = X \cdot I_{\{X>N\}} + X \cdot I_{\{X\leq N\}} \leq X \cdot I_{\{X>N\}} + N,$$

taking expectations on both sides yields

$$\mathbb{E}[X] = \mathbb{E}\left[X \cdot I_{\{X>N\}}\right] + N < N + 1 < +\infty.$$

Q.E.D.

Since any real or complex random variable X is integrable if and only if $|X|$ is integrable,

the above characterization suggests that X is integrable if and only if

$$\lim_{n\to\infty} \mathbb{E}\left[|X| \cdot I_{\{|X|>n\}}\right] = 0.$$

The same goes for real random vectors with $|\cdot|$ now standing for the euclidean norm.

A collection $\mathcal{K}$ of complex random variables or real random vectors is said to be uniformly integrable if

$$\sup_{X\in\mathcal{K}} \mathbb{E}\left[|X| \cdot I_{\{|X|>n\}}\right] \to 0$$

as $n \to \infty$. In other words, every element of $\mathcal{K}$ is integrable "at approximately the same rate". There are many convenient properties of uniformly integrable collections of random variables:

Theorem 1.28 (Properties of Uniform Integrability)
Let $\mathcal{K}$ be a collection of complex random variables and real random vectors. Then, the following hold true:

i) If $\mathcal{K}$ is uniformly integrable, then it is L^1-bounded, that is,

$$\sup_{X\in\mathcal{K}} \mathbb{E}|X| < +\infty.$$

ii) If there exists an integrable non-negative random variable Z such that $|X| \leq Z$ for any $X \in \mathcal{K}$, then $\mathcal{K}$ is uniformly integrable.

iii) If there exists a non-negative function $f : \mathbb{R} \to \mathbb{R}_+$ such that

$$\lim_{x\to\infty} \frac{f(x)}{x} = +\infty \quad \text{and} \quad \sup_{X\in\mathcal{K}} \mathbb{E}\left[f \circ |X|\right] < +\infty,$$

then $\mathcal{K}$ is uniformly integrable.

iv) If, for some $p > 1$, $\mathcal{K}$ is L^p-bounded, that is,

$$\sup_{X\in\mathcal{K}} \mathbb{E}|X|^p < +\infty,$$

then $\mathcal{K}$ is uniformly integrable.

Proof) i) Suppose $\mathcal{K}$ is uniformly integrable. Then, there exists an $N \in N_+$ such that

$$\mathbb{E}\left[|X| \cdot I_{\{|X|>n\}}\right] < 1$$

for any $n \geq N$ and $X \in \mathcal{K}$. For any $X \in \mathcal{K}$, since

$$|X| = |X| \cdot I_{\{|X|>N\}} + |X| \cdot I_{\{|X|\leq N\}} \leq |X| \cdot I_{\{|X|>N\}} + N,$$

it follows that

$$\mathbb{E}|X| \leq \mathbb{E}\left[|X| \cdot I_{\{|X|>n\}}\right] + N < 1 + N.$$

This holds for any $X \in \mathcal{K}$, so

$$\sup_{X \in \mathcal{K}} \mathbb{E}|X| \leq 1 + N < +\infty,$$

and $\mathcal{K}$ is L^1-bounded.

ii) Suppose that the elements of $\mathcal{K}$ are dominated by an integrable non-negative random variable Z. Then, for any $n \in N_+$ and $X \in \mathcal{K}$,

$$|X| \cdot I_{\{|X|>n\}} \leq Z \cdot I_{\{|X|>n\}} \leq Z \cdot I_{\{Z>n\}},$$

where the last inequality follows from the fact that $|X| \leq Z$. Thus,

$$\mathbb{E}\left[|X| \cdot I_{\{|X|>n\}}\right] \leq \mathbb{E}\left[Z \cdot I_{\{Z>n\}}\right]$$

for any $n \in N_+$ and $X \in \mathcal{K}$, implying that

$$\sup_{X \in \mathcal{K}} \mathbb{E}\left[|X| \cdot I_{\{|X|>n\}}\right] \leq \mathbb{E}\left[Z \cdot I_{\{Z>n\}}\right]$$

for any $n \in N_+$. Since Z is integrable, by the characterization of integrability the right hand side goes to 0 as $n \to \infty$; it follows that $\mathcal{K}$ is uniformly integrable.

iii) Suppose that there exists a non-negative function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\lim_{x \to \infty} \frac{f(x)}{x} = +\infty \quad \text{and} \quad \sup_{X \in \mathcal{K}} \mathbb{E}[f \circ |X|] < +\infty.$$

Then, for any $M > 0$ there exists a $b \in \mathbb{R}$ such that

$$\frac{f(x)}{x} > M \quad \text{for any } x > b.$$

By implication, for any $n \in N_+$ such that $n > b$ and $X \in \mathcal{K}$, we have

$$|X| \cdot I_{\{|X|>n\}} \leq \frac{1}{M}\,(f \circ |X|)$$

since $\frac{1}{M} f(|X(\omega)|) > |X(\omega)|$ for any $\omega \in \Omega$ such that $|X(\omega)| > n > b$, and f is

non-negative valued. Taking expectations on both sides yields

$$\mathbb{E}\left[|X| \cdot I_{\{|X|>n\}}\right] \leq \frac{1}{M}\mathbb{E}\left[f \circ |X|\right],$$

and because holds for any $X \in \mathcal{K}$,

$$\sup_{X\in\mathcal{K}} \mathbb{E}\left[|X| \cdot I_{\{|X|>n\}}\right] \leq \frac{1}{M}\left(\sup_{X\in\mathcal{K}} \mathbb{E}\left[f \circ |X|\right]\right).$$

Finally, this holds for any $n > b$, so

$$\limsup_{n\to\infty} \sup_{X\in\mathcal{K}} \mathbb{E}\left[|X| \cdot I_{\{|X|>n\}}\right] \leq \frac{1}{M}\left(\sup_{X\in\mathcal{K}} \mathbb{E}\left[f \circ |X|\right]\right).$$

This in turn holds for any $M > 0$, so that the finiteness of $\sup_{X\in\mathcal{K}} \mathbb{E}\left[f \circ |X|\right]$ implies that

$$\lim_{n\to\infty} \sup_{X\in\mathcal{K}} \mathbb{E}\left[|X| \cdot I_{\{|X|>n\}}\right] = 0.$$

By definition, $\mathcal{K}$ is uniformly integrable.

iv) Suppose that $\mathcal{K}$ is L^p-bounded. Defining $f : \mathbb{R}_+ \to \mathbb{R}_+$ as

$$f(x) = \begin{cases} x^p & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

for any $x \in \mathbb{R}$, f is a non-negative and increasing function such that

$$\lim_{x\to\infty} \frac{f(x)}{x} = \lim_{x\to\infty} x^{p-1} = +\infty,$$

since $p > 1$. By assumption,

$$\sup_{X\in\mathcal{K}} \mathbb{E}\left[f \circ |X|\right] < +\infty,$$

so that, by the preceding result, $\mathcal{K}$ is uniformly integrable.

Q.E.D.

We can offer the following characterizations of uniform integrability; the third criterion, attributed to De la Valleé Poussin, proves particularly useful in the theory of martingales.

Theorem 1.29 (Characterizations of Uniform Integrability)
Let $\mathcal{K}$ be a collection of complex random variables and real random vectors. Then, the following are equivalent:

i) $\mathcal{K}$ is uniformly integrable.

ii) **(Epsilon-Delta Characterization)** $\mathcal{K}$ is L^1-bounded and, for any $\epsilon > 0$ there exists a $\delta > 0$ such that, for any $H \in \mathcal{H}$ satisfying $\mathbb{P}(H) < \delta$, we have

$$\sup_{X \in \mathcal{K}} \mathbb{E}[|X| \cdot I_H] < \epsilon.$$

iii) **(De la Valleé Poussin Characterization)** There exists an increasing, non-negative and convex function $f : \mathbb{R} \to \mathbb{R}_+$ such that

$$\lim_{x \to \infty} \frac{f(x)}{x} = +\infty \quad \text{and} \quad \sup_{X \in \mathcal{K}} \mathbb{E}[f \circ |X|] < +\infty.$$

Proof) We prove the equivalencies one by one.

Epsilon-Delta Characterization

Suppose $\mathcal{K}$ is uniformly integrable. The preceding theorem already shows us that $\mathcal{K}$ is L^1-bounded, so it remains to prove the $\epsilon - \delta$ part of the claim.

Choose any $\epsilon > 0$; by the definition of uniform integrability, there exists an $N \in N_+$ such that

$$\sup_{X \in \mathcal{K}} \mathbb{E}\left[|X| \cdot I_{\{|X| > n\}}\right] < \frac{\epsilon}{2}$$

for any $n \geq N$. Defining $\delta = \frac{\epsilon}{2N}$, choose any $H \in \mathcal{H}$ such that $\mathbb{P}(H) < \delta$. Then, for any $X \in \mathcal{K}$,

$$\begin{aligned}
\mathbb{E}[|X| \cdot I_H] &= \mathbb{E}\left[|X| \cdot I_{\{|X| > N\} \cap H}\right] + \mathbb{E}\left[|X| \cdot I_{\{|X| \leq N\} \cap H}\right] \\
&\leq \mathbb{E}\left[|X| \cdot I_{\{|X| > N\}}\right] + N \cdot \mathbb{P}(\{|X| \leq N\} \cap H) \\
&\leq \frac{\epsilon}{2} + N \cdot \mathbb{P}(H) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

This holds for any $X \in \mathcal{K}$, so

$$\sup_{X \in \mathcal{K}} \mathbb{E}[|X| \cdot I_H] \leq \frac{\epsilon}{2} + N \cdot \mathbb{P}(H) < \epsilon.$$

This in turn holds for any $\epsilon > 0$, so the proof of necessity is complete.

As for sufficiency, suppose that $\mathcal{K}$ is L^1-bounded and that, for any $\epsilon > 0$, there exists a $\delta > 0$ such that $\mathbb{P}(H) < \delta$ implies

$$\sup_{X \in \mathcal{K}} \mathbb{E}\left[|X| \cdot I_H\right] < \epsilon.$$

Choose any $\epsilon > 0$ and $X \in \mathcal{K}$. Since $\mathcal{K}$ is L^1-bounded, it is bounded in probability, so that there exists an $N \in N_+$ such that

$$\sup_{X \in \mathcal{K}} \mathbb{P}\left(|X| > n\right) < \delta$$

for any $n \geq N$. For any $n \geq N$ and $X \in \mathcal{K}$, since $\mathbb{P}\left(|X| > n\right) > \delta$, letting $H = \{|X| > n\}$,

$$\mathbb{E}\left[|X| \cdot I_{|X|>n}\right] \leq \sup_{Y \in \mathcal{K}} \mathbb{E}\left[|Y| \cdot I_H\right] < \epsilon.$$

This holds for any $X \in \mathcal{K}$, so

$$\sup_{X \in \mathcal{K}} \mathbb{E}\left[|X| \cdot I_{|X|>n}\right] \leq \epsilon$$

for any $n \geq N$. This in turn holds for any $\epsilon > 0$, so by definition

$$\lim_{n \to \infty} \sup_{X \in \mathcal{K}} \mathbb{E}\left[|X| \cdot I_{|X|>n}\right] = 0$$

and $\mathcal{K}$ is uniformly integrable.

De La Valleé Poussin Characterization

We proved sufficiency in the previous theorem. To show necessity, we assume that $\mathcal{K}$ is uniformly integrable and construct a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ with the desired properties.

By uniform integrability, we can choose a sequence $\{x_m\}_{m \in N_+}$ of strictly increasing natural numbers such that

$$\sup_{X \in \mathcal{K}} \mathbb{E}\left[|X| \cdot I_{\{|X|>x_m\}}\right] < 2^{-m}$$

for any $m \in N_+$.

Now define the function $f : \mathbb{R} \to \mathbb{R}_+$ as

$$f(x) = \sum_{n=1}^{\infty} (x - x_n)_+$$

for any $x \in \mathbb{R}$, where $(x - x_n)_+$ is defined as

$$(x - x_n)_+ = \max(x - x_n, 0)$$

for any $n \in N_+$. For any $x \in \mathbb{R}_+$, since there exists an $N \in N_+$ such that $x_n > x$ for any $n \geq N$, $f(x)$ is non-negative valued.

f is also clearly increasing, and we can show that it is convex. For any $x, y \in \mathbb{R}$ and $t \in [0,1]$,

$$\begin{aligned}((tx + (1-t)y) - x_n)_+ &= \max(t(x - x_n) + (1-t)(y - x_n), 0) \\ &\leq t \cdot \max(x - x_n, 0) + (1-t)\max(y - x_n, 0) \\ &= t \cdot (x - x_n)_+ + (1-t) \cdot (y - x_n)_+\end{aligned}$$

for any $n \in N_+$ by the convexity of the mapping $z \mapsto \max(z, 0)$ on $\mathbb{R}$, so that

$$\begin{aligned}f(tx + (1-t)y) &= \sum_{n=1}^{\infty} ((tx + (1-t)y) - x_n)_+ \\ &\leq t \cdot \sum_{n=1}^{\infty} (x - x_n)_+ + (1-t) \cdot \sum_{n=1}^{\infty} (y - x_n)_+ = tf(x) + (1-t)f(y).\end{aligned}$$

It remains to show that f satisfies the conditions laid out in the theorem. Note that

$$\frac{f(x)}{x} = \sum_{n=1}^{\infty} \left(1 - \frac{x_n}{x}\right)_+$$

for any $x \in \mathbb{R}_+$. For any $m \in N_+$, define the function $\phi_n : N_+ \to \mathbb{R}_+$ as

$$\phi_n(m) = \left(1 - \frac{x_m}{n}\right)_+ = \max\left(1 - \frac{x_m}{n}, 0\right)$$

for any $m \in N_+$. For any $m, n \in N_+$,

$$\phi_n(m) = \max\left(1 - \frac{x_m}{n}, 0\right) \leq \max\left(1 - \frac{x_m}{n+1}, 0\right) = \phi_{n+1}(m),$$

so that $\{\phi_n\}_{n \in N_+}$ is an increasing sequence of functions whose limit is 1. Letting c be the counting measure on N_+, by the MCT we now have

$$\begin{aligned}\lim_{n \to \infty} \frac{f(n)}{n} &= \lim_{n \to \infty} \sum_{m=1}^{\infty} \left(1 - \frac{x_m}{n}\right)_+ = \lim_{n \to \infty} \int_{N_+} \phi_n dc \\ &= \int_{N_+} \left(\lim_{n \to \infty} \phi_n\right) dc = \int_{N_+} 1 dc = \sum_{m=1}^{\infty} 1 = +\infty.\end{aligned}$$

It follows that

$$\lim_{x\to\infty}\frac{f(x)}{x}=+\infty.$$

For any $X\in\mathcal{K}$ and $\omega\in\Omega$, suppose that $|X(\omega)|\leq x_n$ for any $n\geq N$, where $N\in N_+$. Then,

$$\begin{aligned}(f\circ|X|)(\omega)&=\sum_{n=1}^{\infty}(|X(\omega)|-x_n)_+\\&=\sum_{n=1}^{N}(|X(\omega)|-x_n)\\&\leq\sum_{n=1}^{N}|X(\omega)|=\sum_{n=1}^{\infty}|X(\omega)|\cdot I_{\{|X|>x_n\}}(\omega).\end{aligned}$$

This holds for any $\omega\in\Omega$, so the inequality

$$f\circ|X|\leq\sum_{n=1}^{\infty}|X|\cdot I_{\{|X|>x_n\}}$$

holds. Taking expectations on both sides yields

$$\begin{aligned}\mathbb{E}[f\circ|X|]&=\mathbb{E}\left[\sum_{n=1}^{\infty}|X|\cdot I_{\{|X|>x_n\}}\right]\\&=\sum_{n=1}^{\infty}\mathbb{E}\left[|X|\cdot I_{\{|X|>x_n\}}\right],\end{aligned}$$

where the second inequality follows from Fubini's theorem for non-negative functions. By design,

$$\mathbb{E}\left[|X|\cdot I_{\{|X|>x_n\}}\right]\leq 2^{-n}$$

for any $n\in N_+$, so we can conclude that

$$\mathbb{E}[f\circ|X|]\leq\sum_{n=1}^{\infty}2^{-n}=1<+\infty.$$

This holds for any $X\in\mathcal{K}$, so

$$\sup_{X\in\mathcal{K}}\mathbb{E}[f\circ|X|]\leq 1<+\infty.$$

Q.E.D.

Chapter 2

Conditioning and Information

Here we delve deeper into the idea of σ-algebras as information and expectations formed conditioned on such information. Our investigation culminates in Ionescu-Tulcea's theorem, a discrete version of Kolmogorov's extension theorem, which reveals that there exist probability spaces supporting very general kinds of stochastic processes.

2.1 Conditional Expectations

2.1.1 Conditional Expectations of Non-negative Random Variables

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the underlying probability space, and $\mathcal{F}$ a sub σ-algebra of $\mathcal{H}$. Let X be a non-negative random variable. A version of the conditional expectation of X given $\mathcal{F}$ is a numerical random variable $\overline{X}$ such that

- **Measurability**
  $\overline{X}$ is $\mathcal{F}$-measurable
- **Defining Property**
  For any $H \in \mathcal{F}$,

$$\mathbb{E}[X \cdot I_H] = \mathbb{E}\left[\overline{X} \cdot I_H\right].$$

We often denote the collection of all versions of the conditional expectation of X given $\mathcal{F}$ by $\mathbb{E}[X \mid \mathcal{F}]$.
The following result shows that $\mathbb{E}[X \mid \mathcal{F}]$ is actually an equivalence class of almost surely equal functions:

Lemma 2.1 Let $\mathcal{F}$ be a sub σ-algebra of $\mathcal{H}$, and X, Y $\mathcal{F}$-measurable numerical random variables. Then, $X \leq Y$ almost surely if

$$\mathbb{E}[X \cdot I_H] \leq \mathbb{E}[Y \cdot I_H]$$

for any $H \in \mathcal{F}$, granted that the expectations exist.

Proof) Suppose that

$$\mathbb{E}[X \cdot I_H] \leq \mathbb{E}[Y \cdot I_H]$$

for any $H \in \mathcal{F}$. For any $q, r \in \mathbb{Q}$ such that $q < r$, define

$$H_{qr} = \{Y < q\} \cap \{X > r\},$$

where $H_{qr} \in \mathcal{F}$ because Y, X are $\mathcal{F}$-measurable. Note that

$$H = \{Y < X\} = \bigcup_{q,r \in \mathbb{Q},\ q<r} (\{Y < q\} \cap \{X > r\}) = \bigcup_{q,r \in \mathbb{Q},\ q<r} H_{qr},$$

where the union on the right is over a countable index. Suppose that $\mathbb{P}(H) > 0$. Then, by countable subadditivity,

$$0 < \mathbb{P}(H) \leq \sum_{q,r \in \mathbb{Q},\ q<r} \mathbb{P}(H_{qr}),$$

indicating that there exists some $q, r \in \mathbb{Q}$ such that $q < r$ and $\mathbb{P}(H_{qr}) > 0$. We now have

$$\begin{aligned} r \cdot \mathbb{P}(H_{qr}) = \mathbb{E}\left[r \cdot I_{H_{qr}}\right] &\leq \mathbb{E}\left[X \cdot I_{H_{qr}}\right] \\ &\leq \mathbb{E}\left[Y \cdot I_{H_{qr}}\right] \leq \mathbb{E}\left[q \cdot I_{H_{qr}}\right] = q \cdot \mathbb{P}(H_{qr}). \end{aligned}$$

Dividing both sides by $\mathbb{P}(H_{qr}) > 0$ yields $r \leq q$, a contradiction. Therefore, we must have $\mathbb{P}(H) = 0$, and by implication $X \leq Y$ almost surely.

Q.E.D.

Corollary to Lemma 2.1 Let $\mathcal{F}$ be a sub σ-algebra of $\mathcal{H}$, and X, Y $\mathcal{F}$-measurable numerical random variables. Then, $X = Y$ almost surely if

$$\mathbb{E}[X \cdot I_H] = \mathbb{E}[Y \cdot I_H]$$

for any $H \in \mathcal{F}$, granted that the expectations exist.

Proof) $X = Y$ almost surely if and only if $X \leq Y$ and $X \geq Y$ almost surely. The above lemma then implies that $X \leq Y$ and $X \geq Y$ almost surely if and only if $\mathbb{E}[X \cdot I_H] \leq \mathbb{E}[Y \cdot I_H]$ and $\mathbb{E}[X \cdot I_H] \geq \mathbb{E}[Y \cdot I_H]$ for any $H \in \mathcal{F}$. Putting this all together, $X = Y$ almost surely if and only if

$$\mathbb{E}[X \cdot I_H] = \mathbb{E}[Y \cdot I_H]$$

for any $H \in \mathcal{F}$.

Q.E.D.

For any non-negative random variable X, let $\overline{X}$ and $\tilde{X}$ belong to $\mathbb{E}[X \mid \mathcal{F}]$. Then, for any $H \in \mathcal{F}$,

$$\mathbb{E}\left[\overline{X} \cdot I_H\right] = \mathbb{E}[X \cdot I_H] = \mathbb{E}\left[\tilde{X} \cdot I_H\right],$$

and because $\overline{X}$ and $\tilde{X}$ are both $\mathcal{F}$-measurable, we have $\overline{X} = \tilde{X}$ almost surely by the above lemma. Therefore, versions of conditional expectations are unique up to almost sure equivalence, and we can let $\mathbb{E}[X \mid \mathcal{F}]$ also denote a representative of the equivalence class of versions of conditional expectations of X given $\mathcal{F}$. For this reason, equalities and inequalities usually only hold with probability 1 when dealing with conditional expectations. In what comes below, every equality and inequality is required to hold only almost surely, unless it is important to specify otherwise.

One of the first questions we can ask is whether $\mathbb{E}[X \mid \mathcal{F}]$ is non-empty. The fact that $\mathbb{E}[X \mid \mathcal{F}]$ is non-empty for any non-negative random variable X can be shown using the Radon-Nikodym theorem:

Theorem 2.2 (Existence of Conditional Expectations)
Let $\mathcal{F}$ be a sub σ-algebra of $\mathcal{H}$, and X a non-negative random variable. Then, $\mathbb{E}[X \mid \mathcal{F}] \neq \emptyset$.

Proof) Suppose initially that X is integrable. Since X is non-negative, we can define the measure μ on $(\Omega, \mathcal{F})$ as

$$\mu(H) = \int_H X d\mathbb{P}$$

for any $H \in \mathcal{F}$. Letting $\mathbb{P}'$ be the restriction of $\mathbb{P}$ to $(\Omega, \mathcal{F})$, μ is absolutely continuous with respect to $\mathbb{P}'$, and both are finite measures ($\mu(\Omega) = \mathbb{E}[X] < +\infty$). Therefore, by the Radon-Nikodym theorem there exists a $\mathcal{F}$-measurable function $\overline{X}$ such that

$$\mu(H) = \int_H \overline{X} d\mathbb{P}'$$

for any $H \in \mathcal{F}$. It follows that

$$\mathbb{E}[X \cdot I_H] = \mu(H) = \int_H \overline{X} d\mathbb{P}'$$

for any $H \in \mathcal{F}$.

It remains to verify that the integral on the right is the $\mathbb{P}$-expectation of $\overline{X} \cdot I_H$. To this end, we approximate $\overline{X}$ with an increasing sequence of non-negative and measurable simple functions, as usual. Fix $H \in \mathcal{F}$, and let $\{f_n\}_{n \in N_+}$ be a sequence of non-negative $\mathcal{F}$-measurable simple functions increasing to $\overline{X}$. For any $n \in N_+$, suppose the canonical

form of f_n is given as

$$f_n = \sum_{i=1}^{m} a_i \cdot I_{A_i},$$

where $A_1, \cdots, A_m \in \mathcal{F}$. Then,

$$\begin{aligned}\int_H f_n d\mathbb{P}' &= \int_\Omega \left[\sum_{i=1}^{m} a_i \cdot I_{A_i \cap H}\right] d\mathbb{P}' \\ &= \sum_{i=1}^{m} a_i \cdot \mathbb{P}'(A_i \cap H) = \sum_{i=1}^{m} a_i \cdot \mathbb{P}(A_i \cap H) = \int_H f_n d\mathbb{P},\end{aligned}$$

since each $A_i \cap H \in \mathcal{F}$ and $\mathbb{P} = \mathbb{P}'$ on $\mathcal{F}$. This holds for any $n \in N_+$, so by the MCT,

$$\int_H \overline{X} d\mathbb{P}' = \lim_{n\to\infty} \int_H f_n \mathbb{P}' = \lim_{n\to\infty} \int_H f_n d\mathbb{P} = \int_H \overline{X} d\mathbb{P}.$$

The last term on the right is just $\mathbb{E}\left[\overline{X} \cdot I_H\right]$, so we can see that

$$\mathbb{E}[X \cdot I_H] = \mathbb{E}\left[\overline{X} \cdot I_H\right]$$

for any $H \in \mathcal{F}$.

Now let $\mathbb{E}[X] = +\infty$. In this case, define for any $n \in N_+$

$$X_n = \max(X, n).$$

$\{X_n\}_{n\in N_+}$ is a sequence of bounded and thus integrable non-negative random variables that increases to X. By the preceding result, for any $n \in N_+$ there exists a non-negative $\mathcal{F}$-measurable function $\overline{X}_n$ such that

$$\mathbb{E}[X_n \cdot I_H] = \mathbb{E}\left[\overline{X}_n \cdot I_H\right]$$

for any $H \in \mathcal{F}$. Note that, for any $n \in N_+$,

$$\mathbb{E}\left[\overline{X}_n \cdot I_H\right] = \mathbb{E}[X_n \cdot I_H] \leq \mathbb{E}[X_{n+1} \cdot I_H] = \mathbb{E}\left[\overline{X}_{n+1} \cdot I_H\right]$$

for any $H \in \mathcal{F}$. Since $\overline{X}_n$ and $\overline{X}_{n+1}$ are both $\mathcal{F}$-measurable, this implies by lemma 2.1 that $\overline{X}_n \leq \overline{X}_{n+1}$. In other words, $\{\overline{X}_n\}_{n\in N_+}$ is a sequence that increases almost surely to some limit $\overline{X}$. By the almost everywhere version of the MCT, for any $H \in \mathcal{F}$ we now have

$$\mathbb{E}[X \cdot I_H] = \lim_{n\to\infty} \mathbb{E}[X_n \cdot I_H] = \lim_{n\to\infty} \mathbb{E}\left[\overline{X}_n \cdot I_H\right] = \mathbb{E}\left[\overline{X} \cdot I_H\right].$$

Being the limit of $\mathcal{F}$-measurable functions, $\overline{X}$ is itself $\mathcal{F}$-measurable. Thus, by definition, $\overline{X} = \mathbb{E}[X \mid \mathcal{F}]$.

Q.E.D.

What the above theorem actually does is furnish us with a version of the conditional expectation of X given $\mathcal{F}$ that is non-negative everywhere on Ω. Henceforth, we take $\mathbb{E}[X \mid \mathcal{F}]$ to be the representative of the equivalence class $\mathbb{E}[X \mid \mathcal{F}]$ that is non-negative everywhere. It follows that every version of $\mathbb{E}[X \mid \mathcal{F}]$ is almost surely non-negative.

Conditional expectations retain many properties of expectations, including linearity; these are presented below:

Theorem 2.3 (Expectation Properties of Conditional Expectations)
Let $\mathcal{F}$ be a sub σ-algebra of $\mathcal{H}$. Let X, Y be non-negative random variables. Then, the following hold true:

i) **(Projection Property)** A non-negative $\mathcal{F}$-measurable random variable $\overline{X}$ is a version of the conditional expectation of X given $\mathcal{F}$ if and only if, for any $V \in \mathcal{F}_+$,

$$\mathbb{E}[X \cdot V] = \mathbb{E}\left[\overline{X} \cdot V\right].$$

ii) **(Monotonicity)** If $X \leq Y$, then $\mathbb{E}[X \mid \mathcal{F}] \leq \mathbb{E}[Y \mid \mathcal{F}]$.

iii) **(Linearity)** Let $a \in \mathbb{R}_+$. Then,

$$a \cdot \mathbb{E}[X \mid \mathcal{F}] + \mathbb{E}[Y \mid \mathcal{F}] = \mathbb{E}[aX + Y \mid \mathcal{F}].$$

iv) **(Monotone Convergence Theorem)** Let $\{X_n\}_{n \in N_+}$ be a sequence of (almost surely) increasing non-negative random variables. Letting X be the pointwise limit of $\{X_n\}_{n \in N_+}$,

$$\mathbb{E}[X \mid \mathcal{F}] = \lim_{n \to \infty} \mathbb{E}[X_n \mid \mathcal{F}].$$

v) **(Fatou's Lemma)** Let $\{X_n\}_{n \in N_+}$ be a sequence of (almost surely) non-negative random variables and $X = \liminf_{n \to \infty} X_n$. Then,

$$\mathbb{E}[X \mid \mathcal{F}] \leq \liminf_{n \to \infty} \mathbb{E}[X_n \mid \mathcal{F}].$$

Proof) i) Sufficiency is clear, since I_H is a $\mathcal{F}$-measurable non-negative random variable for any $H \in \mathcal{F}$.

Suppose that $\overline{X} = \mathbb{E}[X \mid \mathcal{F}]$. Then, for any simple non-negative and $\mathcal{F}$-measurable random variable V with canonical form

$$V = \sum_{i=1}^{n} a_i \cdot I_{A_i}$$

for some $A_1, \cdots, A_n \in \mathcal{F}$, we have

$$\begin{aligned}\mathbb{E}[X \cdot V] &= \sum_{i=1}^{n} a_i \cdot \mathbb{E}[X \cdot I_{A_i}] \\ &= \sum_{i=1}^{n} a_i \cdot \mathbb{E}\left[\overline{X} \cdot I_{A_i}\right] = \mathbb{E}\left[\overline{X} \cdot V\right]\end{aligned}$$

by the linearity of integration of non-negative functions.

Now let V be a $\mathcal{F}$-measurable non-negative random variable in general. Then, there exists a sequence $\{V_n\}_{n \in N_+}$ of $\mathcal{F}$-measurable simple functions that increases to V; this indicates that $\{X \cdot V_n\}_{n \in N_+}$ and $\{\overline{X} \cdot V_n\}_{n \in N_+}$ are sequences of non-negative measurable functions that increase to $X \cdot V$ and $\overline{X} \cdot V$, respectively. By the MCT, we now have

$$\mathbb{E}[X \cdot V] = \lim_{n \to \infty} \mathbb{E}[X \cdot V_n] = \lim_{n \to \infty} \mathbb{E}\left[\overline{X} \cdot V_n\right] = \mathbb{E}\left[\overline{X} \cdot V\right].$$

ii) Suppose $X \leq Y$ (as usual, the inequality is almost sure). Then, for any $H \in \mathcal{F}$,

$$\mathbb{E}[X \cdot I_H] \leq \mathbb{E}[Y \cdot I_H]$$

by the monotonicity of integration, which in turn implies by the defining property that

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{F}] \cdot I_H] \leq \mathbb{E}[\mathbb{E}[Y \mid \mathcal{F}] \cdot I_H]$$

for any $H \in \mathcal{F}$. Since $\mathbb{E}[X \mid \mathcal{F}]$ and $\mathbb{E}[Y \mid \mathcal{F}]$ are both $\mathcal{F}$-measurable, by lemma 2.1 we have $\mathbb{E}[X \mid \mathcal{F}] \leq \mathbb{E}[Y \mid \mathcal{F}]$ (also almost surely).

iii) Choose any $a \in \mathbb{R}_+$. The sum $a\mathbb{E}[X \mid \mathcal{F}] + \mathbb{E}[Y \mid \mathcal{F}]$ is a non-negative $\mathcal{F}$-measurable function, and for any $H \in \mathcal{F}$,

$$\begin{aligned}\mathbb{E}[(aX + Y) \cdot I_H] &= a \cdot \mathbb{E}[X \cdot I_H] + \mathbb{E}[Y \cdot I_H] \\ &= a \cdot \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}] \cdot I_H] + \mathbb{E}[\mathbb{E}[Y \mid \mathcal{F}] \cdot I_H] \\ &= \mathbb{E}[(a\mathbb{E}[X \mid \mathcal{F}] + \mathbb{E}[Y \mid \mathcal{F}]) \cdot I_H],\end{aligned}$$

where we used the linearity of the integration of non-negative functions and the defining property of conditional expectations. By definition, $a\mathbb{E}[X \mid \mathcal{F}] + \mathbb{E}[Y \mid \mathcal{F}]$ is a version of the conditional expectation $\mathbb{E}[aX + Y \mid \mathcal{F}]$.

iv) By the monotonicity of conditional expectations, $\{\mathbb{E}[X_n \mid \mathcal{F}]\}_{n \in N_+}$ is an almost surely increasing sequence of non-negative random variables. This means that the (almost sure) pointwise limit of $\{\mathbb{E}[X_n \mid \mathcal{F}]\}_{n \in N_+}$ is well-defined. We must show

that this limit is a version of the conditional expectation of X given $\mathcal{F}$. Denote this limit by $\overline{X}$.

Being the limit of $\mathcal{F}$-measurable functions, $\overline{X}$ is also $\mathcal{F}$-measurable. It remains to show the defining property holds. For any $H \in \mathcal{F}$, $\{X_n \cdot I_H\}_{n \in N_+}$ and $\{\mathbb{E}[X_n \mid \mathcal{F}] \cdot I_H\}_{n \in N_+}$ are sequences that increase (almost surely) to $X \cdot I_H$ and $\overline{X} \cdot I_H$; thus, by the almost everywhere version of the MCT,

$$\mathbb{E}[X \cdot I_H] = \lim_{n \to \infty} \mathbb{E}[X_n \cdot I_H] = \lim_{n \to \infty} \mathbb{E}\left[\overline{X}_n \cdot I_H\right] = \mathbb{E}\left[\overline{X} \cdot I_H\right].$$

By definition,

$$\mathbb{E}[X \mid \mathcal{F}] = \overline{X} = \lim_{n \to \infty} \mathbb{E}[X_n \mid \mathcal{F}].$$

v) Define the sequence $\{Y_n\}_{n \in N_+}$ as

$$Y_n = \inf_{k \geq n} X_n$$

for any $n \in N_+$. Then, $\{Y_n\}_{n \in N_+}$ is an increasing sequence of non-negative random variables with pointwise limit $X = \liminf_{n \to \infty} X_n$.

For any $n \in N_+$ and $H \in \mathcal{F}$, since $Y_n \leq X_k$ for any $k \geq n$, we have

$$\mathbb{E}[Y_n \cdot I_H] \leq \mathbb{E}[X_k \cdot I_H],$$

and by the defining property of conditional expectations,

$$\mathbb{E}[\mathbb{E}[Y_n \mid \mathcal{F}] \cdot I_H] \leq \mathbb{E}[\mathbb{E}[X_k \mid \mathcal{F}] \cdot I_H].$$

This holds for any $H \in \mathcal{F}$, so by lemma 2.1,

$$\mathbb{E}[Y_n \mid \mathcal{F}] \leq \mathbb{E}[X_k \mid \mathcal{F}]$$

for any $k \geq n$, and as such

$$\mathbb{E}[Y_n \mid \mathcal{F}] \leq \inf_{k \geq n} \mathbb{E}[X_k \mid \mathcal{F}].$$

By the conditional version of the MCT,

$$\lim_{n \to \infty} \mathbb{E}[Y_n \mid \mathcal{F}] = \mathbb{E}[X \mid \mathcal{F}].$$

Therefore,

$$\mathbb{E}[X \mid \mathcal{F}] = \lim_{n \to \infty} \mathbb{E}[Y_n \mid \mathcal{F}] \leq \lim_{n \to \infty} \inf_{k \geq n} \mathbb{E}[X_k \mid \mathcal{F}] = \liminf_{n \to \infty} \mathbb{E}[X_n \mid \mathcal{F}].$$

Q.E.D.

Conditional expectations also possess the following unique properties:

Theorem 2.4 (Unique Properties of Conditional Expectations)
Let $\mathcal{F}, \mathcal{G}$ be sub σ-algebras of $\mathcal{H}$, and X, Y non-negative random variables. Then, the following hold true:

i) **(Conditional Determinism)** If X is $\mathcal{F}$-measurable, then

$$\mathbb{E}[XY \mid \mathcal{F}] = X \cdot \mathbb{E}[Y \mid \mathcal{F}].$$

ii) **(Tower Property)** If $\mathcal{F} \subset \mathcal{G}$, then

$$\mathbb{E}[X \mid \mathcal{F}] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}] \mid \mathcal{G}] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{F}].$$

Proof) i) If $X \in \mathcal{F}$, then for any $H \in \mathcal{F}$, since $X \cdot I_H$ is a $\mathcal{F}$-measurable non-negative random variable, by the projection property

$$\mathbb{E}[XY \cdot I_H] = \mathbb{E}[\mathbb{E}[Y \mid \mathcal{F}] \cdot XI_H].$$

This holds for any $H \in \mathcal{F}$, and $\mathbb{E}[Y \mid \mathcal{F}] \cdot X$ is a $\mathcal{F}$-measurable non-negative random variable, so by definition $\mathbb{E}[Y \mid \mathcal{F}] \cdot X$ is a version of $\mathcal{XY} \mid \mathcal{F}$, or in other words,

$$X \cdot \mathbb{E}[Y \mid \mathcal{F}] = \mathbb{E}[XY \mid \mathcal{F}].$$

ii) The first equality follows easily from conditional determinism, since $\mathbb{E}[X \mid \mathcal{F}]$ is $\mathcal{F}$-measurable and thus $\mathcal{G}$-measurable.

To show that the second equality holds, note that, for any $H \in \mathcal{F}$, $H \in \mathcal{G}$ as well, so that

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \cdot I_H] = \mathbb{E}[X \cdot I_H].$$

However, by definition, we also have the equality

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \cdot I_H] = \mathbb{E}[\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{F}] \cdot I_H].$$

Putting the two together,

$$\mathbb{E}[\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{F}] \cdot I_H] = \mathbb{E}[X \cdot I_H]$$

for any $H \in \mathcal{F}$, and since $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{F}]$ is $\mathcal{F}$-measurable, by definition

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{F}] = \mathbb{E}[X \mid \mathcal{F}].$$

Q.E.D.

The tower property, in particular, implies the infamous law of iterated expectations:

$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{F}]] = \mathbb{E}[X],$$

since unconditional expectations can be thought of as conditional expectations with respect to the trivial σ-algebra. In other words, if X is integrable, then so is $\mathbb{E}[X \mid \mathcal{F}]$.

2.1.2 Conditional Expectations of Real Random Variables

Now we extend the above framework to arbitrary real random variables. Let X be a real-valued random variable, with positive and negative parts X^+ and X^-. Since X^+ and X^- are non-negative random variables, by the results we showed above the conditional expectations $\mathbb{E}\left[X^+ \mid \mathcal{F}\right]$ and $\mathbb{E}[X^- \mid \mathcal{F}]$ exist. We define the conditional expectation of X given $\mathcal{F}$ as the collection

$$\mathbb{E}[X \mid \mathcal{F}] = \{\overline{X}^+ - \overline{X}^- \mid \overline{X}^{\pm} \in \mathbb{E}\left[X^{\pm} \mid \mathcal{F}\right], \quad \text{and the difference is almost surely well-defined}\}.$$

We also write this as

$$\mathbb{E}[X \mid \mathcal{F}] = \mathbb{E}\left[X^+ \mid \mathcal{F}\right] - \mathbb{E}\left[X^- \mid \mathcal{F}\right].$$

As in the case of non-negative random variables, we can see that $\mathbb{E}[X \mid \mathcal{F}]$ is an equivalence class of almost surely equal random variables. To see this, let $\overline{X}, \overline{Y} \in \mathbb{E}[X \mid \mathcal{F}]$; then, there exist $\overline{X}^{\pm}, \overline{Y}^{\pm} \in \mathbb{E}[X^{\pm} \mid \mathcal{F}]$ such that

$$\overline{X} = \overline{X}^+ - \overline{X}^- \quad \text{and} \quad \overline{Y} = \overline{Y}^+ - \overline{Y}^-.$$

Since $\overline{X}^+ = \overline{Y}^+$ and $\overline{X}^- = \overline{Y}^-$ almost surely, it follows that $\overline{X} = \overline{Y}$ almost surely as well. Thus, as in the non-negative case, we let $\mathbb{E}[X \mid \mathcal{F}]$ also denote a representative of the equivalence class, and assume that all equalities and inequalities hold almost surely, unless specified otherwise.

For $\mathbb{E}[X \mid \mathcal{F}]$ to be non-empty, it suffices for $\mathbb{E}[X^{\pm} \mid \mathcal{F}]$ to be finite almost surely (since $\mathbb{E}[X^{\pm} \mid \mathcal{F}]$ is an equivalence class, this means that every conditional expectation is also finite almost surely). In this case, we say that X is conditionally integrable given $\mathcal{F}$; note that every random variable in $\mathbb{E}[X \mid \mathcal{F}]$ is almost surely real-valued in this case. Note that this does not imply that X is integrable.

The following show that conditional expectations of real random variables retains much of the properties associated with unconditional expectations, and that they also possess the same unique properties as conditional expectations of non-negative random variables.

Theorem 2.5 (Properties of Conditional Expectations of Real Variables) Let $\mathcal{F},\mathcal{G}$ be sub σ-algebras of $\mathcal{H}$. Let X,Y be real valued random variables that are conditionally integrable given $\mathcal{F}$. Then, the following hold true:

i) **(Defining Property)** Suppose X is integrable. Then, it is conditionally integrable given $\mathcal{F}$. Additionally, $\overline{X} = \mathbb{E}[X \mid \mathcal{F}]$ is a $\mathcal{F}$-measurable and integrable real random variable such that

$$\mathbb{E}[X \cdot V] = \mathbb{E}\left[\overline{X} \cdot V\right]$$

for any integrable V.

ii) **(Monotonicity)** If $X \leq Y$, then $\mathbb{E}[X \mid \mathcal{F}] \leq \mathbb{E}[Y \mid \mathcal{F}]$.

iii) **(Linearity)** For any $a \in \mathbb{R}$,

$$a \cdot \mathbb{E}[X \mid \mathcal{F}] + \mathbb{E}[Y \mid \mathcal{F}] = \mathbb{E}[aX + Y \mid \mathcal{F}].$$

iv) **(Dominated Convergence Theorem)** Let $\{X_n\}_{n \in N_+}$ be a sequence of real random variables that converges pointwise to some real random variable X. Suppose there exists a real random variable Y conditionally integrable given $\mathcal{F}$ such that $|X_n| \leq Y$ for any $n \in N_+$. Then, X is conditionally integrable given $\mathcal{F}$ and

$$\lim_{n \to \infty} \mathbb{E}[X_n \mid \mathcal{F}] = \mathbb{E}[X \mid \mathcal{F}].$$

v) **(Conditional Determinism)** Let X be a $\mathcal{F}$-measurable random variable. Then, XY is conditionally integrable given $\mathcal{F}$ and

$$\mathbb{E}[XY \mid \mathcal{F}] = X \cdot \mathbb{E}[Y \mid \mathcal{F}].$$

vi) **(Tower Property)** Let $\mathcal{F} \subset \mathcal{G}$, and assume that X is also conditionally integrable given $\mathcal{G}$. Then,

$$\mathbb{E}[X \mid \mathcal{F}] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}] \mid \mathcal{G}] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{F}].$$

Proof) i) Suppose X is integrable. Then, by the law of iterated expectations,

$$\mathbb{E}\left[\mathbb{E}\left[X^{\pm} \mid \mathcal{F}\right]\right] = \mathbb{E}\left[X^{\pm}\right] < +\infty,$$

so that, by the finiteness property, $\mathbb{E}[X^{\pm} \mid \mathcal{F}] < +\infty$ almost surely. It follows that X is conditionally integrable given $\mathcal{F}$.

Let $\overline{X} = \mathbb{E}[X \mid \mathcal{F}]$. By definition, there exist $\overline{X}^{+}, \overline{X}^{-}$ such that $\overline{X}^{\pm} = \mathbb{E}[X^{\pm} \mid \mathcal{F}]$

and

$$\overline{X} = \overline{X}^+ - \overline{X}^-;$$

since $\overline{X}^+$ and $\overline{X}^-$ are both integrable, so is $\overline{X}$, and by the finiteness property we can choose versions of them that are finite everywhere. This ensures that the operations below are all well-defined.

For any real integrable $\mathcal{F}$-measurable random variable V, we can see that

$$X \cdot V = (X^+ - X^-)(V^+ - V^-) = X^+V^+ - X^-V^+ - X^+V^- + X^-V^-.$$

Since V^+, V^- are non-negative $\mathcal{F}$-measurable random variables, by the projection property for non-negative variables we have

$$\begin{aligned}\mathbb{E}\left[X^+V^\pm\right] &= \mathbb{E}\left[\overline{X}^+V^\pm\right] \\ \mathbb{E}\left[X^-V^\pm\right] &= \mathbb{E}\left[\overline{X}^-V^\pm\right].\end{aligned}$$

Since these expectations are all finite (they are products of integrable variables), it follows that

$$\begin{aligned}\mathbb{E}[X \cdot V] &= \mathbb{E}\left[X^+V^+\right] - \mathbb{E}\left[X^-V^+\right] - \mathbb{E}\left[X^+V^-\right] + \mathbb{E}\left[X^-V^-\right] \\ &= \mathbb{E}\left[\overline{X}^+V^+\right] - \mathbb{E}\left[\overline{X}^-V^+\right] - \mathbb{E}\left[\overline{X}^+V^-\right] + \mathbb{E}\left[\overline{X}^-V^-\right] = \mathbb{E}\left[\overline{X} \cdot V\right].\end{aligned}$$

ii) Suppose $\overline{X} = \mathbb{E}[X \mid \mathcal{F}]$ and $\overline{Y} = \mathbb{E}[Y \mid \mathcal{F}]$, and let $X \leq Y$. Then,

$$X^+ + Y^- \leq Y^+ + X^-,$$

and by the monotonicity and linearity of conditional expectations of non-negative random variables,

$$\mathbb{E}\left[X^+ \mid \mathcal{F}\right] + \mathbb{E}\left[Y^- \mid \mathcal{F}\right] \leq \mathbb{E}\left[X^- \mid \mathcal{F}\right] + \mathbb{E}\left[Y^+ \mid \mathcal{F}\right].$$

By conditional integrability, $\mathbb{E}[X^\pm \mid \mathcal{F}] < +\infty$ and $\mathbb{E}[Y^\pm \mid mathcalF] < +\infty$ almost surely; as such, the above inequality tells us that

$$\begin{aligned}\mathbb{E}[X \mid \mathcal{F}] &= \mathbb{E}\left[X^+ \mid \mathcal{F}\right] - \mathbb{E}\left[X^- \mid \mathcal{F}\right] \\ &\leq \mathbb{E}\left[Y^+ \mid \mathcal{F}\right] - \mathbb{E}\left[Y^- \mid \mathcal{F}\right] = \mathbb{E}[Y \mid \mathcal{F}].\end{aligned}$$

iii) Choose any $a \in \mathbb{R}$. Suppose that $a \geq 0$. In this case,

$$(aX)^+ = aX^+ \quad \text{and} \quad (aX)^- = aX^-,$$

so that

$$\mathbb{E}\left[(aX)^{\pm} \mid \mathcal{F}\right] = a \cdot \mathbb{E}\left[X^{\pm} \mid \mathcal{F}\right]$$

by the linearity of conditional expectations of non-negative functions. By conditional integrability, $\mathbb{E}[(aX)^{\pm} \mid \mathcal{F}] < +\infty$ almost surely, so that aX is also conditionally integrable given $\mathcal{F}$, and we have

$$\mathbb{E}[aX \mid \mathcal{F}] = a \cdot \mathbb{E}[X \mid \mathcal{F}].$$

If $a < 0$, we can repeat the process above using the observation that

$$(aX)^+ = -aX^- \quad \text{and} \quad (aX)^- = -aX^+.$$

By monotonicity, since $(aX+Y)^+ \leq (aX)^+ + Y^+$,

$$\mathbb{E}\left[(aX+Y)^+ \mid \mathcal{F}\right] \leq \mathbb{E}\left[(aX)^+ \mid \mathcal{F}\right] + \mathbb{E}\left[Y^+ \mid \mathcal{F}\right] < +\infty.$$

The same holds for $(aX+Y)^-$, so $aX+Y$ is conditionally integrable given $\mathcal{F}$. Now note that

$$aX + Y = (aX)^+ - (aX)^- + Y^+ - Y^- = (aX+Y)^+ - (aX+Y)^-,$$

so that

$$(aX+Y)^+ + (aX)^- + Y^- = (aX)^+ + Y^+ + (aX+Y)^-.$$

By the linearity of conditional expectations of non-negative functions, we have

$$\begin{aligned}\mathbb{E}\left[(aX+Y)^+ \mid \mathcal{F}\right] + \mathbb{E}\left[(aX)^- \mid \mathcal{F}\right] + \mathbb{E}\left[Y^- \mid \mathcal{F}\right] \\ = \mathbb{E}\left[(aX+Y)^- \mid \mathcal{F}\right] + \mathbb{E}\left[(aX)^+ \mid \mathcal{F}\right] + \mathbb{E}\left[Y^+ \mid \mathcal{F}\right],\end{aligned}$$

and because all the conditional expectations involved are almost surely finite, rearranging terms yields

$$\mathbb{E}[aX+Y \mid \mathcal{F}] = \mathbb{E}[aX \mid \mathcal{F}] + \mathbb{E}[Y \mid \mathcal{F}] = a \cdot \mathbb{E}[X \mid \mathcal{F}] + \mathbb{E}[Y \mid \mathcal{F}].$$

iv) We first show that X is conditionally integrable. For any $n \in N_+$, $|X_n| \leq Y$ and

Y is conditionally integrable, so that, by monotonicity,

$$\mathbb{E}[|X_n| \mid \mathcal{F}] \leq \mathbb{E}[Y \mid \mathcal{F}] < +\infty$$

for any $n \in N_+$. By the conditional version of Fatou's lemma, it now follows that

$$\mathbb{E}[|X| \mid \mathcal{F}] \leq \liminf_{n\to\infty} \mathbb{E}[|X_n| \mid \mathcal{F}] \leq \mathbb{E}[Y \mid \mathcal{F}] < +\infty.$$

Since $|X| = X^+ + X^-$, the linearity of the conditional expectations of non-negative functions implies that

$$\mathbb{E}\left[X^{\pm} \mid \mathcal{F}\right] < +\infty,$$

and as such that X is conditionally integrable.

Define the sequences $\{Y_n\}_{n\in N_+}$ and $\{Z_n\}_{n\in N_+}$ as

$$Y_n = Y - X_n \quad \text{and} \quad Z_n = X_n + Y$$

for any $n \in N_+$. Since $|X_n| \leq Y$ for any $n \in N_+$ by assumption, $\{Y_n\}_{n\in N_+}$ and $\{Z_n\}_{n\in N_+}$ are sequences of non-negative random variables, so that, by the conditional version of Fatou's lemma,

$$\mathbb{E}[Y - X \mid \mathcal{F}] \leq \liminf_{n\to\infty} \mathbb{E}[Y_n \mid \mathcal{F}]$$
$$\mathbb{E}[Y + X \mid \mathcal{F}] \leq \liminf_{n\to\infty} \mathbb{E}[Z_n \mid \mathcal{F}].$$

Here, Y is conditionally integrable given $\mathcal{F}$, so that, by linearity,

$$\limsup_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{F}] \leq \mathbb{E}[X \mid \mathcal{F}] \leq \liminf_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{F}].$$

It follows that

$$\lim_{n\to\infty} \mathbb{E}[X_n \mid \mathcal{F}] = \mathbb{E}[X \mid \mathcal{F}].$$

v) Suppose X is $\mathcal{F}$-measurable. Then, so are X^+ and X^-. We first show that XY is conditionally integrable given $\mathcal{F}$.

$$XY = X^+Y^+ - X^+Y^- - X^-Y^+ + X^-Y^-,$$

and $XY < 0$ if only one of X or Y is negative, while $XY \geq 0$ if both are non-negative or negative, we can see that

$$(XY)^+ = X^+Y^+ + X^-Y^- \quad \text{and} \quad (XY)^- = X^+Y^- + X^-Y^+.$$

By conditional determinism applied to the positive and negative parts of X and Y, we have

$$\begin{aligned}\mathbb{E}\left[X^{+}Y^{\pm} \mid \mathcal{F}\right] &= X^{+} \cdot \mathbb{E}\left[Y^{\pm} \mid \mathcal{F}\right] \\ \mathbb{E}\left[X^{-}Y^{\pm} \mid \mathcal{F}\right] &= X^{-} \cdot \mathbb{E}\left[Y^{\pm} \mid \mathcal{F}\right].\end{aligned}$$

Then, the linearity of the conditional expectations of non-negative functions implies that

$$\begin{aligned}\mathbb{E}\left[(XY)^{+} \mid \mathcal{F}\right] &= X^{+} \cdot \mathbb{E}\left[Y^{+} \mid \mathcal{F}\right] + X^{-} \cdot \mathbb{E}\left[Y^{-} \mid \mathcal{F}\right] < +\infty \\ \mathbb{E}\left[(XY)^{-} \mid \mathcal{F}\right] &= X^{+} \cdot \mathbb{E}\left[Y^{-} \mid \mathcal{F}\right] + X^{-} \cdot \mathbb{E}\left[Y^{+} \mid \mathcal{F}\right] < +\infty\end{aligned}$$

almost surely. By definition, XY is conditionally integrable given $\mathcal{F}$. It is now immediately evident that

$$\mathbb{E}\left[XY \mid \mathcal{F}\right] = \mathbb{E}\left[(XY)^{+} \mid \mathcal{F}\right] - \mathbb{E}\left[(XY)^{-} \mid \mathcal{F}\right] = X \cdot \mathbb{E}\left[Y \mid \mathcal{F}\right].$$

vi) By the tower property for non-negative random variables,

$$\mathbb{E}\left[X^{\pm} \mid \mathcal{F}\right] = \mathbb{E}\left[\mathbb{E}\left[X^{\pm} \mid \mathcal{F}\right] \mid \mathcal{G}\right] = \mathbb{E}\left[\mathbb{E}\left[X^{\pm} \mid \mathcal{G}\right] \mid \mathcal{F}\right].$$

Since X is conditionally integrable given $\mathcal{F}$ and $\mathcal{G}$, we can see that

$$\begin{aligned}\mathbb{E}\left[X \mid \mathcal{F}\right] &= \mathbb{E}\left[X^{+} \mid \mathcal{F}\right] - \mathbb{E}\left[X^{-} \mid \mathcal{F}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[X^{+} \mid \mathcal{F}\right] \mid \mathcal{G}\right] - \mathbb{E}\left[\mathbb{E}\left[X^{-} \mid \mathcal{F}\right] \mid \mathcal{G}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[X^{+} \mid \mathcal{F}\right] - \mathbb{E}\left[X^{-} \mid \mathcal{F}\right] \mid \mathcal{G}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{F}\right] \mid \mathcal{G}\right]\end{aligned}$$

by linearity, and the same holds for the second equality as well.

Q.E.D.

2.1.3 Conditional Expectations of Real Random Vectors

The conditional expectations of real random vectors are defined analogously to their unconditional expectations. Let $\mathcal{F}$ be a sub σ-algebra of $\mathcal{H}$, and $X = (X_1, \cdots, X_k)$ a k-dimensional random vector. We say that X is conditionally integrable given $\mathcal{F}$ if $X_1, \cdots, X_k$ are all conditionally integrable given $\mathcal{F}$, and in this case we define

$$\mathbb{E}[X \mid \mathcal{F}] = \begin{pmatrix} \mathbb{E}[X_1 \mid \mathcal{F}] \\ \vdots \\ \mathbb{E}[X_k \mid \mathcal{F}]. \end{pmatrix}$$

If X takes values in $\mathbb{R}^k_+$, so that its coordinates are all non-negative random variables, then $\mathbb{E}[X \mid \mathcal{F}]$ always exists and is defined as above. Again, $\mathbb{E}[X \mid \mathcal{F}]$ stands for the representative of an equivalence class.

We first state Jensen's inequality for conditional expectations; we did not state this as part of the preceding theorem because it is mostly used in a multivariate context.

Theorem 2.6 (Conditional Version of Jensen's Inequality)
Let $\mathcal{F}$ be a sub σ-algebra of $\mathcal{H}$, and X a k-dimensional real random vector. For any convex function $f : \mathbb{R}^k \to \mathbb{R}$, if $f \circ X$ and X are conditionally integrable given $\mathcal{F}$, then

$$\mathbb{E}[f \circ X \mid \mathcal{F}] \geq f(\mathbb{E}[X \mid \mathcal{F}]).$$

Proof) The proof proceeds almost identically to the proof of the unconditional version of the inequality.

Let W_f be the set of all affine minorants of f, that is, the collection of all affine functions $h : \mathbb{R}^k \to \mathbb{R}$ such that $h(x) \leq f(x)$ for any $x \in \mathbb{R}^k$. The characterization of convex functions (refer to the text on correspondences and convex analysis) tells us that

$$f(x) = \sup_{h \in W_f} h(x)$$

for any $x \in \mathbb{R}^k$. We use this to establish our claim.

For any $h \in W_f$, there exist $v \in \mathbb{R}^k$ and $c \in \mathbb{R}$ such that

$$h(x) = v'x + c$$

for any $x \in \mathbb{R}^k$, since h is an affine function. Since X is conditionally integrable given $\mathcal{F}$, and constants are trivially conditionally integrable, the linearity of conditional ex-

pectations tells us that $h \circ X$ is also conditionally integrable given $\mathcal{F}$ with

$$\mathbb{E}[h \circ X] = v' \cdot \mathbb{E}[X \mid \mathcal{F}] + c = h(\mathbb{E}[X \mid \mathcal{F}]).$$

Since $h(x) \leq f(x)$ for any $x \in \mathbb{R}^k$, we can see that $h \circ X \leq f \circ X$ and thus, by the monotonicity of conditional expectations,

$$\mathbb{E}[f \circ X] \geq \mathbb{E}[h \circ X] = h(\mathbb{E}[X \mid \mathcal{F}]).$$

This holds for any $h \in W_f$, so

$$\mathbb{E}[f \circ X] \geq \sup_{h \in W_f} h(\mathbb{E}[X \mid \mathcal{F}]).$$

On the other hand, the characterization of f reveals that

$$f(\mathbb{E}[X \mid \mathcal{F}]) = \sup_{h \in W_f} h(\mathbb{E}[X \mid \mathcal{F}]).$$

Therefore,

$$\mathbb{E}[f \circ X] \geq \sup_{h \in W_f} h(\mathbb{E}[X \mid \mathcal{F}]) = f(\mathbb{E}[X \mid \mathcal{F}]),$$

which is our desired result.

Q.E.D.

The result above allows us to draw the following connection between conditional expectations and uniform integrability:

Theorem 2.7 Let X an integrable k-dimensional real random vector, and define the collection

$$\mathcal{K} = \{\mathbb{E}[X \mid \mathcal{F}] \mid \mathcal{F} \text{ is a sub } \sigma\text{-algbra of } \mathcal{H} \}.$$

$\mathcal{K}$ is a uniformly integrable collection of real random vectors.

Proof) X is an integrable random vector, so that the singleton $\{X\}$ is trivially uniformly integrable. By the De La Valleé Poussin characteriztion of uniform integrability, there exists a non-negative increasing and convex function $f : \mathbb{R} \to \mathbb{R}_+$ such that $f(0) = 0$ and

$$\lim_{x \to \infty} \frac{f(x)}{x} = +\infty \quad \text{and} \quad \mathbb{E}[f \circ |X|] < +\infty.$$

Choose any $Z \in \mathcal{K}$; by definition, there exists a sub σ-algebra $\mathcal{F}$ of $\mathcal{H}$ such that

$$Z = \mathbb{E}[X \mid \mathcal{F}].$$

By the conditional version of Jensen's inequality, the convexity of the mapping $x \mapsto |x|$, and the conditional integrability of $|X|$ and X, we can see that

$$|Z| \leq \mathbb{E}[|X| \mid \mathcal{F}].$$

Since $f \circ |X|$ is integrable, it is also conditionally integrable; by the conditional version of Jensen's inequality once more, we have

$$f(\mathbb{E}[|X| \mid \mathcal{F}]) \leq \mathbb{E}[f \circ |X| \mid \mathcal{F}].$$

Finally, f is an increasing function, so

$$f \circ |Z| \leq f(\mathbb{E}[|X| \mid \mathcal{F}]) \leq \mathbb{E}[f \circ |X| \mid \mathcal{F}].$$

The monotonicity of integration now shows us that

$$\mathbb{E}[f \circ |Z|] \leq \mathbb{E}[f \circ |X|].$$

This holds for any $Z \in \mathcal{K}$, so

$$\sup_{Z \in \mathcal{K}} \mathbb{E}[f \circ |Z|] \leq \mathbb{E}[f \circ |X|] < +\infty.$$

By the De La Valleé Poussin characteriztion of uniform integrability once again, we can conclude that $\mathcal{K}$ is uniformly integrable.

Q.E.D.

2.2 Conditional Probabilities

So far, we have dealt with the conditional expectation of random variables given some sub σ-algebra $\mathcal{F}$ of $\mathcal{H}$. We can instead think of the more general concept of conditional probabilities, that is, an expression that summarizes the information in $\mathbb{P}(H \mid \mathcal{F})$ for any $H \in \mathcal{H}$, where we define

$$\mathbb{P}(H \mid \mathcal{F}) := \mathbb{E}[I_H \mid \mathcal{F}].$$

2.2.1 Regular Versions of Conditional Probabilities and Distributions

Transition probability kernels play a central role in the theory of conditional probabilities and distributions. Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. Recall that the function $K : E \times \mathcal{F} \to [0, +\infty]$ is called a transition probability kernel from $(E, \mathcal{E})$ to $(F, \mathcal{F})$ if:

- $K(\cdot, A)$ is a $\mathcal{E}$-measurable non-negative function for any $A \in \mathcal{F}$, and
- $K(x, \cdot)$ is a probability measure on $(F, \mathcal{F})$ for any $x \in E$.

Recall that, for any $x \in E$, we denote the integral of some function $\mathcal{F}$-measurable function f with respect to the probability measure $K(x, \cdot)$ as

$$\int_F f(y) K(x, dy),$$

granted that the integral exists.

Consider a function $Q : \Omega \times \mathcal{H} \to [0, 1]$, where $Q(\cdot, H)$ is taken to be a version of $\mathbb{P}(H \mid \mathcal{F})$ that takes values in $[0, 1]$ for any $H \in \mathcal{H}$, with $Q(\cdot, \emptyset) = 0$ and $Q(\cdot, \Omega) = 1$. Then, each $Q(\cdot, H)$ is an $\mathcal{F}$-measurable non-negative function, and for a disjoint collection $\{H_n\}_{n \in N_+}$ of $\mathcal{H}$-measurable sets with union $H = \bigcup_n H_n$,

$$\begin{aligned} Q(\cdot, H) = \mathbb{E}[I_H \mid \mathcal{F}] &= \mathbb{E}\left[\sum_{n=1}^{\infty} I_{H_n} \mid \mathcal{F}\right] \\ &= \sum_{n=1}^{\infty} \mathbb{E}[I_{H_n} \mid \mathcal{F}] = \sum_{n=1}^{\infty} Q(\cdot, H_n) \end{aligned}$$

by the conditional version of the MCT for series, where the equality holds almost surely.

If we can choose Q so that this equality holds everywhere on Ω for any disjoint collection $\{H_n\}_{n \in N_+} \subset \mathcal{H}$, then Q becomes a transition probability kernel from $(\Omega, \mathcal{F})$ into $(\Omega, \mathcal{H})$ such that

$$Q(\cdot, H) = \mathbb{P}(H \mid \mathcal{F})$$

for any $H \in \mathcal{H}$. In a sense, Q summarizes the information present in all conditional probabilities given $\mathcal{F}$; we call Q a regular version of the conditional probability $\mathbb{P}(\cdot \mid \mathcal{F})$. Heuristically,

we may view Q as representing the conditional analogue of the underlying probability measure $\mathbb{P}$.

Let X be a random variable taking values in the measurable space $(E,\mathcal{E})$, and let $\mathcal{F}$ be a sub σ-algebra of $\mathcal{H}$. Recall that the distribution of X as the pushforward measure of $\mathbb{P}$ with respect to X, that is, as $\mu = \mathbb{P} \circ X^{-1}$. The value of μ given a measurable set$A \in \mathcal{E}$ represents the probability of X taking values in the set A.

We can define the conditional distribution of X given $\mathcal{F}$ in a similar manner. Suppose that $Q : \Omega \times \mathcal{H} \to [0,1]$ is a regular version of the conditional probability given $\mathcal{F}$. Then, letting $L : \Omega \times \mathcal{E} \to [0,1]$ be defined as

$$L(\omega, A) = Q(\omega, X^{-1}(A))$$

for any $(\omega, A) \in \Omega \times \mathcal{E}$, L is referred to as a regular version of the conditional distribution of X given $\mathcal{F}$. Note that L is a transition probability kernel from $(\Omega,\mathcal{F})$ into $(E,\mathcal{E})$:

- For any $A \in \mathcal{E}$, $L(\cdot, A) = Q(\cdot, X^{-1}(A))$ is a $\mathcal{F}$-measurable non-negative function
- For any $\omega \in \Omega$, $L(\omega, \cdot)$ is the pushforward measure of $Q(\omega, \cdot)$ with respect to X and thus a probability measure on $(E,\mathcal{E})$.

For any $A \in \mathcal{E}$,

$$L(\cdot, A) = Q(\cdot, X^{-1}(A)) = \mathbb{P}(X \in A \mid \mathcal{F})$$

so that $L(\cdot, A)$ can be interpreted as the conditional probability of X taking values in A given $\mathcal{F}$. Like the regular version Q represents the conditional analogue of the underlying probability measure $\mathbb{P}$, L represents the conditional analogue of the distribution $\mu = \mathbb{P} \circ X^{-1}$ of X.

Given the parallels beteween the unconditional probability measure/distribution and the conditional versions, we might expect conditional expectations to be formulated in terms of integrals with respect to Q or L, like how expectations can be written as integrals with respect to $\mathbb{P}$ or μ. We confirm below that this is indeed the case:

Theorem 2.8 Let X be a random variable taking values in $(E,\mathcal{E})$, and $\mathcal{F}$ a sub σ-algebra of $\mathcal{H}$. Suppose there exists a regular version Q of the conditional probability given $\mathcal{F}$, using which we can define a regular version L of the conditional distribution of X given $\mathcal{F}$ as above. Then, for any $f \in \mathcal{E}_+$, the mappings

$$\omega \mapsto \int_\Omega (f \circ X)(t) Q(\omega, dt)$$

and

$$\omega \mapsto \int_E f(x) L(\omega, dx)$$

are both versions of the conditional expectation $\mathbb{E}[f \circ X \mid \mathcal{F}]$.

Proof) Define $\Lambda_Q, \Lambda_L : \Omega \to [0, +\infty]$ as

$$\Lambda_Q(\omega) = \int_\Omega (f \circ X)(t) Q(\omega, dt)$$

$$\Lambda_L(\omega) = \int_E f(x) L(\omega, dx)$$

for any $\omega \in \Omega$. Since $L(\omega, \cdot) = Q(\omega, \cdot) \circ X^{-1}$, by theorem 1.1, we can immediately see that

$$\Lambda_L(\omega) = \int_E f(x) L(\omega, dx) = \int_\Omega (f \circ X)(t) Q(\omega, dt) = \Lambda_Q(\omega)$$

for any $\omega \in \Omega$, so that

$$\Lambda(f) := \Lambda_L = \Lambda_Q$$

on Ω. It remains to show that $\Lambda(f)$ is a version of the conditional expectation of $f \circ X$ given $\mathcal{F}$.

Suppose f is a non-negative $\mathcal{E}$-measurable simple function with canonical form

$$f = \sum_{i=1}^{n} a_i \cdot I_{A_i},$$

where $a_1, \cdots, a_n \in [0, +\infty)$ and $A_1, \cdots, A_n \in \mathcal{E}$ form a measurable partition of E. Then, for any $\omega \in \Omega$,

$$\int_E f(x) L(\omega, dx) = \sum_{i=1}^{n} a_i \cdot L(\omega, A_i),$$

by the linearity of integration of non-negative functions, so that

$$\Lambda(f) = \sum_{i=1}^{n} a_i \cdot L(\cdot, A_i).$$

Each $L(\cdot, A_i)$ is $\mathcal{F}$-measurable and non-negative by definition, so it follows that $\Lambda(f)$ is a non-negative $\mathcal{F}$-measurable function. Furthermore, for any $H \in \mathcal{F}$,

$$\begin{aligned} \mathbb{E}[\Lambda(f) \cdot I_H] &= \mathbb{E}\left[\sum_{i=1}^{n} a_i \cdot L(\cdot, A_i) I_H\right] \\ &= \sum_{i=1}^{n} a_i \cdot \mathbb{E}[L(\cdot, A_i) \cdot I_H] \end{aligned}$$

by the linearity of expectations of non-negative functions. Since each $L(\cdot, A_i)$ is a version

of $\mathbb{E}\left[X^{-1}(A_i) \mid \mathcal{F}\right]$, by the defining property of conditional expectations

$$\mathbb{E}[L(\cdot, A_i) \cdot I_H] = \mathbb{E}\left[X^{-1}(A_i) \cdot I_H\right];$$

it follows that

$$\begin{aligned}\mathbb{E}[\Lambda(f) \cdot I_H] &= \sum_{i=1}^{n} a_i \cdot \mathbb{E}\left[X^{-1}(A_i) \cdot I_H\right] \\ &= \mathbb{E}\left[\left[\left(\sum_{i=1}^{n} a_i \cdot I_{A_i}\right) \circ X\right] \cdot I_H\right] = \mathbb{E}[(f \circ X) \cdot I_H].\end{aligned}$$

Therefore, if f is a non-negative simple $\mathcal{F}$-measurable function, $\Lambda(f)$ is a version of $\mathbb{E}[f \circ X \mid \mathcal{F}]$.

Now let f be an arbitrary non-negative $\mathcal{E}$-measurable function. Then, there exists a sequence $\{f_n\}_{n \in N_+}$ of non-negative $\mathcal{E}$-measurable simple functions increasing to f. By the MCT, we can see that, for any $\omega \in \Omega$,

$$\int_E f(x) L(\omega, dx) = \lim_{n \to \infty} \int_E f_n(x) L(\omega, dx),$$

so that

$$\Lambda(f) = \lim_{n \to \infty} \Lambda(f_n).$$

We showed above that $\{\Lambda(f_n)\}_{n \in N_+}$ is a sequence of non-negative $\mathcal{F}$-measurable functions, so it follows that their pointwise limit $\Lambda(f)$ should also be a non-negative $\mathcal{F}$-measurable function. Note also that $\{\Lambda(f_n)\}_{n \in N_+}$ is itself an increasing sequence of functions by the monotonicity of integration.

Similarly, by repeatedly using the MCT, we can see that, for any $H \in \mathcal{F}$,

$$\begin{aligned}\mathbb{E}[\Lambda(f) \cdot I_H] &= \lim_{n \to \infty} \mathbb{E}[\Lambda(f_n) \cdot I_H] \\ &= \lim_{n \to \infty} \mathbb{E}[(f_n \circ X) \cdot I_H] = \mathbb{E}[(f \circ X) \cdot I_H].\end{aligned}$$

By definition, $\Lambda(f)$ is a version of $\mathbb{E}[f \circ X \mid \mathcal{F}]$.

Q.E.D.

We have thus seen that, in almost every sense, the transition probability kernels Q and L play the role of the conditional analogue of the underlying probability measure $\mathbb{P}$ and distribution μ of a random variable X. However, unlike the underlying probability measure $\mathbb{P}$, we do not know whether the regular version Q of the conditional probability $\mathcal{F}$ even exists. Are there any sufficient conditions under which Q exists?

Instead of directly addressing the question of the existence of Q, we can address a related, but distinct, question. Suppose X, Y are two random variables taking values in the measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$ with joint distribution $\pi = \mathbb{P} \circ (X, Y)^{-1}$. Letting μ and v be the distributions of X and Y, recall that we say that X and Y are independent if the joint distribution π can be expressed as the product $\pi = \mu \times v$. However, what if X and Y are not independent? Can we still decompose π into the product of μ and some other mathematical object in this case?

Fortunately, the answer to both questions is yes. The decomposition $\pi = \mu \times v$ turns out to be a special case of the general case where π is decomposed into the "product" of μ and some transition probability kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. In the next section we study the general method of constructing a probability measure on a product space from a measure and a transition probability kernel, and then furnish sufficient conditions for a probability measure on a product space into the product of a measure and a transition probability kernel. We then show that the transition probability kernel in question can be used to construct a regular verison of conditional distribution of Y given X[1], as well as a regular version of the conditional probability given X.

2.2.2 Construction of Probability Measures

One of the main uses of transition probability kernels is in the construction of measures on the product space $(E \times F, \mathcal{E} \bigotimes \mathcal{F})$. To this end, we first show that, for any transition probability kernel K from $(E, \mathcal{E})$ to $(F, \mathcal{F})$ and non-negative $\mathcal{E} \bigotimes \mathcal{F})$-measurable function f, the mapping

$$x \mapsto \int_F f_x(y) K(x, dy)$$

is a non-negative $\mathcal{E}$-measurable function, where $f_x : F \to [0, +\infty]$ is the section of f defined as $f_x(y) = f(x, y)$ for any $y \in F$. Below, we let $(\mathcal{E} \bigotimes \mathcal{F})_{+,b}$ denote the collection of all non-negative or bounded $\mathcal{E} \bigotimes \mathcal{F}$-measurable functions.

Lemma 2.9 Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ be measurable spaces. For any transition probability kernel $K : E \times \mathcal{F} \to [0, 1]$ from $(E, \mathcal{E})$ to $(F, \mathcal{F})$. For any non-negative (bounded) $\mathcal{E} \bigotimes \mathcal{F}$-measurable function f,

$$(T_K f)(x) = \int_F f_x(y) K(x, dy)$$

is well-defined for any $x \in E$ and the function $T_K f : E \to [-\infty, +\infty]$ is a $\mathcal{E}$-measurable non-negative (bounded) function.

Furthermore, the transformation $T_K : (\mathcal{E} \bigotimes \mathcal{F})_{+,b} \to \mathcal{E}_{+,b}$ possesses the following properties:

i) **(Linearity)** For any non-negative (real-valued) scalar a and non-negative (bounded)

[1] When we say given X, we mean given the sub σ-algebra σX of $\mathcal{H}$.

$\mathcal{E} \bigotimes \mathcal{F}$-measurable functions f, g,

$$T_K(af + g) = a \cdot (T_K f) + (T_K g).$$

ii) **(Sequential Continuity)** For any increasing sequence $\{f_n\}_{n \in N_+}$ of non-negative $\mathcal{E} \bigotimes \mathcal{F}$-measurable functions with pointwise limit f, $\{T_K f_n\}_{n \in N_+}$ is an increasing sequence of non-negative $\mathcal{E}$-measurable functions such that

$$T_K f = \sup_{n \in N_+} T_K f_n.$$

Proof) We first show that, for any $f \in (\mathcal{E} \bigotimes \mathcal{F})_{+,b}$, $(T_K f)(x)$ is well-defined for any $x \in E$. First note that the section $f_x : F \to [-\infty, +\infty]$ is a $\mathcal{F}$-measurable function[2]. If f is non-negative, then so is f_x, and therefore the integral

$$(T_K f)(x) = \int_F f_x(y) K(x, dy)$$

is well-defined and takes values in $[0, +\infty]$. Meanwhile, if f is bounded, then there exists an $M > 0$ such that $|f| \leq M$, and by the monotonicity of integration and the fact that $K(x, \cdot)$ is a probability measure, we have

$$\int_F |f_x(y)| K(x, dy) \leq M < +\infty.$$

This shows us that f_x is $K(x, \cdot)$-integrable, and as such that

$$(T_K f)(x) = \int_F f_x(y) K(x, dy)$$

is again well-defined and takes values in $\mathbb{R}$.

Now we show that the mapping T_K from the set of all non-negative or bounded $\mathcal{E} \bigotimes \mathcal{F}$-measurable functions to numerical functions on E possess linearity and sequential continuity properties. Choose any $\mathcal{E} \bigotimes \mathcal{F}$-measurable functions f, g that are both bounded (and thus real-valued), and let $a \in \mathbb{R}$. $af + g$ is again a bounded $\mathcal{E} \bigotimes \mathcal{F}$-measurable function, so $T_K(af + g)$ is a well-defined real valued function on E. For any $x \in E$, by

[2]This can easily be seen as follows. Define $g : F \to E \times F$ as $g(y) = (x, y)$ for any $y \in F$. Then, for any measurable rectangle $A \times B$ on $(E \times F, \mathcal{E} \bigotimes \mathcal{F})$, we have

$$g^{-1}(A \times B) = \begin{cases} \emptyset & \text{if } x \notin A \\ B & \text{if } x \in A \end{cases} \in \mathcal{F},$$

which implies that g is measurable relative to $\mathcal{F}$ and $\mathcal{E} \bigotimes \mathcal{F}$, since the set of all measurable rectangles generates the product σ-algebra $\mathcal{E} \bigotimes \mathcal{F}$. Since $f_x = f \circ g$, the preservation of measurability across compositions and the $\mathcal{E} \bigotimes \mathcal{F}$-measurability of f ensures that f_x is $\mathcal{F}$-measurable.

the linearity of integration for real integrable functions,

$$(T_K(af+g))(x) = \int_F (af+g)_x(y)K(x,dy) = \int_F (a\cdot f_x(y)+g_x(y))\,K(x,dy)$$
$$= a\cdot\int_F f_x(y)K(x,dy)+\int_F g_x(y)K(x,dy) = a\cdot(T_Kf)(x)+(T_Kg)(x).$$

This shows us that $T_K(af+g) = a\cdot(T_Kf)+(T_Kg)$. On the other hand, if f,g are non-negative instead of bounded, and $a\in[0,+\infty]$, then $af+g$ is again a non-negative $\mathcal{E}\bigotimes\mathcal{F}$-measurable function, so that $T_K(af+g)$ is a well-defined non-negative function on E. The linearity result then follows from the linearity of integration for non-negative functions.

Now suppose that $\{f_n\}_{n\in N_+}$ is an increasing seqeunce of non-negative $\mathcal{E}\bigotimes\mathcal{F}$-measurable functions. Then, defining $f=\sup_{n\in N_+} f_n$, f is also in $(\mathcal{E}\bigotimes\mathcal{F})_+$ because measurability is preserved across pointwise supremums, so that T_Kf is a non-negative function on E. Furthermore, for any $x\in E$, since $\{f_{n,x}\}_{n\in N_+}$ is an increasing sequence of non-negative $\mathcal{F}$-measurable functions increasing to f_x, the MCT shows us that

$$(T_Kf)(x) = \int_F f_x(y)K(x,dy) = \lim_{n\to\infty}\int_F f_{n,x}(y)K(x,dy) = \lim_{n\to\infty}(T_Kf_n)(x).$$

The monotonicity of integration also shows us that $T_Kf_n\le T_Kf_{n+1}$ for any $n\in N_+$, so $\{T_Kf_n\}_{n\in N_+}$ is a sequence of non-negative functions on E that increase to T_Kf.

It remains to show that T_Kf is $\mathcal{E}$-measurable for any bounded or non-negative $\mathcal{E}\bigotimes\mathcal{F}$-measurable function. To this end, we make use of the monotone class theorem for functions: define the collection $\mathcal{M}$ as

$$\mathcal{M} = \{f\in\left(\mathcal{E}\bigotimes\mathcal{F}\right)_+ \mid T_Kf \text{ is } \mathcal{E}\text{-measurable.}\}$$

Note that, for any measurable rectangle $A\times B$ on $(E\times F,\mathcal{E}\bigotimes\mathcal{F})$,

$$(T_KI_{A\times B})(x) = \int_F I_A(x)I_B(y)K(x,dy) = I_A(x)\cdot K(x,B)$$

for any $x\in E$, so that $T_KI_{A\times B} = I_A\cdot K(\cdot,B)$. Since I_A and $K(\cdot,B)$ are both $\mathcal{E}$-measurable non-negative functions, so is $T_KI_{A\times B}$; this shows us that $I_{A\times B}\in\mathcal{M}$. We now show that $\mathcal{M}$ is a monotone class of functions:

i) $I_{E\times F}\in\mathcal{M}$ because $E\times F$ is a measurable rectangle.

ii) For any $a,b\in\mathbb{R}$ and $f,g\in\mathcal{M}_b$, since $f,g\in(\mathcal{E}\bigotimes\mathcal{F})_b$, by the linearity result shown above,

$$T_K(af+bg) = a\cdot(T_Kf)+b\cdot(T_Kg).$$

Since $T_K f, T_K g$ are both bounded $\mathcal{E}$-measurable functions because $f, g \in \mathcal{M}$, the preservation of measurability across linear combinations shows us that $T_K(af+bg)$ is also a bounded $\mathcal{E}$-measurable function, and as such that $af+bg \in \mathcal{M}_b$.

iii) For any increasing sequence $\{f_n\}_{n\in N_+} \subset \mathcal{M}_+$, since $\{f_n\}_{n\in N_+}$ is an increasing sequence of non-negative $\mathcal{E}\bigotimes\mathcal{F}$-measurable functions, the sequential continuity result shown above shows us that

$$T_K f = \sup_{n\in N_+} T_K f_n.$$

By assumption, $\{T_K f_n\}_{n\in N_+}$ is a sequence of $\mathcal{E}$-measurable non-negative functions, so it follows that $T_K f \in \mathcal{E}_+$ as well. Therefore, $f \in \mathcal{M}_+$.

Thus, $\mathcal{M}$ is a monotone class of functions that contains all indicator functions of the form $I_{A\times B}$. Since the collection of all measurable rectangles on $E\times F$ forms a π-system that generates the product σ-algebra $\mathcal{E}\bigotimes\mathcal{F}$, the monotone class theorem for functions tells us that any non-negative or bounded $\mathcal{E}\bigotimes\mathcal{F}$-measurable function is contained in $\mathcal{M}$. In other words, for any $f \in (\mathcal{E}\bigotimes\mathcal{F})_{+,b}$, the function $T_K f : E \to [-\infty, +\infty]$ is $\mathcal{E}$-measurable.

Q.E.D.

The next result shows us how to construct a probability measure on $(E\times F, \mathcal{E}\bigotimes\mathcal{F})$ given a probability measure on $(E,\mathcal{E})$ and a transition probability kernel from $(E,\mathcal{E})$ into $(F,\mathcal{F})$.

Theorem 2.10 (Construction of Probability Measures on Product Spaces)
Let $(E,\mathcal{E})$ and $(F,\mathcal{F})$ be measurable spaces. Let μ be a probability measure on $(E,\mathcal{E})$ and K a transition probability kernel from $(E,\mathcal{E})$ into $(F,\mathcal{F})$. Then, there exists a probability measure π on $(E\times F, \mathcal{E}\bigotimes\mathcal{F})$ such that

$$\int_{E\times F} f d\pi = \int_E (T_K f) d\mu$$

for any $f \in (\mathcal{E}\bigotimes\mathcal{F})_+$. Furthermore, π is the unique measure on $(E\times F, \mathcal{E}\bigotimes\mathcal{F})$ such that

$$\pi(A\times B) = \int_A K(\cdot, B) d\mu$$

for any measurable rectangle $A\times B$ on $E\times F$.

Proof) Define the functional $\Lambda : (\mathcal{E} \bigotimes \mathcal{F})_+ \to [0, +\infty]$ as

$$\Lambda(f) = \int_E (T_K f) d\mu$$

for any $f \in (\mathcal{E} \bigotimes \mathcal{F})_+$. We verify that Λ satisfies the following three properties:

- Λ **assigns 0 to the empty set**
  $\Lambda(I_\emptyset) = \int_E (T_K I_\emptyset) d\mu = 0$ because $T_K I_\emptyset = 0$ by the definition of the mapping T_K.

- Λ **is a linear functional**
  For any $a \in [0, +\infty]$ and $f, g \in (\mathcal{E} \bigotimes \mathcal{F})_+$, by the linearity of the mapping $T_K : (\mathcal{E} \bigotimes \mathcal{F})_+ \to \mathcal{E}_+$ and the linearity of the integration of non-negative functions, we have

$$\begin{aligned} \Lambda(af + g) = \int_E T_K (af + g) \, d\mu &= \int_E (a \cdot (T_K f) + T_K g) \, d\mu \\ &= a \cdot \int_E (T_K f) d\mu + \int_E (T_K g) d\mu = a \cdot \Lambda(f) + \Lambda(g). \end{aligned}$$

- Λ **is sequentially continuous**
  For any increasing sequence $\{f_n\}_{n \in N_+}$ of non-negative $\mathcal{E} \bigotimes \mathcal{F}$-measurable functions with pointwise limit $f \in (\mathcal{E} \bigotimes \mathcal{F})_+$, the preceding result tells us that $\{T_K f_n\}_{n \in N_+}$ is an increasing sequence of non-negative $\mathcal{E}$-measurable functions with pointwise limit $T_K f$. By the MCT, this implies that

$$\Lambda(f) = \int_E (T_K f) d\mu = \lim_{n \to \infty} \int_E (T_K f_n) d\mu = \lim_{n \to \infty} \Lambda(f_n).$$

  Furthermore, by the monotonicity of integration, $\{\Lambda(f_n)\}_{n \in N_+}$ is an increasing sequence of functions, so that

$$\Lambda(f_n) \nearrow \Lambda(f).$$

The elementary representation theorem for linear sequentially continuous functionals stated in the measure theory text now implies that there exists a measure π on $(E \times F, \mathcal{E} \bigotimes \mathcal{F})$ such that

$$\int_{E \times F} f d\pi = \Lambda(f)$$

for any $f \in (\mathcal{E} \bigotimes \mathcal{F})_+$. Since

$$\pi(E \times F) = \Lambda(I_{E \times F}) = \int_E K(\cdot, F) d\mu = 1,$$

π is a probability measure, and since the collection of all measurable rectangles on $E \times F$ is a π-system generating $\mathcal{E} \bigotimes \mathcal{F}$, π is equivalent to any probability measure v such that π and v agree on all measurable rectangles $A \times B$ on $E \times F$. Since

$$\pi(A \times B) = \int_A K(\cdot, B) d\mu,$$

this proves the uniqueness claim.

Q.E.D.

We denote the probability measure π constructed above as $\pi = \mu \times K$, and we often write

$$\int_{E \times F} f d\pi = \int_E (T_K f) d\mu = \int_E \int_F f(x,y) K(x, dy) d\mu(x)$$

for any $f \in (\mathcal{E} \bigotimes \mathcal{F})_+$. The measure π is then called the product of the probability measure μ and the transition probability kernel K.

2.2.3 Decomposition of Joint Distributions

What we are actually interested in is the opposite of the above case: specifically, we want to furnish sufficient conditions under which a probability measure π on the product space can be expressed as the product of some probability measure μ on $(E, \mathcal{E})$ and a transition probability kernel K from $(E, \mathcal{E})$ to $(F, \mathcal{F})$. The following result shows how the decomposition problem is intimately related to the existence of regular versions of conditional distributions of random variables:

Theorem 2.11 Let X and Y be random variables taking values in the measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$. Denote the joint distribution of X and Y by π, and suppose that it has the decomposition $\pi = \mu \times K$ for some probability measure μ on $(E, \mathcal{E})$ and transition probability kernel K from $(E, \mathcal{E})$ into $(F, \mathcal{F})$. Then, μ is the marginal distribution of X, and defining $L : \Omega \times \mathcal{F} \to [0,1]$ as

$$L(\omega, A) = K(X(\omega), A)$$

for any $(\omega, A) \in \Omega \times \mathcal{F}$, L is a regular version of the conditional distribution of Y given X. Furthermore, for any $f \in (\mathcal{E} \bigotimes \mathcal{F})_+$,

$$\mathbb{E}[f(X,Y) \mid X] = (T_K f) \circ X = \int_F f(X, y) K(X, dy).$$

Proof) We first show that μ is the marginal distribution of X. For any $A \in \mathcal{E}$, since $\pi = \mu \times K$,

we have

$$\pi(A \times F) = \int_A K(\cdot, F) d\mu = \mu(A)$$

since K is a transition probability kernel. By definition, μ is the marginal distribution of X.

Now we show that L is a regular version of the conditional distribution of Y given X. It is easy to see that L is a transition probability kernel from $(\Omega, \sigma X)$ into $(F, \mathcal{F})$; for any $A \in \mathcal{F}$,

$$L(\cdot, A) = K(\cdot, A) \circ X,$$

so that $L(\cdot, A)$ is σX-measurable by the Doob-Dynkin lemma, while for any $\omega \in \Omega$,

$$L(\omega, \cdot) = K(X(\omega), \cdot)$$

is a probability measure on $(F, \mathcal{F})$. Now choose any $A \in \mathcal{F}$. We can see that, for any $H \in \sigma X$, there exists a $B \in \mathcal{E}$ such that $H = X^{-1}(B)$ and thus

$$\begin{aligned}
\mathbb{E}[L(\cdot, A) \cdot I_H] &= \mathbb{E}[(K(\cdot, A) \cdot I_B) \circ X] \\
&= \int_E (K(\cdot, A) \cdot I_B) \, d\mu \\
&= \int_B K(\cdot, A) d\mu = \pi(B \times A) \\
&= \mathbb{E}[(I_B \circ X)(I_A \circ Y)] = \mathbb{E}[(I_A \circ Y) \cdot I_H],
\end{aligned}$$

where the second and fifth equalities are justified by the definitions of the distribution of random variables/vectors. Since $L(\cdot, A)$ is σX-measurable, the above result tells us that $L(\cdot, A)$ is a version of the conditional expectation $\mathbb{E}\left[I_{Y^{-1}(A)} \mid X\right] = \mathbb{P}(Y \in A \mid X)$. By definition, L is a version of the conditional distribution of Y given X.

Finally, let f be a non-negative $\mathcal{E} \bigotimes \mathcal{F}$-measurable function. Then, for any $H \in \sigma X$, letting $A \in \mathcal{E}$ be chosen so that $H = X^{-1}(A)$, note that

$$\begin{aligned}
\mathbb{E}[f(X, Y) \cdot I_H] &= \mathbb{E}[(f \cdot I_{A \times F}) \circ (X, Y)] \\
&= \int_{E \times F} (f \cdot I_{A \times F}) \, d\pi \\
&= \int_A \int_F f(x, y) K(x, dy) d\mu(x) = \int_A (T_K f) \, d\mu \\
&= \mathbb{E}[((T_K f) \cdot I_A) \circ X] = \mathbb{E}[((T_K f) \circ X) \cdot I_H].
\end{aligned}$$

Since $(T_K f) \circ X$ is trivially σX-measurable by the Doob-Dynkin lemma, this shows us that $(T_K f) \circ X$ is a version of the conditional expectation $\mathbb{E}[f(X, Y) \mid X]$.

Q.E.D.

The quantity K allows us to give a mathematically rigorous definition of the conditional probability of $Y \in A$ given $X = x$; since $L(\cdot, A) = K(\cdot, A) \circ X$ is a version of $\mathbb{P}(Y \in A \mid X)$, it stands to reason that we may define

$$\mathbb{P}(Y \in A \mid X = x) = K(x, A)$$

for any $(x, A) \in E \times \mathcal{F}$. This also allows us to rigorously define the conditional density of Y given $X = x$; suppose that there exists some σ-finite measure v on $(F, \mathcal{F})$ such that $K(x, \cdot)$ is absolutely continuous with respect to v for any $x \in E$. By the Radon-Nikodym theorem, for any $x \in E$ there should exist a function $k_x \in \mathcal{F}_+$ such that

$$K(x, A) = \int_A k_x(y) dv(y)$$

for any $(x, A) \in E \times \mathcal{F}$. If the function $f_{Y|X} : E \times F \to [0, +\infty]$ defined as

$$f_{Y|X}(y \mid x) = k_x(y)$$

for any $(x, y) \in E \times F$ is $\mathcal{E} \bigotimes \mathcal{F}$-measurable, then it is referred to as the conditional density of Y given $X = x$ with respect to v. Note that, in this case,

$$\int_A f_{Y|X}(y \mid X) dv(y)$$

is a version of the conditional probability of $Y \in A$ given X, for any $A \in \mathcal{F}$.

Now that we showed how the decomposition $\pi = \mu \times K$ allows us to rigorously define various useful concepts, we again turn back to the problem of the existence of this decomposition. The next theorem shows that a common assumption made in probability theory is actually sufficient for the above decomposition. Specifically, if we interpret π as the joint distribution of two random variables X and Y, then it is sufficient for π to have a joint density with respect to a product of σ-finite measures for π to be decomposed into the product of a probability measure and a transition probability kernel. The most common way in which this setting occurs in probability theory is through the assumption of the existence of a joint pdf with respect to the Lebesgue measure on a multidimensional euclidean space, where this Lebesgue measure can be seen as the product of Lebesgue measures of lower dimensional euclidean spaces.

Theorem 2.12 (Decomposition of Probability Measures on Product Spaces) Let $(E,\mathcal{E})$ and $(F,\mathcal{F})$ be measurable spaces, and π a probability measure on the product space $(E\times F,\mathcal{E}\bigotimes\mathcal{F})$. Let μ and v be σ-finite measures on $(E,\mathcal{E})$ and $(F,\mathcal{F})$, respectively, and suppose that there exists a density $f\in(\mathcal{E}\bigotimes\mathcal{F})_+$ of π with respect to the product measure $\mu\times v$, which is σ-finite[3]. Let P be the probability measure on $(E,\mathcal{E})$ defined as $P(A)=\pi(A\times F)$ for any $A\in\mathcal{E}$. Then, the following hold true:

i) Defining $g:E\to[0,+\infty]$ as

$$g(x)=\int_F f(x,y)dv(y)$$

for any $x\in E$, g is a non-negative $\mathcal{E}$-measurable function such that

$$P(A)=\int_A g(x)d\mu(x)$$

for any $A\in\mathcal{E}$. By implication, $P<<\mu$.

ii) Defining $k:E\times F\to[0,+\infty]$ as

$$k(x,y)=\begin{cases}\frac{f(x,y)}{g(x)} & \text{if } 0<g(x)<+\infty\\ \int_E f(x,y)d\mu(x) & \text{if } g(x)=0 \text{ or } +\infty\end{cases}$$

for any $(x,y)\in E\times F$, k is a non-negative $\mathcal{E}\bigotimes\mathcal{F}$-measurable function that satisfies

$$f(x,y)=g(x)k(x,y)$$

for any $y\in F$ and P-a.e. $x\in E$.

iii) Defining $K:E\times\mathcal{F}\to[0,1]$ as

$$K(x,A)=\int_A k(x,y)dv(y)$$

for any $(x,A)\in E\times\mathcal{F}$, K is a transition probability kernel from $(E,\mathcal{E})$ into $(F,\mathcal{F})$ such that

$$\pi=P\times K.$$

Proof) The function g defined as above is $\mathcal{E}$-measurable due to Fubini's theorem for non-negative functions. To see that P is the indefinite integral of g with respect to μ, note

[3]In light of the Radon-Nikodym theorem, condition is equivalent to the assumption that π is absolutely continuous with respect to $\mu\times v$

that, for any $A \in \mathcal{E}$,

$$\begin{aligned} P(A) = \pi(A \times F) &= \int_{E \times F} (f \cdot I_{A \times F})\, d(\mu \times v) \\ &= \int_A \int_F f(x,y) dv(y) d\mu(x) \qquad \text{(Fubini's theorem)} \\ &= \int_A g(x) d\mu(x). \end{aligned}$$

It follows immediately that $P << \mu$, that is, P is absolutely continuous with resepct to μ.

Now define the function k as above. Letting $G = \{0 < g < +\infty\} \in \mathcal{E}$, define the functions $g_f : E \to [0, +\infty]$ and $\psi : F \to [0, +\infty]$ as

$$g_f(x) = g(x) \cdot I_G(x) + I_{G^c}(x)$$

and

$$\psi(y) = \int_E f(t,y) d\mu(t)$$

for any $(x,y) \in E \times F$. ψ is $\mathcal{F}$-measurable by Fubini's theorem, and g_f is also trivially $\mathcal{E}$-measurable; note that g_f is non-zero and finite everywhere on E while agreeing with g on the points at which g is non-zero and finite. Furthermore, the extensions $g_f^e : E \times F \to [0, +\infty]$ and $\psi^e : E \times F \to [0, +\infty]$ are $\mathcal{E} \bigotimes \mathcal{F}$-measurable; this follows because, for any $a \in \mathbb{R}$,

$$\{g_f^e < a\} = \{g_f < a\} \times F \in \mathcal{E} \bigotimes \mathcal{F},$$

and similarly for ψ^e. Since k can be written as

$$k = \frac{f}{g_f^e} \cdot I_{G \times F} + \psi^e \cdot I_{G^c \times F},$$

and measurability is preserved across arithmetic operations, it follows that k is a non-negative $\mathcal{E} \bigotimes \mathcal{F}$-measurable function.

For any $(x,y) \in E \times F$, if $0 < g(x) < +\infty$, then

$$f(x,y) = g(x)k(x,y)$$

by definition. Define $G_0 = \{g = 0\}$ and $G_\infty = \{g = +\infty\}$. Since

$$P(E) = \pi(E \times F) = 1 = \int_E g(x) d\mu(x),$$

by the finiteness property of non-negative functions we have $\mu(G_\infty) = 0$. The absolute continuity of P with respect to μ implies that $P(G_\infty) = 0$ as well. In addition,

$$P(G_0) = \int_{G_0} g(x) d\mu(x) = 0$$

because $g(x) = 0$ on G_0; therefore, $P(G^c) = P(G_0 \cup G_\infty) = 0$. It follows that $f(x,y) = g(x)k(x,y)$ holds for any $y \in F$ and P-a.e. $x \in E$.

Finally, let $K : E \times \mathcal{F} \to [0,1]$ be defined as above; it is well-defined because k is a non-negative $\mathcal{E} \bigotimes \mathcal{F}$-measurable function. For any $A \in \mathcal{F}$,

$$K(\cdot, A) = \int_F (k \cdot I_{E \times A})(x,y) dv(y),$$

where $k \cdot I_{E \times A} \in (\mathcal{E} \bigotimes \mathcal{F})_+$, so by Fubini's theorem $K(\cdot, A) \in \mathcal{E}_+$. For any $x \in E$, $K(x, \cdot)$ is the indefinite integral of the section $k(x, \cdot)$, which is a non-negative $\mathcal{F}$-measurable function, with respect to v, and therefore is a measure on $(F, \mathcal{F})$. If $0 < g(x) < +\infty$, then

$$K(x,F) = \int_F k(x,y) dv(y) = \frac{1}{g(x)} \int_F f(x,y) dv(y) = 1,$$

while if $g(x) = 0$ or $g(x) = +\infty$, then

$$K(x,F) = \int_F \int_E f(x,y) d\mu(x) dv(y) = \pi(E \times F) = 1,$$

so that $K(x, \cdot)$ is a probability measure in any case. This proves that K is a transition probability kernel from $(E, \mathcal{E})$ into $(F, \mathcal{F})$.

It remains to show that π is equivalent to the measure $P \times K$. Since both are probability measures, this reduces to checking that they agree on all measurable rectangles $A \times B$ on $E \times F$. Choose any $A \in \mathcal{E}$ and $B \in \mathcal{F}$. Note first that, if $g(x) = 0$, then

$$\int_B f(x,y) dv(y) \leq \int_F f(x,y) dv(y) = g(x) = 0,$$

so that $\int_B f(x,y) dv(y) = 0$ and thus

$$\int_{A \cap G_0} \left(\int_B f(x,y) dv(y) \right) d\mu(x) = 0.$$

Meanwhile, since $\mu(G_\infty) = 0$, we trivially have

$$\int_{A \cap G_\infty} \left(\int_B f(x,y) dv(y) \right) d\mu(x) = 0.$$

Now we can see that

$$\begin{aligned}
\pi(A\times B) &= \int_A\int_B f(x,y)dv(y)d\mu(x)\\
&= \int_{A\cap G}\left(\int_B f(x,y)dv(y)\right)d\mu(x)+\int_{A\cap G^c}\left(\int_B f(x,y)dv(y)\right)d\mu(x)\\
&= \int_{A\cap G}\left(\int_B k(x,y)dv(y)\right)g(x)d\mu(x)\\
&\quad +\int_{A\cap G_0}\left(\int_B f(x,y)dv(y)\right)d\mu(x)+\int_{A\cap G_\infty}\left(\int_B f(x,y)dv(y)\right)d\mu(x)\\
&= \int_{A\cap G}(K(\cdot,B)g)\,d\mu \qquad \text{(By definition of } K\text{)}\\
&= \int_{A\cap G}K(\cdot,B)dP \qquad \text{(By the property of indefinite integrals)}\\
&= \int_A K(\cdot,B)dP. \qquad \text{(Since } P(G^c)=P(G_0\cup G_\infty)=0\text{)}
\end{aligned}$$

This proves that $\pi = P\times K$ on the entire product space.

Q.E.D.

The quantities g, k, P and K that appear during the proof have very clear probabilistic interpretations. To make the connections clearer, we re-state the results of the previous theorem in terms of probabilistic concepts:

Corollary to Theorem 2.12 (Probabilistic Interpretation of Theorem 5.12)
Let X and Y be random variables taking values in, and μ and v σ-finite measures on, the measurable spaces $(E,\mathcal{E})$ and $(F,\mathcal{F})$. Denote the joint distribution of X and Y by π and the marginal distribution of X by P_X. Suppose that π is absolutely continuous with respect to the product measure $\mu \times v$, and that $f_{X,Y}$ is the joint density of X,Y with respect to $\mu \times v$. Then, the following hold true:

i) **(Marginal Density of X)**
Defining $f_X : E \to [0,+\infty]$ as

$$f_X(x) = \int_F f_{X,Y}(x,y)dv(y)$$

for any $x \in E$, f_X is the marginal density of X with respect to μ.

ii) **(Conditional Density of Y given $X = x$)**
Defining $f_{Y|X} : E \times F \to [0,+\infty]$ as

$$f_{Y|X}(y \mid x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)} & \text{if } 0 < f_X(x) < +\infty \\ \int_E f_{X,Y}(x,y)d\mu(x) & \text{if } f_X(x) = 0 \text{ or } +\infty \end{cases}$$

for any $(x,y) \in E \times F$, $f_{Y|X}$ is the conditional density of Y given X with respect to v and satisfies

$$f_{X,Y}(x,y) = f_{Y|X}(y \mid x) f_X(x)$$

for any $y \in F$ and P_X-a.e. $x \in E$.

iii) **(Conditional Distribution of Y given $X = x$)**
Defining $K_{Y|X} : E \times \mathcal{F} \to [0,1]$ as

$$K_{Y|X}(x,A) = \int_F f_{Y|X}(y \mid x)dv(y)$$

for any $(x,A) \in E \times \mathcal{F}$, $K_{Y|X}(x,A)$ is the conditional probability of $Y \in A$ given $X = x$.

In light of theorem 5.11, the last point shows us that there exists a regular version of the conditional distribution of Y given X.

Proof) This is just a re-statement of the results of theorem 2.12.

Q.E.D.

2.2.4 Bayes' Rule

One of the most useful applications of the results of theorem 2.12 in probability theory is Bayes' rule. Note the symmetry of the preceding statement; given the distribution π of X and Y, we chose to first obtain the marginal density of X, and then use it to construct the conditional density of Y given X. However, we could have instead first obtained the marignal density of Y, and used that to construct the conditional density of X given Y. In either case, the joint density of X and Y can be represented as the product of the conditional density and the marginal density for almost every point on E and F. The ensuing equality is precisely the content of Bayes' rule; the formal statement is given below:

Theorem 2.13 (Bayes' Rule)
Let X and Y be random variables taking values in, and μ and v σ-finite measures on, the measurable spaces $(E,\mathcal{E})$ and $(F,\mathcal{F})$. Denote the joint distribution of X and Y by π, and the marginal distributions of X and Y by P_X and P_Y. Suppose that π is absolutely continuous with respect to the product measure $\mu \times v$, and that $f_{X,Y}$ is the joint density of X,Y with respect to $\mu \times v$.

Then, there exist marginal densities f_X and f_Y of X and Y with respect to μ and v, as well as conditional densities $f_{X|Y}$ of X given Y with respect to μ and $f_{Y|X}$ of Y given X with respect to v, such that

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) \cdot f_Y(y)}{f_X(x)}$$

for P_X-a.e. $x \in E$ and P_Y-a.e. $y \in F$.

Proof) As in theorem 2.12, define $f_X : E \to [0,+\infty]$ and $f_Y : F \to [0,+\infty]$ as

$$f_X(x) = \int_F f_{X,Y}(x,y)dv(y) \quad \text{and} \quad f_Y(y) = \int_E f_{X,Y}(x,y)d\mu(x)$$

for any $(x,y) \in E \times F$. We saw above that f_X and f_Y are marginal densities of X and Y with respect to μ and v.

Similarly, we can define $f_{Y|X} : E \times F \to [0,+\infty]$ and $f_{X|Y} : E \times F \to [0,+\infty]$ as

$$f_{Y|X}(y \mid x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)} & \text{if } 0 < f_X(x) < +\infty \\ \int_E f_{X,Y}(x,y)d\mu(x) & \text{if } f_X(x) = 0 \text{ or } +\infty \end{cases}$$

and

$$f_{X|Y}(x \mid y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{if } 0 < f_Y(y) < +\infty \\ \int_F f_{X,Y}(x,y)dv(y) & \text{if } f_Y(y) = 0 \text{ or } +\infty \end{cases}$$

for any $(x, y) \in E \times F$. Theorem 2.12 tells us that $f_{Y|X}$ and $f_{X|Y}$ are conditional densities of Y given X and X given Y with respect to v and μ, that is,

$$\mathbb{P}(Y \in B \mid X) = \int_B f_{Y|X}(\cdot \mid X) dv$$
$$\mathbb{P}(X \in A \mid Y) = \int_A f_{X|Y}(\cdot \mid Y) d\mu$$

for any $A \in \mathcal{E}$ and $B \in \mathcal{F}$.

In theorem 2.12, we also saw that

$$f_{X,Y}(x, y) = f_{Y|X}(y \mid x) f_X(x)$$

for any $y \in F$ and P_X-a.e. $x \in E$. By symmetry,

$$f_{X,Y}(x, y) = f_{X|Y}(x \mid y) f_Y(y)$$

for any $x \in E$ and P_Y-a.e. $y \in F$. Therefore, for P_X-a.e. $x \in E$ and P_Y-a.e. $y \in F$, we have the equality

$$f_{X,Y}(x, y) = f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y).$$

Furthermore, $\{f_X = 0\} \cup \{f_X = +\infty\}$ is a set of measure zero under P_X, so that

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) \cdot f_Y(y)}{f_X(x)}$$

also holds and is well-defined for P_X-a.e. $x \in E$ and P_Y-a.e. $y \in F$.

Q.E.D.

2.3 Construction of Sequences of Random Variables

In this section we consider the generalized version of the problem addressed in section 1.8.2. There, we wondered whether there existed a probability space $(\Omega, \mathcal{H}, \mathbb{P})$ on which a sequence $\{X_n\}_{n\in N_+}$ of independent random variables with the desired distributions could be supported. The proof that such a probability space existed relied on the decomposition of the joint distribution of independent random variables into the product of their respective distributions.

What we studied above is how to decompose joint distribution of arbitrary, not necessarily independent, random variables can be decomposed into the product of a probabiliy measure and a transition probability kernel. Thus, we can now apply this new machinery to show that there exists a probability space $(\Omega, \mathcal{H}, \mathbb{P})$ that can support a sequence $\{X_n\}_{n\in N_+}$ of random variables with any dependence structure.

We work with the following setting. Consider a sequence of measurable spaces $\{(E_t, \mathcal{E}_t)\}_{t\in\mathbb{N}}$, and let μ be a probability measure on $(E, \mathcal{E})$. For any $n \in \mathbb{N}$, define

$$(F_n^0, \mathcal{F}_n^0) = \bigotimes_{t=0}^{n} (E_t, \mathcal{E}_t),$$

and let K_{n+1} be a transition probability kernel from $(F_n^0, \mathcal{F}_n^0)$ into $(E_{n+1}, \mathcal{E}_{n+1})$. In the context of repeated trials, we can think of μ as the distribution of the initial trial, and each K_{n+1} as the conditional distribution of the $n+1$th trial given the results of all the trials that precede it. Our objective is to find a probability space $(\Omega, \mathcal{H}, \mathbb{P})$ and a sequence $\{X_t\}_{t\in\mathbb{N}}$ of random variables such that each X_t takes values in $(E_t, \mathcal{E}_t)$, X_0 has distribution μ, and K_{n+1} is the conditional distribution of X_{n+1} given $X_0 = x_0, \cdots, X_n = x_n$. This can be formally stated as the following theorem:

Theorem 2.14 (Ionescu-Tulcea's Theorem)
Let $\{(E_t, \mathcal{E}_t)\}_{t\in\mathbb{N}}$ be a sequence of measurable spaces, μ a probability measure on $(E, \mathcal{E})$, and K_{n+1} a transition probability kernel from $\bigotimes_{t=0}^{n}(E_t, \mathcal{E}_t)$ into $(E_{n+1}, \mathcal{E}_{n+1})$ for any $n \in \mathbb{N}$.

Then, there exists a probability space $(\Omega, \mathcal{H}, \mathbb{P})$ and a sequence $\{X_t\}_{t\in\mathbb{N}}$ of random variables where each X_t takes values in $(E_t, \mathcal{E}_t)$ such that

i) The distribution of X_0 is μ, and

ii) The conditional probability of $X_{t+1} \in A$ given $X_t, \cdots, X_0$ is

$$K_{t+1}(\cdot, A) \circ (X_0, \cdots, X_t)$$

for any $A \in \mathcal{E}_{t+1}$ and $t \in \mathbb{N}$.

Below we prove the theorem in steps, starting from the definition of the measurable space $(\Omega, \mathcal{H})$ and concluding with the construction of the desired $\mathbb{P}$ from a pre-measure.

2.3.1 Constructing $(\Omega, \mathcal{H})$ and the Sequence $\{X_n\}_{n\in\mathbb{N}}$

To prove the theorem above, we can think of the same construction as in the case of independent sequences of random variables. Namely, we define

$$(\Omega, \mathcal{H}) = \bigotimes_{t\in\mathbb{N}} (E_t, \mathcal{E}_t),$$

and for any $t \in \mathbb{N}$, we define

$$X_t(\omega) = \omega_t$$

for any $\omega = (\omega_t)_{t\in\mathbb{N}} \in \Omega$. This makes $\{X_t\}_{t\in\mathbb{N}}$ a sequence of random variables such that each X_t takes values in $(E_t, \mathcal{E}_t)$. To see this, choose any $A \in \mathcal{E}_t$ and note that

$$X_t^{-1}(A) = \prod_{s\in\mathbb{N}} A_s, \quad A_s = \begin{cases} E_s & \text{if } s \neq t \\ A & \text{if } s = t \end{cases} \quad \forall s \in \mathbb{N}.$$

In other words, $X_t^{-1}(A)$ is a measurable rectangle on Ω, making it an element of $\mathcal{H}$; this tells us that X_t is measurable relative to $\mathcal{H}$ and $\mathcal{E}_t$.

We also define the sequence $\{Y_t\}_{t\in\mathbb{N}}$ as

$$Y_t = (X_0, \cdots, X_t)$$

for any $t \in \mathbb{N}$; clearly, each Y_t is a random variable taking values in $(F_t^0, \mathcal{F}_t^0)$. Let $\mathcal{F}_t = \sigma Y_t$ be the σ-algebra generated by Y_t. Sets in $\mathcal{F}_t$ are called cylinders, since, for any $H \in \mathcal{F}_t$, there exists an $A \in \mathcal{F}_t^0$ such that

$$H = Y_t^{-1}(A) = A \times E_{t+1} \times E_{t+2} \times \cdots.$$

We denote $A = B_t(H)$, and call it the base of the set $H \in \mathcal{F}_t$. Define

$$\mathcal{H}^0 = \bigcup_n \mathcal{F}_n.$$

$\mathcal{H}^0$ is useful because it is an algebra on Ω that generates $\mathcal{H}$. We can directly verify the first of these statements:

- $\emptyset$ is obviously contained in $\mathcal{H}^0$ (take the inverse image of the empty set for any Y_t).
- For any $H \in \mathcal{H}^0$, letting $H \in \mathcal{F}_n$ for some $n \in \mathbb{N}$,

$$H^c = \left(B_n(H) \times \left(\prod_{t\geq n+1} E_t \right) \right)^c = B_n(H)^c \times \left(\prod_{t\geq n+1} E_t \right),$$

  where $B_n(H) \in \mathcal{F}_n^0$ and thus $B_n(H)^c \in \mathcal{F}_n^0$ as well. This shows us that $H^c \in \mathcal{F}_n \subset \mathcal{H}^0$.
- For any finite collection $\{H_1, \cdots, H_k\} \subset \mathcal{H}^0$, there must exist some $n \in \mathbb{N}$ such that $\{H_1, \cdots, H_k\} \subset$

$\mathcal{F}_n$. For each $1 \leq i \leq k$, we can write H_i as

$$H_i = B_n(H_i) \times \left(\prod_{t \geq n+1} E_t \right),$$

so we can see that

$$\bigcup_{i=1}^{k} H_i = \left(\bigcup_{i=1}^{k} B_n(H_i) \right) \times \left(\prod_{t \geq n+1} E_t \right),$$

where $\bigcup_{i=1}^{k} B_n(H_i) \in \mathcal{F}_n^0$. Therefore, $\bigcup_{i=1}^{k} H_i \in \mathcal{F}_n \subset \mathcal{H}^0$.

It is also easy to see that the algebra $\mathcal{H}^0$ generates $\mathcal{H}$. Clearly, the σ-algebra generated by $\mathcal{H}^0$ is contained in $\mathcal{H}$ because $\mathcal{H}^0$ is itself contained in $\mathcal{H}$. To see the reverse inclusion, consider a measurable rectangle $\prod_{t \in \mathbb{N}} A_t$ on Ω. Since at most finitely many A_t are not equal to E_t, let

$$n = \max\{t \in \mathbb{N} \mid A_t \neq E_t\} \in \mathbb{N}.$$

It follows that

$$\prod_{t \in \mathbb{N}} A_t \in \mathcal{F}_n \subset \mathcal{H}^0 \subset \sigma\mathcal{H}^0,$$

and because the collection of all measurable rectangles generates $\mathcal{H}$, it follows that $\mathcal{H} \subset \sigma\mathcal{H}^0$, which shows that $\mathcal{H}^0$ generates $\mathcal{H}$.

2.3.2 Applying the Hahn-Kolmogorov Extension Theorem

Now define the sequence $\{\pi_t\}_{t \in \mathbb{N}}$ of probability measures, where each π_t is defined on $(F_t^0, \mathcal{F}_t^0)$, iteratively as $\pi_0 = \mu$ and

$$\pi_t = \pi_{t-1} \times K_t$$

for any $t \in N_+$. Suppose that π_t is the distribution of Y_t for any $t \in \mathbb{N}$. In this case, for any $t \in \mathbb{N}$, because $Y_{t+1} = (Y_t, X_{t+1})$, the joint distribution of Y_t and X_{t+1} is equal to π_{t+1}, which can be decomposed into the product $\pi_t \times K_{t+1}$. In light of theorem 2.11, K_{t+1} then becomes the conditional distribution of X_{t+1} given $Y_t = y$, which is exactly the desired result. Therefore, our problem reduces to finding a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{H})$ such that the distribution of each Y_t is equal to π_t. Specifying the distribution of each Y_t in this manner is referred to as specifying the probability law of the sequence $\{X_t\}_{t \in \mathbb{N}}$.

To this end, define the function $\mathbb{P}' : \mathcal{H}^0 \to [0, 1]$ as

$$\mathbb{P}'(H) = \pi_t(B_t(H))$$

for any $H \in \mathcal{F}_t$ for some $t \in \mathbb{N}$. We must first verify that this function is well-defined by showing

that it assigns the same value to a cylinder set $H \in \mathcal{H}^0$ regardless of how we specify its base. Suppose that $H \in \mathcal{F}_n \cap \mathcal{F}_m$ for $n, m \in \mathbb{N}$, and that $n < m$ without loss of generality. Then,

$$B_m(H) = B_n(H) \times E_{n+1} \times \cdots \times E_m;$$

because $\pi_t = \pi_{t-1} \times K_t$ for any $t \in N_+$, we have

$$\begin{aligned}
\pi_m(B_m(H)) &= \pi_m(B_n(H) \times E_{n+1} \times \cdots \times E_m) \\
&= \int_{B_n(H) \times E_{n+1} \times \cdots \times E_{m-1}} K(x, E_m) d\pi_{m-1}(x) \\
&= \pi_{m-1}(B_n(H) \times E_{n+1} \times \cdots \times E_{m-1}) \\
&= \cdots = \pi_n(B_n(H))
\end{aligned}$$

by repeated applications of the definition of the product measure $\pi_{t-1} \times K_t$.

$\mathbb{P}'$ clearly assigns value 0 to the empty set, since the empty set in $\mathcal{H}$ is the inverse image of the empty set with respect to any Y_t, and is thus a cylinder set with an empty base. In addition, for any finite collection $\{H_1, \cdots, H_k\} \subset \mathcal{H}^0$ of disjoint sets, their union is contained in the algebra $\mathcal{H}^0$, and letting $\{H_1, \cdots, H_k\} \subset \mathcal{F}_n$ for some $n \in \mathbb{N}$,

$$\mathbb{P}'\left(\bigcup_{i=1}^{k} H_i\right) = \pi_n\left(\bigcup_{i=1}^{k} B_n(H_i)\right).$$

Since $B_n(H_1), \cdots, B_n(H_k)$ are disjoint sets contained in $\mathcal{F}_n$ (they are disjont because $H_1, \cdots, H_k$ are not disjoint otherwise), by the finite additivity of π_n we can now see that

$$\mathbb{P}'\left(\bigcup_{i=1}^{k} H_i\right) = \sum_{i=1}^{k} \pi_n\left(B_n(H_i)\right) = \sum_{i=1}^{k} \mathbb{P}'(H_i).$$

In other words, $\mathbb{P}'$ is a pre-measure on the algebra $\mathcal{H}^0$.

Our construction is now almost complete. Suppose we show that $\mathbb{P}'$ is also σ-additive, that is,

$$\mathbb{P}'\left(\bigcup_{n} H_n\right) = \sum_{n=1}^{\infty} \mathbb{P}'(H_n)$$

for any disjoint collection $\{H_n\}_{n \in N_+} \subset \mathcal{H}^0$ whose union is also contained in $\mathcal{H}^0$. Then, since $\mathbb{P}'$ is a σ-additive pre-measure on $\mathcal{H}^0$, by the Hahn-Kolmogorov extension theorem there exists a σ-algebra $\mathcal{M}$ on Ω and a measure $\mathbb{P}''$ on $(\Omega, \mathcal{M})$ such that $\mathcal{H}^0 \subset \mathcal{M}$ and

$$\mathbb{P}''(H) = \mathbb{P}'(H)$$

for any $H \in \mathcal{H}^0$. Since

$$\mathbb{P}'(\Omega) = \pi_0(E_0) = 1 = \mathbb{P}''(\Omega),$$

$\mathbb{P}''$ is a probability measure on $(\Omega, \mathcal{M})$. Finally, since $\mathcal{H}^0$ generates $\mathcal{H}$, $\mathcal{H} \subset \mathcal{M}$, and we define $\mathbb{P}$ as the restriction of $\mathbb{P}''$ to $\mathcal{H}$. Now it follows that

$$\left(\mathbb{P} \circ Y_t^{-1}\right)(A) = \mathbb{P}(Y_t^{-1}(A)) = \mathbb{P}'(A) = \pi_t(A)$$

for any $t \in \mathbb{N}$ and $A \in \mathcal{F}_t^0$, so that the distribution of each Y_t under $\mathbb{P}$ is exactly π_t. This shows us that $(\Omega, \mathcal{H}, \mathbb{P})$ and $\{X_t\}_{t\in\mathbb{N}}$ are exactly the probability space and sequence of random variables with the desired (conditional) distributions.

2.3.3 The σ-additivity of the Pre-Measure

The fact that $\mathbb{P}'$ is a σ-additive pre-measure on $\mathcal{H}^0$ can be seen as follows. First, for any $H \in \mathcal{F}_n$, we define the sequence $\{Q_m(\cdot, H)\}_{m\in\mathbb{N}}$ of measurable functions $Q_m(\cdot, H) : F_m^0 \to [0,1]$ inductively as follows. For any $m \geq n$, we define

$$Q_m((x_0, \cdots, x_m), H) = I_{B_n(H)}(x_0, \cdots, x_n)$$

for any $(x_0, \cdots, x_m) \in F_m^0$. $Q_m(\cdot, H)$ is an indicator function for the $\mathcal{F}_m^0$-measurable set $B_n(H) \times E_{n+1} \times \cdots \times E_m$, so it follows that it is $\mathcal{F}_m^0$-measurable. Now, suppose that for some $0 < m \leq n$, we have shown that $Q_m(\cdot, H)$ is an $\mathcal{F}_m^0$-measurable function taking values in $[0,1]$. Then, we define $Q_{m-1}(\cdot, H) : F_{m-1}^0 \to [0,1]$ as

$$Q_{m-1}((x_0, \cdots, x_{m-1}), H) = \int_{E_m} Q_m((x_0, \cdots, x_m), H) K_m((x_0, \cdots, x_{m-1}), dx_m)$$

for any $(x_0, \cdots, x_{m-1}) \in F_{m-1}^0$. Then, integral on the right hand side is well-defined because $Q_m(\cdot, H) : F_{m-1}^0 \times E_m \to [0,1]$ is $\mathcal{F}_m^0 = \mathcal{F}_{m-1}^0 \bigotimes \mathcal{E}_m$-measurable. Furthermore, in light of lemma 2.9, we can write

$$Q_{m-1}(\cdot, H) = T_{K_m} Q_m(\cdot, H),$$

since K_m is a transition probability kernel from $(F_{m-1}^0, \mathcal{F}_{m-1}^0)$ into $(E_m, \mathcal{E}_m)$, and as such $Q_{m-1}(\cdot, H)$ is a non-negative $\mathcal{F}_{m-1}^0$-measurable function. In addition, since $Q_m(\cdot, H)$ is bounded above by 1 and K_m is a transition probability kernel, $Q_{m-1}(\cdot, H)$ is also bounded above by 1.

Having constructed this sequence $\{Q_m(\cdot, H)\}_{m\in\mathbb{N}}$ of measurable functions mapping into $[0,1]$, we can now express $\mathbb{P}'(H)$ in terms of the function $Q_0(\cdot, H)$. This follows by noting that

$$\begin{aligned}
\mathbb{P}'(H) &= \pi_n(B_n(H)) \\
&= \int_{E_0} \int_{E_1} \cdots \underbrace{\int_{E_n} \underbrace{I_{B_n(H)}(x_0, \cdots, x_n)}_{Q_n((x_0,\cdots,x_n),H)} K_n((x_0, \cdots, x_{n-1}), dx_n) \cdots K_1((x_0, x_1), dx_1) d\mu(x_0)}_{Q_{n-1}((x_0,\cdots,x_{n-1},H)}
\end{aligned}$$

$$= \int_{E_0} \int_{E_1} \cdots \underbrace{\int_{E_{n-1}} Q_{n-1}((x_0, \cdots, x_{n-1}), H) K_{n-1}((x_0, \cdots, x_{n-2}), dx_{n-1})}_{Q_{n-2}((x_0, \cdots, x_{n-2}), H)} \cdots K_1((x_0, x_1), dx_1) d\mu(x_0)$$

$$= \cdots = \int_{E_0} \underbrace{\int_{E_1} Q_1((x_0, x_1), H) K_1((x_0, x_1), dx_1)}_{Q_0(x_0, H)} d\mu(x_0)$$

$$= \int_{E_0} Q_0(\cdot, H) d\mu.$$

We can also see that, for any $H_1, H_2 \in \mathcal{H}^0$ such that $H_1 \subset H_2$, we have $Q_m(\cdot, H_1) \leq Q_m(\cdot, H_2)$ for any $m \in \mathbb{N}$. This can again be seen inductively. Letting $H_1, H_2 \in \mathcal{F}_n$ for some $n \in \mathbb{N}$, since $B_n(H_1) \subset B_n(H_2)$, for any $m \geq n$ we have

$$Q_m((x_0, \cdots, x_m), H_1) = I_{B_n(H_1)}(x_0, \cdots, x_n) \leq I_{B_n(H_2)}(x_0, \cdots, x_n) = Q_m((x_0, \cdots, x_m), H_2).$$

for any $(x_0, \cdots, x_m) \in F_m^0$. Now suppose that $Q_m(\cdot, H_1) \leq Q_m(\cdot, H_2)$ for some $0 < m \leq n$. Then,

$$Q_{m-1}(\cdot, H_1) = T_{K_m} Q_m(\cdot, H_1) \leq T_{K_m} Q_m(\cdot, H_2) = Q_{m-1}(\cdot, H_2)$$

by the monotonicity of the operator $T_{K_m} : \left(\mathcal{F}_{m-1}^0 \bigotimes \mathcal{E}_m\right)_{+,b} \to \left(\mathcal{F}_{m-1}^0\right)_{+,b}$.

We now return to the original problem of showing the σ-additivity of $\mathbb{P}'$. Choose any sequence $\{H_k\}_{k \in \mathbb{N}}$ such that $H_k \searrow \emptyset$. Let us show that $\{\mathbb{P}'(H_k)\}_{k \in \mathbb{N}}$ decreases to 0. For the sake of contradiction, suppose

$$\lim_{k \to \infty} \mathbb{P}'(H_k) = \inf_{k \in \mathbb{N}} \mathbb{P}'(H_k) > 0.$$

Then, because $\{Q_0(\cdot, H_k)\}_{k \in \mathbb{N}}$ is a decreasing sequence of bounded $\mathcal{E}_0$-measurable functions (and thus has a pointwise limit equal to its infimum) and μ is a finite measure, by the BCT we have

$$0 < \lim_{k \to \infty} \mathbb{P}'(H_k) = \lim_{k \to \infty} \int_{E_0} Q_0(\cdot, H_k) d\mu = \int_{E_0} \left(\inf_{k \in \mathbb{N}} Q_0(\cdot, H_k) \right) d\mu.$$

If $\inf_{k \in \mathbb{N}} Q_0(x_0, H_k) = 0$ for every $x_0 \in E_0$, then the integral on the right will be equal to 0, a contradiction. It follows that there must exist some $x_0^* \in E_0$ such that

$$\inf_{k \in \mathbb{N}} Q_0(x_0^*, H_k) > 0.$$

Suppose that we have found $(x_0^*, \cdots, x_m^*) \in F_m^0$ such that

$$\inf_{k \in \mathbb{N}} Q_m((x_0^*, \cdots, x_m^*), H_k) > 0$$

for some $m \geq 0$. Then, since the sequence of mappings $x_{m+1} \mapsto Q_{m+1}((x_0^*, \cdots, x_m^*, x_{m+1}), H_k)$ is a bounded and decreasing sequence, and the measure $K_{m+1}((x_0^*, \cdots, x_m^*), \cdot)$ is finite, by the BCT

again we have

$$\begin{aligned}0 < \inf_{k\in\mathbb{N}} Q_m((x_0^*,\cdots,x_m^*),H_k) &= \lim_{k\to\infty} Q_m((x_0^*,\cdots,x_m^*),H_k)\\ &= \lim_{k\to\infty}\int_{E_{m+1}} Q_{m+1}((x_0^*,\cdots,x_m^*,x_{m+1}),H_k)K_{m+1}((x_0^*,\cdots,x_m^*),dx_{m+1})\\ &= \int_{E_{m+1}}\left(\inf_{k\in\mathbb{N}} Q_{m+1}((x_0^*,\cdots,x_m^*,x_{m+1}),H_k)\right)K_{m+1}((x_0^*,\cdots,x_m^*),dx_{m+1}).\end{aligned}$$

Thus, by the same reasoning as above, there must exist an $x_{m+1}^* \in E_{m+1}$ such that

$$\inf_{k\in\mathbb{N}} Q_{m+1}((x_0^*,\cdots,x_m^*,x_{m+1}^*),H_k) > 0.$$

We have inductively constructed a point $x^* = (x_t^*)_{t\in\mathbb{N}} \in \Omega = \prod_{t\in\mathbb{N}} E_t$ such that

$$\inf_{k\in\mathbb{N}} Q_m((x_0^*,\cdots,x_m^*),H_k) > 0$$

for any $m \in \mathbb{N}$. For any $k \in N_+$, let $H_k \in \mathcal{F}_n$; then,

$$Q_n((x_0^*,\cdots,x_n^*),H_k) = I_{B_n(H_k)}(x_0^*,\cdots,x_n^*) = I_{H_k}(x^*) > 0,$$

so it follows that

$$x^* \in H_k.$$

This holds for any $k \in N_+$, so

$$x^* \in \bigcap_k H_k.$$

However, this is a contradiction because $\{H_k\}_{k\in N_+}$ decreases to the empty set; therefore, it must be the case that $\lim_{k\to\infty}\mathbb{P}'(H_k) = 0$.

Let $\{A_k\}_{k\in N_+}$ be a collection of disjoint sets in $\mathcal{H}^0$ such that $A = \bigcup_k A_k \in \mathcal{H}^0$. Define

$$H_k = A \setminus \left(\bigcup_{i=1}^k A_i\right)$$

for any $k \in N_+$. Then, each H_k is in $\mathcal{H}^0$, since $\mathcal{H}^0$, being an algebra of sets, is closed under finite unions, intersections and complements, and $\{H_k\}_{k\in N_+}$ decreases to the empty set. By what we showed above,

$$\lim_{k\to\infty}\mathbb{P}'(H_k) = 0.$$

By the finite additivity of $\mathbb{P}'$, we can see that

$$\mathbb{P}'(A) = \mathbb{P}'\left(\left(\bigcup_{i=1}^k A_i\right)\cup H_k\right) = \mathbb{P}'\left(\bigcup_{i=1}^k A_i\right) + \mathbb{P}'(H_k)$$

$$= \sum_{i=1}^{k} \mathbb{P}'(A_i) + \mathbb{P}'(H_k).$$

Then, taking $k \to \infty$ on both sides yields

$$\mathbb{P}'(A) = \sum_{k=1}^{\infty} \mathbb{P}'(A_k),$$

which shows us that $\mathbb{P}'$ is σ-additive. This completes the proof of Ionescu-Tulcea's theorem.

Chapter 3

Convergence of Random Variables

In this chapter we will be discusisng three main ways in which random variables taking values in some metric space can converge. The first, almost sure convergence, is essentially an almost everywhere version of the pointwise convergence of measurable functions, and is the strongest form of convergence we will study. Convergence in probability and in L^p, which will be discussed next, are slightly weaker versions of convergence, but are often used due to their tractability and ubiquity (for instance, compare the difficulty of proving the WLLN against that of the SLLN).

3.1 Random Variables in Metric Spaces

From now on, we will mostly concern ourselves with random variables taking values in a metric space, instead of an abstract measurable space. This is because the introduction of a metric allows us to explicitly measure the distance between random variables and thus formulate limit theorems in much more easily than, say, if we assume the random variables of interest take values in a Hausdorff space, in which convergence is defined in an abstract manner.
Our formal framework for analysis is given as follows. Let $(\Omega, \mathcal{H}, \mathbb{P})$ be the underlying probability space, (E, d) a metric space, τ the metric topology on E induced by the metric d, and $\mathcal{E}$ the Borel σ-algebra on E generated by τ. The following are the two most important properties of metric spaces:

- **Completeness**
  (E, d) is said to be complete if any Cauchy sequence $\{x_n\}_{n \in N_+}$ in E is also convergent, that is,

  $$\lim_{m,n \to \infty} d(x_n, x_m) = 0 \quad \text{implies} \quad \lim_{n \to \infty} d(x_n, x) = 0$$

  for some $x \in E$.

- **Separability**
  (E, d) is said to be a separable metric space if the induced topological space (E, τ) is separable; recall that (E, τ) is separable if there exists a countable subset E_0 of E that is dense

in E, that is, $\overline{E_0} = E$.

Recall also that (E, τ) is said to be second countable if there exists a countable base $\mathbb{B} \subset \tau$ on E that generates the topology τ. We proved in the measure theory text that an arbitrary topological space is separable if it is second countable, while the converse also holds if it is also metrizable (as in the case of (E, τ)). Since we are dealing exclusively with metric spaces, for our purposes the properties of second countability and separability are equivalent.

Note that any euclidean space is separable. Specifically, the euclidean k-space $\mathbb{R}^k$ is the closure of the countable subset $\mathbb{Q}^k$, which is the collection of all points in $\mathbb{R}^k$ with rational coordinates.

Below we study in detail some important implications of separability for metric spaces. First, we define and study the product of metrics and how they relate to product topologies.

3.1.1 Product Metrics

It is also sometimes necessary to talk of the product of two metrics. Let (E, d) and (F, ρ) be two metric spaces, and let their metric topologies be τ and s, respectively. The product metric $d \times \rho$ on $E \times F$ is defined as

$$(d \times \rho)(x, y) = \max(d(x_1, y_1), \rho(x_2, y_2))$$

for any $x, y \in E \times F$. In our measure theory text, we saw that the product of k metrics on $\mathbb{R}$ generates the same topology as the euclidean metric on $\mathbb{R}^k$, which is exactly the product topology on $\mathbb{R}^k$. That $d \times \rho$ defines an actual metric on $E \times F$ is evident from the individual properties of d and ρ:

- $d \times \rho$ takes values in $[0, +\infty)$ because d and ρ do.
- For any $x, y \in E \times F$, because of the symmetry of d and ρ, we have

$$\begin{aligned}(d \times \rho)(x, y) &= \max(d(x_1, y_1), \rho(x_2, y_2)) \\ &= \max(d(y_1, x_1), \rho(y_2, x_2)) = (d \times \rho)(y, x).\end{aligned}$$

- For any $x, y \in E \times F$, $(d \times \rho)(x, y) = 0$ implies that $d(x_1, y_1) = \rho(x_2, y_2) = 0$, so that $x = y$. Conversely, if $x = y$, then clearly $(d \times \rho)(x, y) = 0$. Thus, $(d \times \rho)(x, y) = 0$ if and only if $x = y$.
- For any $x, y, z \in E \times F$, suppose without loss of generality that $(d \times \rho)(x, y) = d(x_1, y_1)$.

Then,

$$\begin{aligned}(d\times\rho)(x,y) = d(x_1,y_1) &\le d(x_1,z_1)+d(z_1,y_1)\\ &\le (d\times\rho)(x,z)+(d\times\rho)(z,y)\end{aligned}$$

by the triangle inequality of d; the same applies when the maximum is $\rho(x_2,y_2)$ rather than $d(x_1,y_1)$.

The important part is that the product metric $d\times\rho$ metrizes the product topology $\tau\times s$. This means that any function on $E\times F$ that is continuous with respect to the product metric $d\times\rho$ can also be considered a continuous function on the topological space $(E\times F,\tau\times s)$. This greatly simplifies proofs that require us to show that some function on $E\times F$ is continuous; indeed, this property is used to show the measurability result for separable metric spaces stated above.

Lemma 3.1 Let (E,d) and (F,ρ) be metric spaces with metric topologies τ and s. Then, the product metric $d\times\rho$ metrizes the product topology $\tau\times s$.

Proof) Let $\mathbb{B}$ be the collection of all open rectangles on $E\times F$; by definition, $\mathbb{B}$ is a base on $E\times F$ that generates the product topology $\tau\times s$. On the other hand, let $\mathbb{D}$ be the collection of all sets on $E\times F$ of the form

$$B_{d\times\rho}(x,\epsilon) = \{y\in E\times F \mid (d\times\rho)(x,y)<\epsilon\}$$

for any $x\in E\times F$ and $\epsilon>0$. Our goal is to show that $\mathbb{D}$ generates the product topology $\tau\times s$. We proceed in steps:

Step 1: $\mathbb{D}$ is a collection of open sets in $\tau\times s$

Choose any $x\in E\times F$, $\epsilon>0$, and a point y in $B_{d\times\rho}(x,\epsilon)$. Then,

$$d(x_1,y_1),\rho(x_2,y_2)\le(d\times\rho)(x,y)<\epsilon,$$

so that, defining $\delta_1=\epsilon-d(x_1,y_1)>0$ and $\delta_2=\epsilon-\rho(x_2,y_2)>0$,

$$y\in B_d(y_1,\delta_1)\times B_\rho(y_2,\delta_2)\subset B_{d\times\rho}(x,\epsilon).$$

The first inclusion is clear, while the second inclusion follows from the observation that, if $z\in B_d(y_1,\delta_1)\times B_\rho(y_2,\delta_2)$, then $d(z_1,y_1)<\delta_1$ and $\rho(z_2,y_2)<\delta_2$, so that

$$d(x_1,z_1)\le d(x_1,y_1)+d(y_1,z_1)<d(x_1,y_1)+\delta_1=\epsilon$$

and likewise, $\rho(x_2,z_2)<\epsilon$, which together imply that

$$(d\times\rho)(x,z)=\max(d(x_1,z_1),\rho(x_2,z_2))<\epsilon,$$

or that $z \in B_{d\times\rho}(x,\epsilon)$.

This holds for any $y \in B_{d\times\rho}(x,\epsilon)$, so labeling $A_y = B_d(y_1,\delta_1)\times B_\rho(y_2,\delta_2) \in \mathbb{B}$,

$$B_{d\times\rho}(x,\epsilon) = \bigcup_{y\in B_{d\times\rho}(x,\epsilon)} A_y.$$

Each A_y is open with respect to $\tau\times s$, and the arbitrary union of open sets is open, so $B_{d\times\rho}(x,\epsilon)$ is open with respect to $\tau\times s$.

Step 2: $\mathbb{D}$ is a base on $E\times F$ that generates $\tau\times s$

$\mathbb{D}$ clearly covers $E\times F$: fixing any $x\in E\times F$, for any $y\in E\times F$, since $(d\times\rho)(x,y) = M < +\infty$, it follows that

$$y \in B_{d\times\rho}(x, M+1) \subset \bigcup_{A\in\mathbb{D}} A.$$

Choose any $A\in\tau\times s$ and an $x\in A$. Since $\mathbb{B}$ generates $\tau\times s$, there exists an open rectangle $B_1\times B_2\in\mathbb{B}$ such that

$$x \in B_1\times B_2 \subset A.$$

Since B_1 and B_2 are open sets in the metric spaces (E,d) and (F,ρ), there also exist $\epsilon>0$, $\delta>0$ such that

$$x_1 \in B_d(x_1,\epsilon)\subset B_1 \quad \text{and} \quad x_2\in B_\rho(x_2,\delta)\subset B_2.$$

Letting $\eta = \min(\epsilon,\delta) > 0$, it follows that

$$\begin{aligned} x \in B_{d\times\rho}(x,\eta) &\subset B_d(x_1,\epsilon)\times B_\rho(x_2,\delta) \\ &\subset B_1\times B_2 \subset A. \end{aligned}$$

This holds for any open set $A\in\tau\times s$ and an element $x\in A$, so by definition $\mathbb{D}$ is a base on $E\times F$ (this can be shown by choosing A as the intersection of two elements of $\mathbb{D}$) that generates $\tau\times s$.

Q.E.D.

3.1.2 Separable Metric Spaces

In the context of the convergence of random variables, separability plays an important role because it ensures that, for any two random variables X, Y taking values in (E,d), the composition $d \circ (X,Y)$ is a non-negative real valued random variable, the measurability of which proves integral to the definition of convergence in probability.

Theorem 3.2 Let (E,d) be a separable metric space with Borel σ-algebra $\mathcal{E}$ generated by the metric topology induced by d, and X, Y random variables that take values in $(E,\mathcal{E})$. Then, $d(X,Y)$ is a non-negative real valued random variable.

Proof) Since (E,d) is a separable metric space, the induced topological space (E,τ) is second countable. It was shown in the measure theory text that, in this case, the Borel σ-algebra generated by the product topology τ^2 is equivalent to the product of two Borel σ-algebras $\mathcal{E}$;

$$\mathcal{B}(E^2,\tau^2) = \mathcal{B}(E,\tau)^2 = \mathcal{E}^2.$$

We first show that the metric $d : E^2 \to [0,+\infty)$ is continuous on E^2 with respect to the product metric $d \times d$. To see this, note that, for any $(x,y),(z,w) \in E^2$, we have

$$\begin{aligned} |d(x,y) - d(z,w)| &\leq |d(x,y) - d(z,y)| + |d(z,y) - d(z,w)| \\ &\leq d(x,z) + d(y,w) \leq 2 \cdot \max(d(x,z), d(y,w)) = 2 \cdot (d \times d)((x,y),(z,w)), \end{aligned}$$

where the second inequality follows from the triangle inequality. Therefore, d is uniformly continuous on E^2 under the metric $d \times d$. Since the product topology τ^2 is metrizable by $d \times d$, it follows that $d^{-1}(A) \in \tau^2$ for any Borel set $A \subset \mathbb{R}$, and that d is thus measurable relative to $\mathcal{B}(E^2,\tau^2) = \mathcal{E}^2$ and $\mathcal{B}(\mathbb{R})$.
Next, we must show that the mapping $\omega \mapsto (X(\omega), Y(\omega))$ is measurable relative to $\mathcal{H}$ and $\mathcal{E}^2$. However, this follows immediately, since X and Y are random variables taking values in $\mathcal{E}^2$. Therefore, by the preservation of measurability across compositions,

$$d(X,Y) = d \circ (X,Y)$$

must be measurable relative to $\mathcal{H}$ and $\mathcal{B}(\mathbb{R})$. By definition, $d(X,Y)$ is now a real random variable.
Q.E.D.

The following theorem is an application of the above result. Specifically, it furnishes sufficient conditions for pointwise convergence to be uniform.

Theorem 3.3 (Egorov's Theorem)
Let (E,d) be a separable metric space, τ the metric topology induced by d and $\mathcal{E}$ the Borel σ-algebra on E generated by τ. Suppose $\{X_n\}_{n\in N_+}$ is a sequence of random variables on $(E,\mathcal{E})$ that converges almost surely to the random variable X.
Then, for any $\epsilon > 0$ there exists a measurable set $\Omega_0 \in \mathcal{H}$ such that $\mathbb{P}(\Omega_0) < \epsilon$ and $\{X_n\}_{n\in N_+}$ converges to X uniformly on $\Omega \setminus \Omega_0$.

Proof) Let $X_n \to X$ pointwise on a set $H \in \mathcal{H}$ such that $\mathbb{P}(H) = 1$. Choose any $\epsilon > 0$. For any $n, k \in N_+$, we define the sets

$$A_{n,k} = \bigcup_{m\geq n} \Big\{ d(X_m, X) \geq \frac{1}{k} \Big\}.$$

Each $A_{n,k}$ is $\mathcal{H}$-measurable because the separability of the metric space (E,d) implies that $d(X_m, X)$ is a random variable for any $m \in N_+$.
Note that $A_{n+1,k} \subset A_{n,k}$, and that, if $\omega \in H$, then $\omega \notin \bigcap_n A_{n,k}$ because the fact that $X_m(\omega) \to X(\omega)$ as $m \to \infty$ means that the distance between $X_m(\omega)$ and $X(\omega)$ will eventually become smaller than $\frac{1}{k}$ for large m. Therefore, $\bigcap_n A_{n,k} \subset H^c$, which implies that

$$\mathbb{P}\left(\bigcap_n A_{n,k} \right) = 0.$$

Because $\mathbb{P}$ is a probability measure and $\{A_{n,k}\}_{n\in N_+}$ is a decreasing sequence of measurable sets, by sequential continuity

$$\lim_{n\to\infty} \mathbb{P}(A_{n,k}) = \mathbb{P}\left(\bigcap_n A_{n,k} \right) = 0,$$

which implies that there exists an $N_k \in N_+$ such that

$$\mathbb{P}(A_{n,k}) \leq \frac{\epsilon}{2^{k+1}}$$

for any $n \geq N_k$.

Now define $\Omega_0 = \bigcup_k A_{N_k,k}$; in effect, Ω_0 collects all the outcomes such that the distance between X_n and X does not decrease below some small positive threshold for any large $n \in N_+$. It stands to reason that the sequence $\{X_n\}_{n\in N_+}$ converges uniformly to X outside of Ω_0, which is exactly what will be shown below.
For any $\delta > 0$, let $k \in N_+$ satisfy $\frac{1}{k} < \delta$. Then, for any $n \geq k$,

$$d(X_n(\omega), X(\omega)) < \frac{1}{k} < \delta$$

for any $\omega \in \Omega \setminus \Omega_0$, since $\omega \in A^c_{N_k,k}$ in this case. This holds for any $\delta > 0$, so by definition $\{X_n\}_{n\in N_+}$ converges to X uniformly on $\Omega \setminus \Omega_0$.

Finally, note that

$$\mathbb{P}(\Omega_0) \leq \sum_{k=1}^{\infty} \mathbb{P}(A_{N_k,k}) \leq \frac{\epsilon}{2} < \epsilon,$$

which completes the proof.
Q.E.D.

3.1.3 The Distance Function and Urysohn's Lemma

We can also think of a generalization of a metric that yields the distance from a point to a set, which is called the distance function. Let (E,d) be a metric space and A a nonempty subset of E. Then, the distance from a point $x \in E$ to A is defined as

$$d(x,A) = \inf\{d(x,y) \mid y \in A\}.$$

The infimum on the right exists by the least upper bound property of the real line and the fact that the set $\{d(x,y) \mid y \in A\}$ is bounded below at 0. The properties of the distance function, as well as the strong version of Urysohn's lemma that makes use of the distance function, were studied in the measure theory text. Here we present only the important results:

Lemma 3.3 Let (E,d) be a metric space and A a non-empty subset of E. Then, the following hold true:

i) $d(x,A) = 0$ if and only if $x \in \overline{A}$.

ii) The distance function $d(\cdot,A)$ is continuous on E.

Theorem 3.4 (Urysohn's Lemma for Metric Spaces)
Let (E,d) be a metric space. For any closed set F and open set V such that $F \subset V$, therre exists a continuous function $f : E \to \mathbb{R}$ such that

$$f(x) \in \begin{cases} \{0\} & \text{if } x \notin V \\ [0,1] & \text{if } x \in V \setminus F \\ \{1\} & \text{if } x \in F \end{cases}$$

for any $x \in E$.

In addition, if the distance between F and V^c is bounded below by a non-zero value, that is, if there exists a $\delta > 0$ such that $d(x,y) \geq \delta$ for any $x \in F$ and $y \in V^c$, then we can choose f to be Lipschitz continuous.

3.1.4 Compactness

The relationship between various forms of compactness will often prove relevant when dealing with metric spaces. Let (E, τ) be a topological space and A a subset of E. A is

- **Limit Point Compact**
  If any infinite set in A has a limit point in A
- **Countably Compact**
  If any countable open cover of A has a finite subcover
- **Sequentially Compact**
  If any sequence in A has a convergent subsequence with limit in A
- **Relatively Compact**
  If the closure $\overline{A}$ of A is compact.

Clearly, a compact subset of a Hausdorff space is relatively compact (because it is closed and thus is equivalent to its closure) and any compact set is countably compact.

A related concept is that of a **Lindelöf Space**. A topological space (E, τ) is Lindelöf if any open cover of E has a countable subcover. Note how any compact space is also Lindelöf. We state below some relevant results, which are proven in the measure theory text.

Theorem 3.5 The following hold true:

i) Compact sets are limit point compact.

ii) Limit point compact sets in a metric space are sequentially compact.

iii) Metrizable sequentially compact spaces are second countable.

iv) Second countable topological spaces are Lindelöf.

v) Sequentially compact sets are countably compact.

vi) A subset of a metric space is compact if and only if it is sequentially compact.

Corollary to Theorem 3.5 (The Bolzano-Weierstrass Theorem)
Let $F = \mathbb{R}^n$ or $\mathbb{C}$. Then, any bounded sequence $\{x_n\}_{n \in N_+}$ in F contains a convergent subsequence.

Theorem 3.6 Let (E, d) be a metric space, and A a subset of E. Then, A is relatively compact if and only if any sequence in A has a convergent subsequence.

3.2 Almost Sure Convergence

Let (E,d) be a separable metric space, τ the topology induced by the metric d, and $\mathcal{E}=\mathcal{B}(\tau)$ the Borel σ-algebra on E generated by τ. For any sequence $\{X_n\}_{n\in N_+}$ of random variables taking values in E, we say that $\{X_n\}_{n\in N_+}$ converges almost surely to some random variable X taking values in E if there exists a set $\Omega_0\in\mathcal{H}$ such that $\mathbb{P}(\Omega_0)=1$ and

$$\lim_{n\to\infty} d(X_n(\omega),X(\omega))=0$$

for any $\omega\in\Omega_0$. We denote this mode of convergence succinctly as

$$X_n \overset{a.s.}{\to} X.$$

Because almost sure convergence is basically pointwise convergence, it is one of the most powerful forms of convergence that we deal with in probability theory. Being so strong, however, it only holds in rare cases. Nevertheless, there are some relevant results converning the almost sure convergence of random variables.

From the definition, we can obtain the following are useful characterizations of almost sure convergence:

Theorem 3.7 Let (E,d) be a separable metric space, and define the Borel σ-algebra $\mathcal{E}$ as above. For any sequence $\{X_n\}_{n\in N_+}$ of random variables and a random variable X taking values in E, the following are equivalent:

i) $\{X_n\}_{n\in N_+}$ converges almost surely to X.

ii) $\limsup_{n\to\infty} d(X_n,X)=0$ almost surely.

iii) For any $\epsilon>0$,

$$\sum_{n=1}^{\infty} I_{\{d(X_n,X)>\epsilon\}} < +\infty$$

almost surely.

Proof) Note that the set

$$\Omega_0=\{\limsup_{n\to\infty} d(X_n,X)=0\}$$

is $\mathcal{H}$-measurable because $\{d(X_n,X)\}_{n\in N_+}$ is a sequence of real valued random variables (due to the separability of the underlying metric space) and the limit superior of a sequence of real $\mathcal{H}$-measurable functions $d(X_n,X)$ is also $\mathcal{H}$-measurable. The equivalence of i) and ii) follows because $\{d(X_n,X)\}_{n\in N_+}$ is bounded below by 0 for any outcome.

To see the equivalence of i) and iii), for any $\epsilon > 0$ let Ω_ϵ be the set of outcomes for which

$$\sum_{n=1}^{\infty} I_{\{d(X_n,X)>\epsilon\}} < +\infty;$$

again, the set Ω_ϵ is $\mathcal{H}$-measurable because

$$\sum_{n=1}^{\infty} I_{\{d(X_n,X)>\epsilon\}} = \sum_{n=1}^{\infty} I_{(\epsilon,+\infty)} \circ d(X_n, X)$$

is the limit of the sum of $\mathcal{H}$-measurable non-negative functions.

Suppose that $X_n \overset{a.s.}{\to} X$. Then, there exists an almost sure set $\Omega_0 \in \mathcal{H}$ such that $X_n(\omega) \to X(\omega)$ in the metric d for any $\omega \in \Omega_0$. Choose any $\omega \in \Omega_0$ and $\epsilon > 0$. By definition, there exists an $N \in N_+$ such that

$$d(X_n(\omega), X(\omega)) \leq \epsilon$$

for any $n \geq N$, so that

$$\sum_{n=1}^{\infty} I_{\{d(X_n(\omega),X(\omega))>\epsilon\}} \leq N < +\infty.$$

This holds for any $\omega \in \Omega_0$, so it follows that

$$\Omega_0 \subset \Omega_\epsilon$$

and thus $\mathbb{P}(\Omega_\epsilon) = 1$. This in turn holds for any $\epsilon > 0$, so we have proven that i) -¿ iii).

Conversely, suppose that $\mathbb{P}(\Omega_\epsilon) = 1$ for any $\epsilon > 0$. Define

$$\Omega_0 = \bigcap_n \Omega_{1/n} \in \mathcal{H};$$

since each $\Omega_{1/n}$ is of probability 1,

$$\mathbb{P}(\Omega_0^c) = \mathbb{P}\left(\bigcup_n \Omega_{1/n}^c\right) \leq \sum_{n=1}^{\infty} \mathbb{P}\left(\Omega_{1/n}^c\right) = 0$$

and Ω_0 is also an almost sure set. For any $\omega \in \Omega_0$ and $\epsilon > 0$, choose $k \in N_+$ such that $\frac{1}{k} < \epsilon$. Since $\omega \in \Omega_{1/k}$, we can see that

$$\sum_{n=1}^{\infty} I_{\{d(X_n(\omega),X(\omega))>1/k\}} < +\infty.$$

This tells us that there exists an $N \in N_+$ such that

$$d(X_n(\omega), X(\omega)) \leq \frac{1}{k} < \epsilon$$

for any $n \geq N$, since otherwise, the above series would be the sum of an infinite number of 1s and thus diverge to $+\infty$. This holds for any $\epsilon > 0$, so the sequence $\{X_n(\omega)\}_{n \in N_+}$ converges in the metric d to $X(\omega)$. This in turn holds for any ω in the almost sure set Ω_0, so $X_n \overset{a.s.}{\to} X$.

Q.E.D.

Closely related to almost sure convergence is the Borel-Cantelli lemma, which deals with the probability for events to occur infinitely often. Given a sequence of events $\{H_n\}_{n \in N_+} \subset \mathcal{H}$, we say that an outcome $\omega \in \Omega$ occurs infinitely often in the sequence $\{H_n\}$ if it is contained in the set

$$\limsup_{n \to \infty} H_n = \bigcap_n \bigcup_{k=n}^{\infty} H_k \in \mathcal{H}.$$

Heuristically, an outcome in the set $\limsup_{n \to \infty} H_n$ appears in the sequence $\{H_n\}$ no matter how many of the events $H_1, H_2, \cdots$ have already come to pass; this is the intuition behind the term "infinitely often". We also write $\{H_n \; i.o.\}$ for $\limsup_{n \to \infty} H_n$.

Below is the Borel-Cantelli lemma, which relates the sum of the probabilities of a sequence of events to the probability of infinitely many of them occuring.

Lemma 3.8 (Borel-Cantelli Lemma)
For any sequence $\{H_n\}_{n \in N_+} \subset \mathcal{H}$ of events, if

$$\sum_{n=1}^{\infty} \mathbb{P}(H_n) < +\infty,$$

then

$$\sum_{n=1}^{\infty} I_{H_n} < +\infty$$

almost surely and

$$\mathbb{P}\left(\limsup_{n \to \infty} H_n\right) = 0.$$

Proof) Define the non-negative real random variable X as

$$X = \sum_{n=1}^{\infty} I_{H_n}.$$

Then, by the MCT for series,

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{E}[I_{H_n}] = \sum_{n=1}^{\infty} \mathbb{P}(H_n) < +\infty.$$

The finiteness property for non-negative functions now implies that

$$\mathbb{P}(X < +\infty) = 1,$$

that is, X is almost surely finite.

Choose any

$$\omega \in \limsup_{n\to\infty} H_n = \bigcap_n \bigcup_{k=n}^{\infty} H_k.$$

We can immediately see that there exists an $n_1 \in N_+$ such that $\omega \in H_{n_1}$. Suppose that we have found $n_1 < \cdots < n_k$ for some $k \in N_+$. Then, because

$$\omega \in \bigcup_{m=n_k+1}^{\infty} H_m,$$

there exists an $n_{k+1} > n_k$ such that $\omega \in H_{n_{k+1}}$. Constructing the subsequence $\{H_{n_k}\}_{k\in N_+}$ in this manner, we can see that ω is contained in every single one of the sets in H_{n_k}; therefore,

$$+\infty = \sum_{k=1}^{\infty} I_{H_{n_k}}(\omega) \leq X(\omega)$$

and $\omega \in \{X = +\infty\}$. Therefore,

$$\limsup_{n\to\infty} H_n \subset \{X = +\infty\}$$

and the monotonicity of measures now shows us that

$$\mathbb{P}\left(\limsup_{n\to\infty} H_n\right) = 0.$$

Q.E.D.

The Borel-Cantelli lemma can be used to furnish sufficient conditions for almost sure convergence in terms of probabilities:

Theorem 3.9 Let (E, d) be a separable metric space, and define the Borel σ-algebra $\mathcal{E}$ as above. For any sequence $\{X_n\}_{n\in N_+}$ of random variables and a random variable X taking values in E, the following hold true:

i) If, for any $\epsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(d(X_n, X) > \epsilon\right) < +\infty,$$

then $X_n \overset{a.s.}{\to} X$.

ii) If there exists a sequence $\{\epsilon_n\}_{n\in N_+}$ of positive scalars converging to 0 such that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(d(X_n, X) > \epsilon_n\right) < +\infty,$$

then $X_n \overset{a.s.}{\to} X$.

iii) If there exists a sequence $\{\epsilon_n\}_{n\in N_+}$ of positive scalars such that

$$\sum_{n=1}^{\infty} \epsilon_n < +\infty \quad \text{and} \quad \sum_{n=1}^{\infty} \mathbb{P}\left(d(X_n, X) > \epsilon_n\right) < +\infty,$$

then $X_n \overset{a.s.}{\to} X$.

Proof) Suppose that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(d(X_n, X) > \epsilon\right) < +\infty,$$

for any $\epsilon > 0$. The Borel-Cantelli lemma tells us that this implies

$$\sum_{n=1}^{\infty} I_{\{d(X_n,X)>\epsilon\}} < +\infty$$

almost surely for any $\epsilon > 0$. The characterization in theorem 3.7 now allows us to conclude that $X_n \overset{a.s.}{\to} X$.

Now suppose that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(d(X_n, X) > \epsilon_n\right) < +\infty,$$

for some sequence $\{\epsilon_n\}_{n\in N_+}$ of positive scalars converging to 0. Again, by the Borel-

Cantelli lemma,

$$\sum_{n=1}^{\infty} I_{\{d(X_n,X)>\epsilon_n\}} < +\infty$$

almost surely. Letting $\Omega_0 \in \mathcal{H}$ be this almost sure set, for any $\omega \in \Omega_0$, we have

$$\sum_{n=1}^{\infty} I_{\{d(X_n(\omega),X(\omega))>\epsilon_n\}} < +\infty,$$

so that there exists an $N \in N_+$ such that

$$d(X_n(\omega), X(\omega)) \leq \epsilon_n$$

for any $n \geq N$. Taking $n \to \infty$ on both sides now implies that $d(X_n(\omega), X(\omega)) \to 0$, so that $X_n \overset{a.s.}{\to} X$.

Finally, suppose that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(d(X_n, X) > \epsilon_n\right) < +\infty,$$

for some sequence $\{\epsilon_n\}_{n\in N_+}$ of positive scalars such that $\sum_{n=1}^{\infty} \epsilon_n < +\infty$. By the nth term test for the convergence of series, the fact that $\sum \epsilon_n$ converges indicates that $\epsilon_n \to 0$ as $n \to \infty$, and as such $X_n \overset{a.s.}{\to} X$ follows from ii).

Q.E.D.

3.3 Convergence in Probability

Let (E, d) be a separable metric space, and define the Borel σ-algebra $\mathcal{E}$ as in the previous section. For any sequence $\{X_n\}_{n \in N_+}$ of random variables taking values in E, we say that $\{X_n\}_{n \in N_+}$ converges in probability to some random variable X in E if, for any $\delta > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(d(X_n, X) > \delta\right) = 0.$$

We denote this mode of convergence succinctly as

$$X_n \xrightarrow{p} X.$$

This form of convergence is the one that we encounter most often in probability, statistics and econometrics. Its popularity is due to its relative strength and ubiquity; it is stronger than weak convergence (as we will see in a later section) but weaker than almost sure convergence.

One fact that is immediately noticable is that the probability limit of a sequence of random variables is unique up to almost sure equivalence. To see this, let X and X' be probability limits of some sequence $\{X_n\}_{n \in N_+}$ of random variables taking values in E. Then,

$$d(X, X') \leq d(X, X_n) + d(X', X_n)$$

for any $n \in N_+$, so that, for any $k \in N_+$,

$$\mathbb{P}\left(d(X, X') > \frac{1}{k}\right) \leq \mathbb{P}\left(d(X_n, X) > \frac{1}{2k}\right) + \mathbb{P}\left(d(X_n, X') > \frac{1}{2k}\right)$$

for any $n \in N_+$. The terms on the right hand side go to 0 as $n \to \infty$, so it follows that

$$\mathbb{P}\left(d(X, X') > \frac{1}{k}\right) = 0.$$

This holds for any $k \in N_+$, and denoting $H_k = \{d(X, X') > \frac{1}{k}\}$, since

$$H = \{d(X, X') > 0\} = \bigcup_k H_k,$$

by countable subadditivity we have $\mathbb{P}\left(d(X, X') > 0\right) = 0$. Thus, $d(X, X') = 0$ and therefore $X = X'$ almost surely.

From the definition of convergence of probability we can see that there must be a close connection between almost sure convergence and convergence in probability, since the term $\mathbb{P}\left(d(X_n, X) > \delta\right)$ also appears in the infinite sums in theorem 3.9. We confirm below that this is indeed the case:

Theorem 3.10 Let (E,d) be a separable metric space, and define the Borel σ-algebra $\mathcal{E}$ as above. For any sequence $\{X_n\}_{n\in N_+}$ of random variables and a random variable X taking values in E, the following hold true:

i) If $X_n \overset{a.s.}{\to} X$, then $X_n \overset{p}{\to} X$.

ii) If $X_n \overset{p}{\to} X$, then there exists a subsequence $\{X_{n_k}\}_{k\in N_+}$ such that $X_{n_k} \overset{a.s.}{\to} X$ as $k \to \infty$.

iii) If every subsequence of $\{X_n\}_{n\in N_+}$ has a further subsequence that converges to X almost surely, then $X_n \overset{p}{\to} X$.

Proof) Suppose that $X_n \overset{a.s.}{\to} X$. Then, for any $\delta > 0$, the characterization theorem for almost sure convergence tells us that

$$\sum_{n=1}^{\infty} I_{\{d(X_n,X)>\delta\}} < +\infty$$

almost surely. Letting this almost sure set be $\Omega_0 \in \mathcal{H}$, this implies by the nth root test that

$$\lim_{n\to\infty} I_{\{d(X_n(\omega),X(\omega))>\delta\}} = 0$$

for any $\omega \in \Omega$. Therefore, defining $Y_n = I_{\{d(X_n,X)>\delta\}}$ for any $n \in N_+$, $\{Y_n\}_{n\in N_+}$ is a sequence of non-negative real valued random variables such that $Y_n \overset{a.s.}{\to} 0$. Since $|Y_n| \le 1$ for any $n \in N_+$, by the almost sure version of the bounded convergence theorem we can see that

$$\lim_{n\to\infty} \mathbb{E}[Y_n] = \mathbb{E}\left[\lim_{n\to\infty} Y_n\right] = 0.$$

Here, $\mathbb{E}[Y_n] = \mathbb{P}(d(X_n,X) > \delta)$ for any $n \in N_+$, so we have

$$\lim_{n\to\infty} \mathbb{P}(d(X_n,X) > \delta) = 0.$$

This holds for any $\delta > 0$, so $X_n \overset{p}{\to} X$ by definition.

Now suppose that $X_n \overset{p}{\to} X$. Then, there exists an $n_1 \in N_+$ such that

$$\mathbb{P}(d(X_{n_1},X) > 1) \le 2^{-1},$$

since $\lim_{n\to\infty} \mathbb{P}(d(X_n,X) > 1) = 0$ by definition. Assuming that we have chosen $n_1 < \cdots < n_k$ for some $k \in N_+$, we can choose $n_{k+1} > n_k$ so that

$$\mathbb{P}\left(d(X_{n_{k+1}},X) > \frac{1}{k+1}\right) \le 2^{-(k+1)},$$

since we again have $\lim_{n\to\infty}\mathbb{P}\left(d(X_n,X)>\frac{1}{k+1}\right)=0$. The subsequence $\{X_{n_k}\}_{k\in N_+}$ constructed in this manner satisfies

$$\mathbb{P}\left(d(X_{n_k},X)>\frac{1}{k}\right)\leq 2^{-k}$$

for any $K\in N_+$, so that

$$\sum_{k=1}^{\infty}\mathbb{P}\left(d(X_{n_k},X)>\frac{1}{k}\right)\leq\sum_{k=1}^{\infty}2^{-k}=1<+\infty.$$

$\{1/k\}_{k\in N_+}$ is a sequence of positive scalars converging to 0, so by the second Borel-Cantelli condition for almost sure convergence, we can see that $X_{n_k}\overset{a.s.}{\to}X$ as $k\to\infty$.

Finally, suppose that every subsequence of $\{X_n\}_{n\in N_+}$ has a further subsequence that converges almost surely to X. For any $\delta>0$, we want to show that $\{\mathbb{P}(d(X_n,X)>\delta)\}_{n\in N_+}$ converges to 0. To this end, we can make use of the selection theorem for subsequences and try to show that every subsequence of $\{\mathbb{P}(d(X_n,X)>\delta)\}_{n\in N_+}$ has a further subsequence converging to 0.

Choose some subsequence $\{\mathbb{P}(d(X_{n_k},X)>\delta)\}_{k\in N_+}$ of $\{\mathbb{P}(d(X_n,X)>\delta)\}_{n\in N_+}$. By assumption, $\{X_{n_k}\}_{k\in N_+}$ has a further subsequence $\{X_{n_{k_m}}\}_{m\in N_+}$ that converges almost surely to X. Since almost sure convergence implies convergence in probability, this means that $\{X_{n_{k_m}}\}_{m\in N_+}$ converges in probability to X, so that, by definition, $\{\mathbb{P}\left(d(X_{n_{k_m}},X)>\delta\right)\}_{m\in N_+}$ converges to 0. This shows us that $\{\mathbb{P}(d(X_{n_k},X)>\delta)\}_{k\in N_+}$ has a subsequence that converges to 0. By the selection theorem, we can conclude that $\{\mathbb{P}(d(X_n,X)>\delta)\}_{n\in N_+}$ itself converges to 0 and therefore that $X_n\overset{p}{\to}X$.

Q.E.D.

3.3.1 The Continuous Mapping Theorem

Convergence in probability proves useful in so many applications because, unlike convergence in distribution, convergence in probability is preserved under arithmetic operations such as addition and multiplication. Generally speaking, convergence in probability is preserved under any continuous transformation. This impressive result is referred to as the continuous mapping theorem, and we state it below:

Theorem 3.11 (Continuous Mapping Theorem)
Let (E,d), (F,ρ) and (G,d_G) be separable metric spaces, and define the Borel σ-algebra $\mathcal{E}$ as above, and $\mathcal{F}$ similarly using ρ. Let $\{X_n\}_{n\in N_+}$ be a sequence of random variables and X a random variable taking values in E. Likewise, let $\{Y_n\}_{n\in N_+}$ be a sequence of random variables and Y a random variable taking values in F. Letting $\mu = \mathbb{P}\circ X^{-1}$ be the distribution of X, the following hold true:

i) If $X_n \xrightarrow{a.s.} X$, then for any function $f: E\to F$ that is continuous μ-a.e., we have $f\circ X_n \xrightarrow{a.s.} f\circ X$.

ii) If $X_n \xrightarrow{p} X$, then for any function $f: E\to F$ that is continuous μ-a.e., we have $f\circ X_n \xrightarrow{p} f\circ X$.

iii) If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then sequence $\{(X_n,Y_n)\}_{n\in N_+}$ taking values in the product space $(E\times F, \mathcal{E}\bigotimes\mathcal{F})$ converges in probability to (X,Y) in the product metric $d\times\rho$.

iv) If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then for any function $f: E\times F\to G$ that is continuous with respect to the product metric $d\times\rho$ and d_G, $f(X_n,Y_n)\xrightarrow{p} f(X,Y)$.

Proof) i) Suppose that $X_n \xrightarrow{a.s.} X$ and let $f: E\to F$ be a function that is continuous μ-a.e. Denote by $D\in\mathcal{E}$ the discontinuity points of f; by assumption, $\mu(D^c) = \mathbb{P}\left(X^{-1}(D^c)\right) = 1$.

Letting $\Omega_0\in\mathcal{H}$ be the almost sure set on which pointwise convergence occurs, the set $\Omega_0\cap X^{-1}(D^c)$ also occurs with probability equal to 1. For any $\omega\in\Omega_0\cap X^{-1}(D^c)$ and $\epsilon>0$, since $X(\omega)$ is a continuity point of f by design, there exists a $\delta>0$ such that

$$\rho(f(x),f(X(\omega)))<\epsilon$$

for any $x\in E$ such that $d(x,X(\omega))<\delta$. Since $X_n(\omega)\to X(\omega)$ in the metric d, there exists an $N\in N_+$ such that

$$d(X_n(\omega),X(\omega))<\delta,$$

and in turn

$$\rho(f(X_n(\omega)),f(X(\omega)))<\epsilon,$$

for any $n \geq N$. This holds for any $\epsilon > 0$, so $f(X_n(\omega)) \to f(X(\omega))$ in the metric ρ. This in turn holds for any outcome ω in the almost sure set $\Omega_0 \cap X^{-1}(D^c)$, so we can see that $f \circ X_n \overset{a.s.}{\to} f \circ X$.

ii) Suppose that $X_n \overset{p}{\to} X$ and let $f : E \to F$ be a function that is continuous μ-a.e. We make use of results ii) and iii) in the preceding theorem. Choose any subsequence $\{f \circ X_{n_k}\}_{k \in N_+}$ of $\{f \circ X_n\}_{n \in N_+}$. Since $\{X_{n_k}\}_{k \in N_+}$ converges in probability to X, it has a subsequence $\{X_{n_{k_m}}\}_{m \in N_+}$ that converges almost surely. By the continuous mapping theorem for almost sure convergence,

$$f \circ X_{n_{k_m}} \overset{a.s.}{\to} f \circ X.$$

In other words, $\{f \circ X_{n_k}\}_{k \in N_+}$ has a further subsequence that converges almost surely to $f \circ X$. By implication, every subsequence of $\{f \circ X_n\}_{n \in N_+}$ has a further subsequence that converges almost surely to $f \circ X$, and therefore $f \circ X_n \overset{p}{\to} f \circ X$.

iii) Let τ and s be the topologies on E and F induced by d and ρ, respectively. The separability of (E, d) and (F, ρ) implies that the product topology $\tau \times s$ is metrized by the product metric $d \times \rho$. Furthermore, the second countability of (E, d) and (F, ρ), implied by their separability, shows us that the product σ-algebra $\mathcal{E} \bigotimes \mathcal{F} = \mathcal{B}(E, \tau) \bigotimes \mathcal{B}(F, s)$ is equivalent to the Borel σ-algebra generated by the product topology, $\mathcal{B}(E \times F, \tau \times s)$[1].

Now consider the sequence of random vectors $\{(X_n, Y_n)\}_{n \in N_+}$ and the random vector (X, Y) taking values in the product space $(E \times F, \mathcal{E} \bigotimes \mathcal{F})$. In light of the connection between $\mathcal{E} \bigotimes \mathcal{F}$ and the product metric $d \times \rho$, $\{(d \times \rho)((X_n, Y_n), (X, Y))\}_{n \in N_+}$ is a sequence of measurable real-valued random variables. For any $\delta > 0$ and $n \in N_+$, by the definition of the product metric,

$$(d \times \rho)((X_n, Y_n), (X, Y)) = \max\left[d(X_n, X), \rho(Y_n, Y)\right],$$

so that

$$\{(d \times \rho)((X_n, Y_n), (X, Y)) > \delta\} = \{d(X_n, X) > \delta\} \cup \{\rho(Y_n, Y) > \delta\}.$$

Therefore,

$$\mathbb{P}\left((d \times \rho)((X_n, Y_n), (X, Y)) > \delta\right) \leq \mathbb{P}\left(d(X_n, X) > \delta\right) + \mathbb{P}\left(\rho(Y_n, Y) > \delta\right).$$

[1] This result is one of the first proven in the chapter on product spaces in the measure theory text.

It follows that, if $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, both quantities on the right converge to 0 as $n \to \infty$, so that

$$\mathbb{P}\left((d \times \rho)((X_n, Y_n), (X, Y)) > \delta\right) \to 0$$

as $n \to \infty$ as well. This holds for any $\delta > 0$, so $(X_n, Y_n) \xrightarrow{p} (X, Y)$ in the product metric in this case.

iv) Suppose $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, and let $f : E \times F \to G$ be a function that is continuous relative to the metrics $d \times \rho$ and d_G. The preceding result shows us that $(X_n, Y_n) \xrightarrow{p} (X, Y)$ in the product metric $d \times \rho$, so the continuous mapping theorem for convergence in probability implies that

$$f(X_n, Y_n) \xrightarrow{p} f(X, Y)$$

in the metric d_G.

Q.E.D.

The last result, in particular, is of great practical use. Consider the case where $\{X_n\}_{n \in N_+}$ and $\{Y_n\}_{n \in N_+}$ are sequences of real random variables, and let d denote the euclidean metric on $\mathbb{R}$. Then, since the box metric and euclidean metric induce the same topology in euclidean spaces, the mappings $(x, y) \mapsto x + y$ and $(x, y) \mapsto xy$ are continuous with respect to the product metric $d \times d$ and the metric d. Applying result iii) now shows that, if $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ for some real random variables X and Y, we have

$$X_n + Y_n \xrightarrow{p} X + Y \quad \text{and} \quad X_n Y_n \xrightarrow{p} XY.$$

Meanwhile, if each X_n and X are non-zero, the operation $x \mapsto \frac{1}{x}$ is continuous μ-a.e. on $\mathbb{R}$, where μ is the distribution of X. In this case, by the continuous mapping theorem, $\frac{1}{X_n} \xrightarrow{p} \frac{1}{X}$, and the above result applies so that we may conclude

$$\frac{Y_n}{X_n} \xrightarrow{p} \frac{Y}{X}.$$

3.3.2 Big and Little O Notation in Probability

Convergence in probability is so important that mathematicians have developed a separate notation to indicate the speed of convergence in probability. We first introduce the big and little O notations for deterministic functions. Consider a complex-valued function $f : \mathbb{R} \to \mathbb{C}$ and a positive-valued function $g : \mathbb{R} \to (0, +\infty)$. We say that

$$f(x) = O(g(x)) \text{ as } x \to \infty \quad \text{if } \limsup_{x \to \infty} \frac{|f(x)|}{g(x)} < +\infty.$$

Heuristically, $f(x)$ is $O(g(x))$ if $f(x)$ and $g(x)$ grow (decrease) at around the same speed as $x \to \infty$. The little O notation is reserved for a more acute case; for any $a \in [0, +\infty]$, we say that

$$f(x) = o(g(x)) \text{ as } x \to a \quad \text{if } \lim_{x \to a} \frac{|f(x)|}{g(x)} = 0,$$

that is, $f(x)$ is $o(g(x))$ if $f(x)$ decreases to 0 (grows to ∞) faster (slower) than $g(x)$ as $x \to a$. Note that $f(x)$ is $O(g(x))$ as $x \to \infty$ if it is $o(g(x))$ as $x \to \infty$.

Big and little O notations appear in many areas of analysis. For instance, consider a smooth function $f : [a, b] \to \mathbb{C}$, that is, a function that is infinitely differentiable on the interval (a, b). Then, for any $n \in N_+$ and $x, x + h \in (a, b)$, the nth-order Taylor expansion of $f(x + h)$ around x is given as

$$f(x + h) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!} h^k + r_n(h),$$

where

$$\lim_{h \to 0} \frac{|r_n(h)|}{|h|^n} = 0.$$

This shows us that

$$r_n(h) = o(|h|^n) \text{ as } h \to 0,$$

and as such we can rewrite the Taylor expansion as

$$f(x + h) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!} h^k + o(|h|^n).$$

We can also use the big and little O notations to discuss the speed of convergence of a complex valued sequence. Let $\{a_n\}_{n \in N_+}$ be a complex-valued sequence that converges to 0; note that this sequence is simply a complex-valued function with domain equal to the set of all natural numbers N_+. Consider some positive-valued function $g : N_+ \to (0, +\infty)$ such that $g(n) \to 0$ as $n \to \infty$; usually, we put $g(n) = \frac{1}{n^p}$ for some $0 < p < +\infty$. We say that

$$\{a_n\}_{n \in N_+} = O(g(n)) \text{ as } n \to \infty \quad \text{if } \limsup_{n \to \infty} \frac{|a_n|}{g(n)} < +\infty$$

$$\{a_n\}_{n\in N_+} = o(g(n)) \text{ as } n \to \infty \quad \text{if } \lim_{n\to\infty} \frac{|a_n|}{g(n)} = 0.$$

In other words, $\{a_n\}_{n\in N_+}$ is $O(g(n))$ if it converges to 0 at around the same speed as $g(n)$, while it is $o(g(n))$ if it converges to 0 faster than $g(n)$.

Big and little O notation in probability is defined in a similar manner to big and little O notation for deterministic functions and sequences. A sequence $\{X_n\}_{n\in N_+}$ of complex random variables is said to be bounded in probability if, for any $\epsilon > 0$, there exist an $M > 0$ and $N \in N_+$ such that

$$\mathbb{P}(|X_n| > M) < \epsilon$$

for any $n \geq N$. We show below that we can take $N = 1$ if each X_n is integrable:

Lemma 3.12 Let $\{X_n\}_{n\in N_+}$ be a sequence of integrable complex random variables. Then, $\{X_n\}_{n\in N_+}$ is bounded in probability if and only if, for any $\epsilon > 0$, there exists an $M > 0$ such that

$$\sup_{n\in N_+} \mathbb{P}(|X_n| > M) < \epsilon.$$

Proof) Sufficiency follows immediately. As such, we turn our attention to necessity. Suppose that $\{X_n\}_{n\in N_+}$ is bounded in probability, so that, for any $\epsilon > 0$, there exist an $M > 0$ and $N \in N_+$ such that

$$\mathbb{P}(|X_n| > M) < \epsilon$$

for any $n \geq N$. Since each X_n is integrable, by the characterization of integrability studied above,

$$\lim_{k\to\infty} \mathbb{E}\left[|X_n| \cdot I_{\{|X_n|>k\}}\right] = 0.$$

Therefore, for each $1 \leq n \leq N$ there exists an $M_n \in N_+$ such that

$$\mathbb{E}\left[|X_n| \cdot I_{\{|X_n|>M_n\}}\right] < \epsilon,$$

and as such

$$\mathbb{P}(|X_n| > M_n) \leq M_n \cdot \mathbb{P}(|X_n| > M_n) \leq \mathbb{E}\left[|X_n| \cdot I_{\{|X_n|>M_n\}}\right] < \epsilon.$$

Define

$$M^* = \max(M, M_1, \cdots, M_N) > 0.$$

We can now easily see that

$$\mathbb{P}(|X_n| > M^*) < \epsilon$$

for any $n \in N_+$.

Q.E.D.

Given a sequence $\{X_n\}_{n \in N_+}$ of complex or non-negative random variables and a sequence $\{a_n\}_{n \in N_+}$ of positive real numbers, we say that

$$X_n = O_p(a_n) \text{ as } n \to \infty \quad \text{if } \{X_n / a_n\}_{n \in N_+} \text{ is bounded in probability.}$$

In other words, X_n is $O_p(a_n)$ if $\{X_n\}_{n \in N_+}$ and $\{a_n\}_{n \in N_+}$ converge or diverge at around the same speed as $n \to \infty$.

On the other hand, we say that

$$X_n = o_p(a_n) \text{ as } n \to \infty \quad \text{if} \quad \frac{X_n}{a_n} \xrightarrow{p} 0.$$

Heuristically, X_n is $o_p(a_n)$ if $\{X_n\}_{n \in N_+}$ converges to 0 faster than $\{a_n\}_{n \in N_+}$.

Below we present some important results concerning big and little O notation in probability:

Theorem 3.13 (Properties of Big and Little O Notation in Probability) The following results hold true:

i) An $o_p(1)$ sequence is also $O_p(1)$.

ii) $O_p(1)O_p(1) = O_p(1)$; if $X_n = O_p(1)$ and $Y_n = O_p(1)$, then $X_nY_n = O_p(1)$.

iii) $o_p(1)O_p(1) = o_p(1)$; if $X_n = o_p(1)$ and $Y_n = O_p(1)$, then $X_nY_n = o_p(1)$.

iv) $O_p(1) + O_p(1) = O_p(1)$; if $X_n = O_p(1)$ and $Y_n = O_p(1)$, then $X_n + Y_n = O_p(1)$.

v) $O_p(1) + o_p(1) = O_p(1)$; if $X_n = O_p(1)$ and $Y_n = o_p(1)$, then $X_n + Y_n = O_p(1)$.

vi) If $\{a_n\}_{n\in N_+}$ and $\{b_n\}_{n\in N_+}$ are positive real-valued sequences such that $\frac{a_n}{b_n} \to 0$, then

$$O_p(a_n) + O_p(b_n) = O_p(b_n) \quad \text{and} \quad O_p(a_n)O_p(b_n) = O_p(a_nb_n).$$

vii) If $\{a_n\}_{n\in N_+}$ is a positive real-valued sequence such that $a_n \nearrow +\infty$, then any $O_p(1)$ sequence is also $o_p(a_n)$.

viii) Any L^1-bounded sequence of complex random variables is $O_p(1)$.

ix) Any uniformly integrable sequence of complex random variables is $O_p(1)$.

x) Let $\{X_n\}_{n\in N_+}$ be a sequence of complex random variables and $\{a_n\}_{n\in N_+}$ a sequence of positive real numbers. If $\{a_n\}_{n\in N_+}$ is convergent in $\mathbb{R}$ and $|X_n| \leq a_n$ for any $n \in N_+$, then $X_n = O_p(1)$.

xi) Let $\{X_n\}_{n\in N_+}$ be a sequence of complex random variables and $\{Y_n\}_{n\in N_+}$ a sequence of non-negative random variables. If $Y_n = O_p(1)$ and $|X_n| \leq Y_n$ for any $n \in N_+$, then $X_n = O_p(1)$.

Proof) i) Suppose that $\{X_n\}_{n\in N_+}$ is $o_p(1)$, so that $X_n \xrightarrow{p} 0$. Since

$$\lim_{n\to\infty} \mathbb{P}(|X_n| > 1) = 0$$

by definition, it follows that, for any $\epsilon > 0$, there exists an $N \in N_+$ such that

$$\mathbb{P}(|X_n| > 1) < \epsilon$$

for any $n \geq N$. Thus, $X_n = O_p(1)$ with $M = 1$ for any $\epsilon > 0$.

ii) Let $X_n = O_p(1)$ and $Y_n = O_p(1)$. Then, for any $\epsilon > 0$ there exist $M_X, M_Y > 0$ and $N_X, N_Y \in N_+$ such that

$$\mathbb{P}(|X_n| > M_X) < \frac{\epsilon}{2}, \quad \text{and} \quad \mathbb{P}(|Y_n| > M_Y) < \frac{\epsilon}{2}$$

for any $n \geq N = \max(N_X, N_Y)$. Defining $M = M_X M_Y > 0$, note that

$$\{|X_n| \leq M_X\} \cap \{|Y_n| \leq M_Y\} \subset \{|X_n Y_n| \leq M\};$$

therefore, for any $n \geq N$,

$$\mathbb{P}(|X_n Y_n| > M) \leq \mathbb{P}(|X_n| > M_X) + \mathbb{P}(|Y_n| > M_Y) < \epsilon.$$

Such an $M > 0$ and $N \in N_+$ exist for any $\epsilon > 0$, so by definition $X_n Y_n = O_p(1)$.

iii) Let $X_n = o_p(1)$ and $Y_n = O_p(1)$. Choose any $\delta > 0$. For any $\epsilon > 0$ there exist $M >$ and $N_1 \in N_+$ such that

$$\mathbb{P}(|Y_n| > M) < \frac{\epsilon}{2}$$

for any $n \geq N_1$. By definition of convergence in probability,

$$\lim_{n \to \infty} \mathbb{P}\left(|X_n| > \frac{\delta}{M}\right) = 0,$$

so there exists an $N \geq N_1$ such that

$$\mathbb{P}\left(|X_n| > \frac{\delta}{M}\right) < \frac{\epsilon}{2}$$

for any $n \geq N$. Since

$$\{|X_n| \leq \frac{\delta}{M}\} \cap \{|Y_n| \leq M\} \subset \{|X_n Y_n| \leq \delta\},$$

we can see that, for any $n \geq N$,

$$\mathbb{P}(|X_n Y_n| > \delta) \leq \mathbb{P}\left(|X_n| > \frac{\delta}{M}\right) + \mathbb{P}(|Y_n| > M) < \epsilon.$$

This holds for any $\epsilon > 0$, so it follows that

$$\lim_{n\to\infty} \mathbb{P}(|X_n Y_n| > \delta) = 0.$$

This in turn holds for any $\delta > 0$, so by definition $X_n Y_n \xrightarrow{p} 0$ and thus $X_n Y_n = o_p(1)$.

iv) The proof is very similar to that of ii). We state it here for the sake of completeness. Let $X_n = O_p(1)$ and $Y_n = O_p(1)$. Then, for any $\epsilon > 0$ there exist $M_X, M_Y > 0$ and $N_X, N_Y \in N_+$ such that

$$\mathbb{P}(|X_n| > M_X) < \frac{\epsilon}{2}, \quad \text{and} \quad \mathbb{P}(|Y_n| > M_Y) < \frac{\epsilon}{2}$$

for any $n \geq N = \max(N_X, N_Y)$. Defining $M = M_X + M_Y > 0$, note that

$$\{|X_n| \leq M_X\} \cap \{|Y_n| \leq M_Y\} \subset \{|X_n + Y_n| \leq M\},$$

since

$$|X_n(\omega) + Y_n(\omega)| \leq |X_n(\omega)| + |Y_n(\omega)| \leq M_X + M_Y = M$$

if the outcome ω is contained in the intersection on the left hand side. Therefore, for any $n \geq N$,

$$\mathbb{P}(|X_n + Y_n| > M) \leq \mathbb{P}(|X_n| > M_X) + \mathbb{P}(|Y_n| > M_Y) < \epsilon.$$

Such an $M > 0$ and $N \in N_+$ exist for any $\epsilon > 0$, so by definition $X_n + Y_n = O_p(1)$.

v) If $X_n = O_p(1)$ and $Y_n = o_p(1)$, then $Y_n = O_p(1)$ by i). This implies that $X_n + Y_n = O_p(1)$ by the preceding result.

vi) Suppose $\{a_n\}_{n\in N_+}$ and $\{b_n\}_{n\in N_+}$ are positive real-valued sequences such that $\frac{a_n}{b_n} \to 0$. Let $X_n = O_p(a_n)$ and $Y_n = O_p(b_n)$. Defining

$$Z_n = \frac{a_n}{b_n}$$

for any $n \in N_+$, $\{Z_n\}_{n \in N_+}$ is a sequence of (degenerate) real random variables that converges almost surely and thus in probability to 0 as $n \to \infty$. Therefore, $Z_n = o_p(1)$.

Note that, by definition,

$$\frac{X_n}{a_n} = O_p(1) \quad \text{and} \quad \frac{Y_n}{b_n} = O_p(1).$$

It follows that

$$\frac{X_n + Y_n}{b_n} = \frac{a_n}{b_n} \cdot \frac{X_n}{a_n} + \frac{Y_n}{b_n} = o_p(1)O_p(1) + O_p(1) = o_p(1) + O_p(1) = O_p(1)$$

by combining the preceding results. By definition, we can see that $X_n + Y_n = O_p(b_n)$.

Similarly,

$$\frac{X_n Y_n}{a_n b_n} = \frac{X_n}{a_n} \cdot \frac{Y_n}{b_n} = O_p(1)O_p(1) = O_p(1),$$

so that $X_n Y_n = O_p(a_n b_n)$.

vii) Suppose that $\{a_n\}_{n \in N_+}$ is a positive real valued sequence such that $a_n \nearrow +\infty$, and let $X_n = O_p(1)$. Choose any $\delta > 0$. For any $\epsilon > 0$, there exists an $M > 0$ and $N_1 \in N_+$ such that

$$\mathbb{P}(|X_n| > M) < \epsilon$$

for any $n \geq N_1$. Since $\{a_n\}_{n \in N_+}$ increases to $+\infty$, there exists an $N \geq N_1$ such that $a_n > \frac{M}{\delta}$ for any $n \geq N$. It then follows that

$$\mathbb{P}\left(\left|\frac{X_n}{a_n}\right| > \delta\right) = \mathbb{P}(|X_n| > \delta \cdot a_n) \leq \mathbb{P}(|X_n| > M) < \epsilon$$

for any $n \geq N$. This holds for any $\epsilon > 0$, so it follows that

$$\lim_{n \to \infty} \mathbb{P}\left(\left|\frac{X_n}{a_n}\right| > \delta\right) = 0.$$

This in turn holds for any $\delta > 0$, so by definition, $\frac{X_n}{a_n} \xrightarrow{p} 0$ and thus $X_n = o_p(a_n)$.

viii) Let $\{X_n\}_{n \in N_+}$ be an L^1-bounded sequence, that is,

$$K := \sup_{n \in N_+} \mathbb{E}|X_n| < +\infty.$$

For any $\epsilon > 0$, by Markov's inequality

$$\mathbb{P}\left(|X_n| > \frac{K}{\epsilon}\right) \leq \frac{\epsilon}{K}\mathbb{E}|X_n| \leq \epsilon$$

for any $n \in N_+$. Therefore, $X_n = O_p(1)$, or bounded in probability, with $M = \frac{K}{\epsilon}$ and $N = 1$ for each $\epsilon > 0$.

ix) If $\{X_n\}_{n\in N_+}$ is uniformly integrable, then it is also L^1-bounded, so that $X_n = O_p(1)$ by the preceding result.

x) Let $\{X_n\}_{n\in N_+}$ be a sequence of complex random variables and $\{a_n\}_{n\in N_+}$ a sequence of positive real numbers. Suppose that $a_n \to a$ for some $a \in \mathbb{R}$ in the euclidean metric on $\mathbb{R}$, and that $|X_n| \leq a_n$ for any $n \in N_+$. Then, $\{a_n\}_{n\in N_+}$ is a bounded sequence, so that there exists an $M > 0$ such that $0 < a_n < M$ for any $n \in N_+$. By implication, $|X_n| \leq M$ for any $n \in N_+$, so that $\{X_n\}_{n\in N_+}$ is L^1-bounded:

$$\sup_{n\in N_+} \mathbb{E}|X_n| \leq M < +\infty.$$

Therefore, $X_n = O_p(1)$.

xi) Let $\{X_n\}_{n\in N_+}$ be a sequence of complex random variables and $\{Y_n\}_{n\in N_+}$ a sequence of non-negative random variables such that $Y_n = O_p(1)$ and $|X_n| \leq Y_n$ for any $n \in N_+$. For any $\epsilon > 0$, there exists an $M > 0$ and $N \in N_+$ such that

$$\mathbb{P}(Y_n > M) < \epsilon$$

for any $n \geq N$. Since $|X_n| > M$ implies $Y_n > M$, we have

$$\mathbb{P}(|X_n| > M) \leq \mathbb{P}(Y_n > M) < \epsilon$$

for any $n \geq N$. This holds for any $\epsilon > 0$, so by definition $X_n = O_p(1)$.

Q.E.D.

3.3.3 Metric for Convergence in Probability

It is convenient to establish a metric for convergence in probability. In other words, given a separable metric space (E,d) with corresponding Borel σ-algebra $\mathcal{E}$, we want to find a metric d_{prob} on the space $\mathcal{H}/\mathcal{E}$ of random variables taking values in E such that, for any $\{X_n\}_{n\in N_+}$ and X in $\mathcal{H}/\mathcal{E}$,

$$X_n \to X \text{ in the metric } d_{prob}$$

if and only if $X_n \xrightarrow{p} X$. d_{prob} is useful because it allows us to apply results that hold for the convergence of sequences in metric spaces to convergence in probability, although we did not define the latter in terms of any metric on $\mathcal{H}/\mathcal{E}$.

One way to define d_{prob} in terms of the metric d of the underlying space E is

$$d_{prob}(X,Y) = \mathbb{E}\left[\min(d(X,Y),1)\right]$$

for any random variables X,Y taking values in E. $d_{prob}(X,Y)$ is well-defined and bounded above by 1 because the separability of (E,d) implies that $\min(d(X,Y),1)$ is a non-negative random variable bounded above by 1. We can also see that d_{prob} is a metric on $\mathcal{H}/\mathcal{E}$, given that we identify almost surely equivalent random variables:

- $d_{prob}(X,Y) = 0$ **if and only if** $X = Y$ **almost surely**
  Suppose that $d_{prob}(X,Y) = 0$ for some $X,Y \in \mathcal{H}/\mathcal{E}$. Then,

  $$\mathbb{E}\left[\min(d(X,Y),1)\right] = 0,$$

  and by the vanishing property of non-negative functions, $\min(d(X,Y),1) = d(X,Y) = 0$ almost surely. By implication, $X = Y$ almost surely.

  Conversely, if $X = Y$ almost surely, then $\min(d(X,Y),1) = 0$ almost surely and therefore $d_{prob} = \mathbb{E}\left[\min(d(X,Y),1)\right] = 0$.

- **Symmetry**
  For any $X,Y \in \mathcal{H}/\mathcal{E}$, $d(X,Y) = d(Y,X)$ and as such

  $$d_{prob}(X,Y) = \mathbb{E}\left[\min(d(X,Y),1)\right] = \mathbb{E}\left[\min(d(Y,X),1)\right] = d_{prob}(Y,X).$$

- **Triangle Inequality**
  For any $X,Y,Z \in \mathcal{H}/\mathcal{E}$,

  $$d(X,Y) \leq d(X,Z) + d(Z,Y)$$

because d satisfies the triangle inequality, which tells us that

$$\begin{aligned}\min(d(X,Y),1) &\leq \min\left(d(X,Z)+d(Z,Y),1\right)\\ &\leq \min(d(X,Z),1)+\min(d(Z,Y),1).\end{aligned}$$

By the monotonicity and linearity of the integration of non-negative functions, we now have

$$\begin{aligned}d_{prob}(X,Y) &= \mathbb{E}\left[\min(d(X,Y),1)\right]\\ &\leq \mathbb{E}\left[\min(d(X,Z),1)\right]+\mathbb{E}\left[\min(d(Z,Y),1)\right] = d_{prob}(X,Z)+d_{prob}(Z,Y).\end{aligned}$$

The next result shows us that convergence of probability is equivalent to convergence in the metric space $(\mathcal{H}/\mathcal{E}, d_{prob})$:

Theorem 3.14 Let (E,d) be a metric space and $\mathcal{E}$ the corresponding Borel σ-algebra. A sequence $\{X_n\}_{n\in N_+}$ of random variables taking values in E converges to some random variable X taking values in E if and only if $\{X_n\}_{n\in N_+}$ converges to X in the metric d_{prob}.

Proof) Suppose that $X_n \xrightarrow{p} X$. Choose any $0<\delta\leq 1$. Then, for any $n\in N_+$,

$$\begin{aligned}d_{prob}(X_n,X) &= \mathbb{E}\left[\min(d(X_n,X),1)\right]\\ &= \mathbb{E}\left[d(X_n,X)\cdot I_{\{d(X_n,X)\leq\delta\}}\right]+\mathbb{E}\left[\min(d(X_n,X),1)\cdot I_{\{d(X_n,X)>\delta\}}\right]\\ &\leq \delta\cdot\mathbb{P}\left(d(X_n,X)\leq\delta\right)+\mathbb{P}\left(d(X_n,X)>\delta\right)\\ &\leq \delta+\mathbb{P}\left(d(X_n,X)>\delta\right).\end{aligned}$$

By the definition of convergence in probability,

$$\lim_{n\to\infty}\mathbb{P}\left(d(X_n,X)>\delta\right)=0,$$

so taking $n\to\infty$ on both sides shows us that

$$\limsup_{n\to\infty} d_{prob}(X_n,X)\leq\delta.$$

This holds for any $0<\delta\leq 1$, so it follows that

$$\lim_{n\to\infty} d_{prob}(X_n,X)=0,$$

and as such that $X_n\to X$ in the metric d_{prob}.

Conversely, suppose that $X_n\to X$ in the metric d_{prob}. Choose any $0<\delta<1$. Then, because $d(X_n,X)>\delta$ is equivalent to $\min(d(X_n,X),1)>\delta$ due to the fact that $\delta<1$,

we can see that

$$\begin{aligned}\mathbb{P}(d(X_n, X) > \delta) &= \mathbb{P}(\min(d(X_n, X), 1) > \delta) \\ &\leq \frac{1}{\delta}\mathbb{E}[\min(d(X_n, X), 1)] = \frac{1}{\delta}d_{prob}(X_n, X)\end{aligned}$$

for any $n \in N_+$ by Markov's inequality. Taking $n \to \infty$ on both sides shows us that

$$\lim_{n\to\infty} \mathbb{P}(d(X_n, X) > \delta) = 0.$$

If $\delta \geq 1$, then

$$\mathbb{P}(d(X_n, X) > \delta) \leq \mathbb{P}\left(d(X_n, X) > \frac{1}{2}\right)$$

for any $n \in N_+$, and since we just showed that the term on the right hand side goes to 0 as $n \to \infty$, we again have

$$\lim_{n\to\infty} \mathbb{P}(d(X_n, X) > \delta) = 0.$$

This holds for any $\delta > 0$, so by definition $X_n \xrightarrow{p} X$.

Q.E.D.

3.4 Convergence in L^p

We now focus on a useful form of convergence for complex valued random variables. For $1 \leq p < +\infty$, let $\{X_n\}_{n \in N_+}$ be a sequence of complex variables in the vector space $L^p(\mathcal{H}, \mathbb{P})$ over the complex field, that is,

$$\mathbb{E}|X_n|^p = \int_\Omega |X_n|^p d\mathbb{P} < +\infty$$

for each $n \in N_+$. Let $\|\cdot\|_p$ be the L^p norm on $L^p(\mathcal{H}, \mathbb{P})$ defined as

$$\|X\|_p = (\mathbb{E}|X|^p)^{\frac{1}{p}}$$

for any $X \in L^p(\mathcal{H}, \mathbb{P})$. We say that $\{X_n\}_{n \in N_+}$ converges in L^p to some random variable $X \in L^p(\mathcal{H}, \mathbb{P})$ if

$$\lim_{n \to \infty} \|X_n - X\|_p = 0.$$

This mode of convergence is denoted $X_n \xrightarrow{L^p} X$. Denoting $d_p : L^p(\mathcal{H}, \mathbb{P})^2 \to [0, +\infty)$ the metric induced by the L^p norm, $X_n \xrightarrow{L^p} X$ is essentially saying that $X_n \to X$ in the metric d_p. Recall that the Riesz-Fischer theorem, one of the most important results in L^p theory, states that the metric space $(L^p(\mathcal{H}, \mathbb{P}), d_p)$ is a complete metric space.

Convegence in L^p has a useful characterization in terms of convergence in probability and uniform integrability. Indeed, it is not an understatement to say that this result is the rmain eason for studying uniform integrability.

Theorem 3.15 (Characterization fo Convergence in L^p)
Let $\{X_n\}_{n \in N_+}$ be a sequence of complex random variables in $L^p(\mathcal{H}, \mathbb{P})$ for $1 \leq p < +\infty$. Then, the following statements are equivalent:

i) There exists an $X \in L^p(\mathcal{H}, \mathbb{P})$ such that $X_n \xrightarrow{L^p} X$.

ii) $\{X_n\}_{n \in N_+}$ is Cauchy in d_p, that is,

$$\lim_{n,m \to \infty} \|X_n - X_m\|_p = 0.$$

iii) $\{|X_n|^p\}_{n \in N_+}$ is uniformly integrable and $X_n \xrightarrow{p} X$ for some complex random variable X.

Proof) The equivalence of i) and ii) follows from the Riesz-Fischer theorem. As such, we focus on proving that i) and iii) are equivalent.

Suppose that $X_n \xrightarrow{L^p} X$ for some $X \in L^p(\mathcal{H}, \mathbb{P})$. We first show that $X_n \xrightarrow{p} X$. By Markov's inequality,

$$\mathbb{P}(|X_n - X| > \delta) \leq \frac{1}{\delta^p}\mathbb{E}|X_n - X|^p = \frac{1}{\delta^p}\|X_n - X\|_p^p$$

for any $\delta > 0$. Since $X_n \xrightarrow{L^p} X$, the expression on the right hand side goes to 0 as $n \to \infty$, so that

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \delta) = 0.$$

This holds for any $\delta > 0$, so by definition $X_n \xrightarrow{p} X$.

We now show that $\{|X_n|^p\}_{n\in N_+}$ must be uniformly integrable using the $\epsilon - \delta$ characterization of uniform integrability. We will make extensive use of the inequality

$$|x + y|^p \leq 2^{p-1}(|x|^p + |y|^p)$$

for any $x, y \in \mathbb{C}$. Choose any $\epsilon > 0$ Then, for any $n \in N_+$ and $H \in \mathcal{H}$,

$$\begin{aligned}\mathbb{E}[|X_n|^p \cdot I_H] &\leq 2^{p-1}[\mathbb{E}[|X_n - X|^p \cdot I_H] + \mathbb{E}[|X|^p \cdot I_H]) \\ &\leq 2^{p-1}\|X_n - X\|_p^p + \mathbb{E}[|X|^p \cdot I_H].\end{aligned}$$

The singleton $\{|X|^p\}$ is uniformly integrable because $|X|^p$ is integrable; therefore, there exists a $\delta_0 > 0$ such that

$$\mathbb{E}[|X|^p \cdot I_H] < \frac{\epsilon}{2}$$

for any $H \in \mathcal{H}$ such that $\mathbb{P}(H) < \delta_0$. That $X_n \xrightarrow{L^p} X$ implies that there exist a $N \in N_+$ such that

$$\|X_n - X\|_p < \frac{1}{2^{p-1}}\left(\frac{\epsilon}{2}\right)^{\frac{1}{p}}$$

for any $n \geq N$. As such, we can see that, for any $H \in \mathcal{H}$ such that $\mathbb{P}(H) < \delta_0$,

$$\sup_{n\geq N} \mathbb{E}[|X_n|^p \cdot I_H] \leq \epsilon.$$

Since the collection $\{|X_1|^p, \cdots, |X_N|^p\}$ is uniformly integrable, there exists a $\delta_1 > 0$ such that

$$\sup_{1\leq i\leq N} \mathbb{E}[|X_i|^p \cdot I_H] \leq \epsilon$$

for any $H \in \mathcal{H}$ such that $\mathbb{P}(H) < \delta_1$. Defining $\delta = \min(\delta_0, \delta_1) > 0$, it follows that, for

any $H \in \mathcal{H}$ such that $\mathbb{P}(H) < \delta$, we have

$$\sup_{n \in N_+} \mathbb{E}\left[|X_n|^p \cdot I_H\right] \leq \epsilon.$$

This holds for any $\epsilon > 0$, and thus $\{|X_n|^p\}_{n \in N_+}$ is uniformly integrable.

Conversely, suppose that $\{|X_n|^p\}_{n \in N_+}$ is uniformly integrable and that $X_n \xrightarrow{p} X$ for some complex random variable X. We must show that $X_n \xrightarrow{L^p} X$. By the characterization of convergence in probability, there exists a subsequence $\{X_{n_k}\}_{k \in N_+}$ such that $X_{n_k} \xrightarrow{a.s.} X$; by the continuous mapping theorem, $|X_{n_k}|^p \xrightarrow{a.s.} |X|^p$. It follows from the almost sure version of Fatou's lemma that

$$\mathbb{E}|X|^p = \mathbb{E}\left[\liminf_{k \to \infty} |X_{n_k}|^p\right] \leq \liminf_{n \to \infty} \mathbb{E}|X_{n_k}|^p \leq \sup_{n \in N_+} \mathbb{E}|X_n|^p < +\infty,$$

where the last inequality follows because $\{|X_n|^p\}_{n \in N_+}$, being uniformly integrable, is L^1-bounded. Therefore, $X \in L^p(\mathcal{H}, \mathbb{P})$. It follows that $\{|X_n|^p\}_{n \in N_+} \cup \{|X|^p\}$ is uniformly integrable.

For any $\epsilon > 0$, by uniform integrability there exists a $\delta > 0$ such that

$$\sup_{n \in N_+} \mathbb{E}\left[|X_n|^p \cdot I_H\right] < \epsilon, \quad \mathbb{E}\left[|X|^p \cdot I_H\right] < \epsilon$$

for any $H \in \mathcal{H}$ such that $\mathbb{P}(H) < \delta$. The fact that $X_n \xrightarrow{p} X$ implies that $|X_n - X|^p \xrightarrow{p} 0$ by the continuous mapping theorem, and as such there exists an $N \in N_+$ such that

$$\mathbb{P}\left(|X_n - X|^p > \epsilon\right) < \delta$$

for any $n \geq N$. Then, for any $n \geq N$,

$$\sup_{n \in N_+} \mathbb{E}\left[|X_n|^p \cdot I_{\{|X_n - X|^p > \epsilon\}}\right] < \epsilon, \quad \mathbb{E}\left[|X|^p \cdot I_{\{|X_n - X|^p > \epsilon\}}\right] < \epsilon,$$

and we have

$$\begin{aligned}
\mathbb{E}|X_n - X|^p &\leq \mathbb{E}\left[|X_n - X|^p \cdot I_{\{|X_n - X|^p > \epsilon\}}\right] + \mathbb{E}\left[|X_n - X|^p \cdot I_{\{|X_n - X|^p \leq \epsilon\}}\right] \\
&\leq \mathbb{E}\left[|X_n - X|^p \cdot I_{\{|X_n - X|^p > \epsilon\}}\right] + \epsilon \\
&\leq 2^{p-1}\left[\mathbb{E}\left[|X_n|^p \cdot I_{\{|X_n - X|^p > \epsilon\}}\right] + \mathbb{E}\left[|X|^p \cdot I_{\{|X_n - X|^p > \epsilon\}}\right]\right] + \epsilon \\
&\leq (2^p + 1)\,\epsilon.
\end{aligned}$$

This holds for any $\epsilon > 0$, so

$$\lim_{n \to \infty} \mathbb{E}|X_n - X|^p = 0$$

and $X_n \overset{L^p}{\to} X$ by definition.

Q.E.D.

Chapter 4

Convergence of Measures

So far we have introduced various forms in which random variables can converge. Sometimes, however, we wish to study the convergence, not of the random variables themselves, but of their distributions. In particular, we focus on a type of convergence referred to as weak convergence, or convergence in distribution. One particular advantage this approach has over the previous ones is that we do not necessarily need to define a probability space beforehand, since the convergence is of the distributions instead of the variables. This leads to a variety of interesting results, including one that demonstrates that suitably choosing the probability space leads to the equivalence of weak convergence and almost sure convergence.

Let (E,τ) be a topological space and $\mathcal{E}$ the Borel σ-algebra generated by τ. We say that a sequence $\{\mu_n\}_{n\in N_+}$ of probability measures on $(E,\mathcal{E})$ converges weakly to another probability meausre μ on $(E,\mathcal{E})$ if, for any bounded and continuous real function f on E,

$$\lim_{n\to\infty}\int_E f d\mu_n = \int_E f d\mu,$$

and we write this relationship $\mu_n \to \mu$ weakly. The set of all such functions is often denoted by $C_b(E,\mathbb{R})$.
Suppose that there exists a sequence $\{X_n\}_{n\in N_+}$ of random variables in $(E,\mathcal{E})$ such that the distribution of each X_n is μ_n, and suppose X is another random variable in $(E,\mathcal{E})$ with distribution μ. In this case, if $\mu_n \to \mu$ weakly, then we say that X_n converges to X in distribution, and write $X_n \xrightarrow{d} X$. The defining property of weak convergence can then be formulated as

$$\lim_{n\to\infty}\mathbb{E}[f\circ X_n] = \mathbb{E}[f\circ X]$$

for any bounded and continuous real valued function f on E.

4.1 The Portmanteau Theorem

The definition of weak convergence is often very difficult to work with, since the space of all bounded and continuous real valued functions on E is large and abstract. We thus rely on the

following characterization:

Theorem 4.1 (The Portmanteau Theorem)
Let (E,d) be a metric space, τ the metric topology induced by d, and $\mathcal{E}$ the Borel σ-algebra on E generated by τ. Let $\{\mu_n\}_{n\in N_+}$ be a sequence of probability measures on $(E,\mathcal{E})$ and μ another probability measure on $(E,\mathcal{E})$. Then, the following are equivalent:

i) $\mu_n \to \mu$ weakly.

ii) For any Lipschitz continuous and bounded real-valued function f on E,

$$\lim_{n\to\infty}\int_E f d\mu_n = \int_E f d\mu.$$

iii) For any closed subset F of E,

$$\limsup_{n\to\infty}\mu_n(F) \le \mu(F).$$

iv) For any open subset V of E,

$$\liminf_{n\to\infty}\mu_n(V) \ge \mu(V).$$

v) For any Borel subset A of E such that $\mu(\partial A)=0$ (we call these sets μ-continuity sets),

$$\lim_{n\to\infty}\mu_n(A) = \mu(A).$$

vi) (If $E=\mathbb{R}$ and $\tau=\tau_{\mathbb{R}}$) Letting F_n be the distribution function corresponding to each μ_n and F that of μ, for any $x\in\mathbb{R}$ at which F is continuous,

$$\lim_{n\to\infty}F_n(x) = F(x).$$

Proof) **i) → ii)**
This follows by definition, since any Lipschitz continuous bounded real valued function on E is also a continuous bounded real valued function on E.

ii) → iii)
Suppose that ii) holds. Choose any closed set $F\subset E$. The distance function $x\mapsto d(x,F)$ is continuous on E, which implies that the sequence $\{V_n\}_{n\in N_+}$ of subsets of E defined as

$$V_n = \left\{x\in E \mid d(x,F) < \frac{1}{n}\right\} = (d(\cdot,F))^{-1}\left(\left(-\infty,\frac{1}{n}\right)\right)$$

for any $n \in N_+$ is a sequence of open sets (each $(-\infty, 1/n)$ is an open set in $\mathbb{R}$).

Furthermore, $\{V_n\}_{n \in N_+}$ is a sequence of open sets that decreases to F; clearly, $V_{n+1} \subset V_n$ for any $n \in N_+$, and $F \subset \bigcap_n V_n$. Conversely, for any $x \in \bigcap_n V_n$, if $x \notin F$ then $d(x, F) > 0$, which implies that $x \notin V_n$ for some $n \in N_+$, a contradiction. Thus, $\bigcap_n V_n \subset F$ and thus $F = \bigcap_n V_n$. By sequential continuity and the finiteness of μ, we can see that

$$\lim_{n \to \infty} \mu(V_n) = \mu(F).$$

By implication, for any $\epsilon > 0$, there exists an $N \in N_+$ such that

$$\mu(V_N \setminus F) = \mu(V_N) - \mu(F) < \epsilon,$$

where the first equality follows because $F \subset V_N$. Denote $V = V_N$.
Finally, we can also see that, for any $x \in V^c$ and $y \in F$,

$$d(x, y) \geq d(x, F) \geq \frac{1}{N},$$

which tells us that the distance between V^c and F is bounded below by a non-zero value. V is an open set that contains the closed set F, so together with the boundedness condition above, Urysohn's lemma for metric spaces tells us that there exists a Lipschitz continuous function $f : E \to \mathbb{R}$ such that

$$f(x) \in \begin{cases} 0 & \text{if } x \notin V \\ [0, 1] & \text{if } x \in V \setminus F \\ 1 & \text{if } x \in F \end{cases}$$

for any $x \in E$. This function satisfies $I_F \leq f \leq I_V$ on E, so it follows that

$$\mu_n(F) \leq \int_E f d\mu_n \leq \mu_n(V)$$

for any $n \in N_+$. Because f is a bounded Lipschitz continuous function, by assumption

$$\lim_{n \to \infty} \int_E f d\mu_n = \int_E f d\mu,$$

so we have the inequality

$$\limsup_{n \to \infty} \mu_n(F) \leq \int_E f d\mu.$$

Since $f \leq I_V$ on E, by monotonicity we have

$$\int_E f d\mu \leq \mu(V) < \epsilon + \mu(F);$$

putting together all these inequalities yields

$$\limsup_{n\to\infty} \mu_n(F) < \epsilon + \mu(F).$$

This holds for any $\epsilon > 0$, so it follows that

$$\limsup_{n\to\infty} \mu_n(F) \leq \mu(F),$$

as desired.

iii) $\rightarrow$ iv)

Suppose iii) holds. Then, for any open set $V \subset F$,

$$1 = \mu_n(E) = \mu_n(V) + \mu_n(V^c)$$

for any $n \in N_+$, so that $\mu_n(V) = 1 - \mu_n(V^c)$. Since V^c is closed, by ii) we have

$$\limsup_{n\to\infty} \mu_n(V^c) \leq \mu(V^c) = 1 - \mu(V).$$

It follows that

$$1 - \liminf_{n\to\infty} \mu_n(V) \leq 1 - \mu(V),$$

and as such iv) holds.

iv) $\rightarrow$ v)

Now suppose iv) holds. Then, by the same process as above, we can show that iii) holds as well.

Choose any $A \in \mathcal{E}$ such that $\mu(\partial A) = 0$. Since

$$\partial A = \overline{A} \setminus A^o,$$

where $\overline{A}$ and A^o are the closure and interior of A, repsectively, we can see that

$$\mu(\overline{A}) - \mu(A^o) = \mu(\overline{A} \setminus A^o) = \mu(\partial A) = 0.$$

$A^o \subset A \subset \overline{A}$, so this result implies that $\mu(A^o) = \mu(A) = \mu(\overline{A})$.

iii) and iv) show us that

$$\limsup_{n\to\infty} \mu_n(\overline{A}) \leq \mu(\overline{A})$$
$$\liminf_{n\to\infty} \mu_n(A^o) \geq \mu(A^o),$$

and because $\mu(\overline{A}) = \mu(A^o) = \mu(A)$,

$$\begin{aligned}
\limsup_{n\to\infty} \mu_n(A) &\leq \limsup_{n\to\infty} \mu_n(\overline{A}) && (A \subset \overline{A}) \\
&\leq \mu(A) \leq \liminf_{n\to\infty} \mu_n(A^o) \\
&\leq \liminf_{n\to\infty} \mu_n(A^o). && (A^o \subset A)
\end{aligned}$$

By implication,

$$\limsup_{n\to\infty} \mu_n(A) = \liminf_{n\to\infty} \mu_n(A) = \mu(A),$$

and thus v) holds.

v) $\rightarrow$ i)

Suppose that v) holds. Let f be a bounded non-negative continuous function. Then, letting $f(x) < M$ for any $x \in E$ and some $M > 0$,

$$\begin{aligned}
\int_E f d\mu_n &= \int_E \int_{\mathbb{R}} I_{[0,f(x))}(t) dt d\mu_n(x) \\
&= \int_E \int_0^\infty I_{(t,+\infty)}(f(x)) dt d\mu_n(x) \\
&= \int_0^\infty \mu_n(\{f > t\}) dt && \text{(Fubini's Theorem)} \\
&= \int_0^M \mu_n(\{f > t\}) dt && (\{f > t\} = \emptyset \text{ for } t > M)
\end{aligned}$$

for any $n \in N_+$, and likewise for μ.

We now want to show that the set

$$D = \{t \in \mathbb{R} \mid \mu(\partial\{f > t\}) > 0\}$$

is at most countable. It is here that the continuity of f is critical. To illustrate this, suppose that $E = \mathbb{R}$, and consider the discontinuous function $g = \frac{1}{2} I_{[0,1/2]} + I_{(1/2,1]}$. In this case, $\partial\{g > t\} = \{1/2\}$ for any $t \in (1/2, 1)$; if $\mu(\{1/2\}) > 0$, then $\mu(\partial\{g > t\}) > 0$ for any $t \in (1/2, 1)$, or in other words, for uncountably many t. Fortunately, the continuity of f precludes such a case.

By the continuity of f, $\{f > t\} = f^{-1}((t, +\infty))$ is open for any $t \in \mathbb{R}$. On the other hand, since $[t, +\infty) = \mathbb{R} \setminus (-\infty, t)$, where $(-\infty, t)$ is an open set, $[t, +\infty)$ is closed, and therefore $f^{-1}([t, +\infty)) = \{f \geq t\}$ is closed as well by continuity. Thus, $\{f \geq t\}$ is a closed set containing the open set $\{f > t\}$, which implies that the closure of $\{f > t\}$ is contained in $\{f \geq t\}$. Together, these imply that

$$\begin{aligned}
\partial\{f > t\} &= \overline{\{f > t\}} \setminus \{f > t\}^o \\
&= \overline{\{f > t\}} \setminus \{f > t\} && (\{f > t\} \text{ is open}) \\
&\subset \{f \geq t\} \setminus \{f > t\} = \{f = t\}.
\end{aligned}$$

Since μ is a finite measure, there can be at most countably many $t \in \mathbb{R}$ such that $\mu(\{f = t\}) > 0$ (this follows from a similar reasoning as that in lemma 1.9).
Let D' be the set of all real numbers t such that $\mu(\{f = t\}) > 0$. Since $\partial\{f > t\} \subset \{f = t\}$ for any $t \in \mathbb{R}$, we have $(D')^c \subset D^c$, that is, if $\mu(\{f = t\}) = 0$, then $\mu(\partial\{f > t\}) = 0$ as well. This implies that $D \subset D'$, and because D' is countable, so is D.

It follows that D has measure 0 under the Lebesgue measure (since it is countable), and by result v),

$$\lim_{n\to\infty} \mu_n(\{f > t\}) = \mu(\{f > t\})$$

for any $t \in [0, M] \setminus D$. Therefore,

$$\begin{aligned}
\int_E f d\mu_n &= \int_0^M \mu_n(\{f_n > t\}) dt \\
&= \int_{[0,M]\setminus D} \mu_n(\{f_n > t\}) dt
\end{aligned}$$

for any $n \in N_+$ and likewise for μ, and because $\mu_n(\{f_n > t\}) \in [0, 1]$ for any $n \in N_+$ and $[0, M] \setminus D$ has finite Lebesgue measure, by the BCT we have

$$\begin{aligned}
\lim_{n\to\infty} \int_E f d\mu_n &= \lim_{n\to\infty} \int_{[0,M]\setminus D} \mu_n(\{f > t\}) dt \\
&= \int_{[0,M]\setminus D} \mu(\{f > t\}) dt = \int_E f d\mu.
\end{aligned}$$

Now let f be an arbitrary bounded and continuous real valued function. By the above result,

$$\lim_{n\to\infty} \int_E f^{\pm} d\mu_n = \int_E f^{\pm} d\mu,$$

so by the linearity of integration,

$$\lim_{n\to\infty} \int_E f d\mu_n = \lim_{n\to\infty} \int_E f^+ d\mu_n - \lim_{n\to\infty} \int_E f^- d\mu_n$$

$$= \int_E f^+ d\mu - \int_E f^- d\mu = \int_E f d\mu.$$

By definition, $\mu_n \to \mu$ weakly.

We have thus shown that i) to v) are equivalent.
Now consider the case where $(E, \mathcal{E}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Suppose that $\mu_n \to \mu$ weakly. Then, v) holds, so that

$$\lim_{n\to\infty} \mu_n(A) = \mu(A)$$

for any $A \in \mathcal{E}$ such that $\mu(\partial A) = 0$.
Choose any $x \in \mathbb{R}$ such that F is continuous at x. This means that

$$\mu(\{x\}) = F(x) - F(x-) = 0,$$

and because $\partial(-\infty, x] = \{x\}$, we have $\mu(\partial(-\infty, x]) = 0$. By the result above, then,

$$\begin{aligned} \lim_{n\to\infty} F_n(x) &= \lim_{n\to\infty} \mu_n((-\infty, x]) \\ &= \mu((-\infty, x]) = F(x). \end{aligned}$$

Therefore, F_n converges to F at all continuity points of F.

Conversely, suppose that F_n converges to F at all continuity points of F. As before, note that, for any bounded and non-negative continuous function f on $\mathbb{R}$ bounded above by $M > 0$,

$$\int_E f d\mu_n = \int_0^M \mu_n(\{f > t\}) dt = \int_0^M (1 - F_n(t)) dt$$

for any $n \in N_+$, and likewise,

$$\int_E f d\mu = \int_0^M (1 - F(t)) dt.$$

Since F has at most countably many discontinuities, letting the collection of all discontinuity points of F be D, and noting that $F_n(t) \to F(t)$ for any $t \notin D$, by the BCT

$$\begin{aligned} \lim_{n\to\infty} \int_E f d\mu_n &= \lim_{n\to\infty} \int_{[0,M]\setminus D} (1 - F_n(t)) dt \\ &= \int_{[0,M]\setminus D} (1 - F(t)) dt = \int_E f d\mu. \end{aligned}$$

The case for general bounded and continuous real-valued function f on $\mathbb{R}$ follows easily, and by definition $\mu_n \to \mu$ weakly.
Q.E.D.

4.1.1 Application: Uniqueness of Weak Limits

The portmanteau theorem has a variety of implications, including the fact that any sequence of probability measures on a complete metric space converges to at most one weak limit. This can be easily derived as a consequence of the following, more fundamental result:

Lemma 4.2 Let (E,d) be a metric space, τ the metric topology induced by d, and $\mathcal{E}$ the Borel σ-algebra on E generated by τ. Let μ and v be two probability measures on $(E,\mathcal{E})$ such that

$$\int_E f d\mu = \int_E f dv$$

for any bounded and continuous real-valued function f on E. Then, $\mu = v$ on $\mathcal{E}$.

Proof) For any $n \in N_+$, define $\mu_n = \mu$ for any $n \in N_+$. Then, for any bounded and continuous real-valued function f on E, because $\int_E f d\mu_n = \int_E f d\mu$ for any $n \in N_+$,

$$\lim_{n\to\infty} \int_E f d\mu_n = \int_E f d\mu = \int_E f dv$$

and thus $\mu_n \to v$ weakly. By the portmanteau theorem, for any open set A,

$$\liminf_{n\to\infty} \mu_n(A) \geq v(A),$$

but because $\mu_n = \mu$ for any $n \in N_+$, we have

$$\mu(A) = \liminf_{n\to\infty} \mu_n(A) \geq v(A).$$

This holds when the roles of μ and v are replaced as well, so $\mu(A) = v(A)$ for any open set $A \in \tau$. Becauase μ and v are both probability measures and τ is a π-system generating $\mathcal{E}$, it follows that $\mu = v$ on $\mathcal{E}$.
Q.E.D.

The above result tells us that a probability measure μ on a metric space is determined by its values on the set $C_b(E,\mathbb{R})$ (that is, the integrals of any bounded and continuous real-valued function with respect to μ). The uniqueness of weak limits following immediately from this result: letting $\{\mu_n\}_{n\in N_+}$ be a sequence of probability measures on the metric space (E,d) converging weakly to both μ and v, since

$$\int_E f d\mu = \lim_{n\to\infty} \int_E f d\mu_n = \int_E f dv$$

for any $f \in C_b(E,\mathbb{R})$, it follows that $\mu = v$.

4.1.2 Application: Convergence in Probability and in Distribution

The Portmanteau theorem also helps shed light on the relationship between convergence in probability and distribution. It shows that convergence in probability implies convergence in distribution, and that the converse holds only when the limit distribution is degenerate.

Theorem 4.3 Let (E,d) be a separable metric space, τ the metric topology induced by d, and $\mathcal{E}$ the Borel σ-algebra on E generated by τ. Let $\{X_n\}_{n\in N_+}$ be a sequence of random variables in $(E,\mathcal{E})$, and X a random variable in $(E,\mathcal{E})$. The following hold true:

i) If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.

ii) If $X_n \xrightarrow{d} X$ and X is a degenerate random variable, then $X_n \xrightarrow{p} X$.

Proof) Suppose $X_n \xrightarrow{p} X$. Then, for any bounded and continuous real valued function f that is bounded above by $M \in (0,+\infty)$, the sequence $\{|f\circ X_n|^2\}_{n\in N_+}$ is bounded above by M^2; as such,

$$\sup_{n\in N_+} \mathbb{E}\left[|f\circ X_n|^2\right] \leq M^2 < +\infty.$$

By implication, $\{|f\circ X_n|\}_{n\in N_+}$ is L^2-bounded and therefore uniformly integrable. The continuous mapping theorem also tells us that $f\circ X_n \xrightarrow{p} f\circ X$, so together with uniform integrability, we have $f\circ X_n \xrightarrow{L^1} f\circ X$. For any $n\in N_+$,

$$|\mathbb{E}[f\circ X_n] - \mathbb{E}[f\circ X]| \leq \mathbb{E}|f\circ X_n - f\circ X|,$$

where the right hand side goes to 0 as $n\to\infty$, so

$$\lim_{n\to\infty} \mathbb{E}[f\circ X_n] = \mathbb{E}[f\circ X].$$

This holds for any $f\in C_b(E,\mathbb{R})$, so by definition $X_n \xrightarrow{d} X$.

Now suppose $X_n \xrightarrow{d} X$, where X is a degenerate random variable, that is, $X=k$ almost surely on Ω for some $k\in E$. Define $h: E\to\mathbb{R}$ as

$$h(x) = d(x,k)$$

for any $x\in E$. Then, h is continuous on E, so that $h\in\mathcal{E}/\mathcal{B}(\mathbb{R})$; for any $\delta>0$ and $n\in N_+$,

$$\begin{aligned}\mathbb{P}(d(X_n,X)>\delta) &= \mathbb{P}(d(X_n,k)>\delta) && (X=k \text{ almost surely})\\ &\leq \mathbb{P}(d(X_n,k)\geq\delta) = \mu_n\left(h^{-1}([\delta,+\infty))\right),\end{aligned}$$

where μ_n is the distribution of X_n. Since $h^{-1}([\delta,+\infty))$ is a closed set ($[\delta,+\infty)$ is closed and h is continuous), taking $n \to \infty$ on both sides yields, by the Portmanteau theorem,

$$\limsup_{n\to\infty} \mathbb{P}\left(d(X_n, X) > \delta\right) \leq \limsup_{n\to\infty} \mu_n\left(h^{-1}([\delta,+\infty))\right) \leq \mu\left(h^{-1}([\delta,+\infty))\right),$$

where μ is the distribution of X. Taking a look at the quantity on the right hand side,

$$\mu\left(h^{-1}([\delta,+\infty))\right) = \mathbb{P}\left(d(X,k) \geq \delta\right) \leq \mathbb{P}\left(X \neq k\right),$$

where the last inequality follows because $d(x,k) \geq \delta > 0$ implies $x \neq k$ for any $x \in E$. Since $X = k$ almost surely, $\mathbb{P}(X \neq k) = 0$ and we thus have

$$0 \leq \limsup_{n\to\infty} \mathbb{P}\left(d(X_n, X) > \delta\right) \leq 0,$$

which implies

$$\lim_{n\to\infty} \mathbb{P}\left(d(X_n, X) > \delta\right) = 0.$$

This holds for any $\delta > 0$, so

$$X_n \xrightarrow{p} X.$$

Q.E.D.

The above result can be used to show, among other results, that any sequence of complex random variables that converges in distribution is bounded in probability and thus $O_p(1)$:

Theorem 4.4 Let $\{X_n\}_{n\in N_+}$ be a sequence of complex random variables that converges in distribution to some complex random variable X. Then, $X_n = O_p(1)$.

Proof) Let $\{\mu_n\}_{n\in N_+}$ and μ be probability measures on $(\mathbb{C}, \mathcal{B}(\mathbb{C}))$ such that each μ_n is the distribution of X_n and μ the distribution of X. It follows by definition from $X_n \xrightarrow{d} X$ that $\mu_n \to \mu$ weakly.

Choose any $\epsilon > 0$. Note first that the sequence $\{B_k\}_{k\in N_+}$ defined as

$$B_k = \{|X| \geq k\} \in \mathcal{H}$$

for any $k \in N_+$ is decreasing with intersection $B = \bigcap_k B_k = \{|X| = +\infty\} = \emptyset$. It follows from sequential continuity and the finiteness of $\mathbb{P}$ that

$$\lim_{k\to\infty} \mathbb{P}\left(|X| \geq k\right) = \lim_{k\to\infty} \mathbb{P}\left(B_k\right) = \mathbb{P}\left(B\right) = 0.$$

Therefore, there exists an $M \in N_+$ such that

$$\mathbb{P}(|X| \geq M) < \epsilon.$$

Define the set $Z \subset \mathbb{C}$ as

$$Z = \{z \in \mathbb{C} \mid |z| \geq M\}.$$

Z, being the complement of the open set $\{z \in \mathbb{C} \mid |z| < M\}$, is a closed set in Z. By the definition of μ_n and μ,

$$\mu_n(Z) = \mathbb{P}(|X_n| \geq M) \quad \text{and} \quad \mu(Z) = \mathbb{P}(|X| \geq M)$$

for any $n \in N_+$, so by the portmanteau theorem,

$$\limsup_{n\to\infty} \mathbb{P}(|X_n| \geq M) \leq \mathbb{P}(|X| \geq M) < \epsilon.$$

Since

$$\limsup_{n\to\infty} \mathbb{P}(|X_n| \geq M) = \inf_{n\in N_+} \sup_{k\geq n} \mathbb{P}(|X_k| \geq M),$$

we can see from the definition of the infimum that there exists an $N \in N_+$ such that

$$\sup_{n\geq N} \mathbb{P}(|X_n| \geq M) < \epsilon,$$

and as such

$$\mathbb{P}(|X_n| \geq M) < \epsilon$$

for any $n \geq N$. We can find such an $M > 0$ and $N \in N_+$ for any $\epsilon > 0$, so it follows that $X_n = O_p(1)$, that is, $\{X_n\}_{n\in N_+}$ is bounded in probability.

Q.E.D.

4.2 Skorokhod's Representation Theorem

Note that no mention of an underlying probability space or random variables were made during the proof of the Portmanteau theorem. This indicates that we may be able to derive, from the weak convergence of probability measures, results concerning the convergence of the corresponding random variables. This is precisely the content of the next theorem, which allows us to construct an underlying probability space such that weak convergence of the distributions translates into the almost sure convergence of the associated random variables.
We first prove a preliminary result:

Lemma 4.5 Let (E,d) be a separable metric space, τ the metric topology induced by d, and $\mathcal{E}$ the Borel σ-algebra on E generated by τ. Suppose μ is a probability measure on $(E,\mathcal{E})$.
Then, for any $\eta > 0$ there exists a countable base $\mathbb{B}'$ of open balls with radius less than η generating τ such that $\mu(\partial A) = 0$ for any $A \in \mathbb{B}'$.

Proof) Choose $\eta > 0$.

Let $\mathbb{B}$ be the collection of all open balls on E, that is, sets of the form $B_d(x,\epsilon)$ for some $x \in E$ and $\epsilon > 0$. We first show that there are only countably many of these balls around some $x \in E$ whose boundary has positive measure under μ.

For any $x \in E$, consider two open balls $B_d(x,\epsilon)$ and $B_d(x,\delta)$ for $\epsilon, \delta > 0$ such that $\epsilon \neq \delta$. Assuming without loss of generality that $\epsilon > \delta$, there exists a rational number $q \in \mathbb{Q}$ such that $\delta < q < \epsilon$; since the closed ball $\bar{B}_d(x,q)$ is a closed set such that

$$B_d(x,\delta) \subset \bar{B}_d(x,q) \subset B_d(x,\epsilon),$$

we can see that

$$\overline{B_d(x,\delta)} \subset \bar{B}_d(x,q) \subset B_d(x,\epsilon),$$

which implies that $\partial B_d(x,\delta)$, being a subset of the closure of $B_d(x,\delta)$, is also contained in $B_d(x,\epsilon)$. Finally,

$$\partial B_d(x,\epsilon) = \overline{B_d(x,\epsilon)} \setminus B_d(x,\epsilon)$$

tells us that $\partial B_d(x,\delta)$ and $\partial B_d(x,\epsilon)$ are disjoint. In other words, the set

$$\{\partial B_d(x,\epsilon) \mid \epsilon > 0\}$$

is a collection of disjoint measurable sets. Therefore, by the finiteness of μ, there can exist at most countably many elements of the above set that have positive measure under μ.

This allows us to define the countable collection $D_x = \{B_{x,n}\}_{n \in N_+}$ of open balls as

follows: for any $n \in N_+$, if $\mu(\partial B_d(x, \eta \cdot 2^{-n})) = 0$, then we put $B_{x,n} = B_d(x, \eta \cdot 2^{-n})$. On the other hand, if $\mu(\partial B_d(x, \eta \cdot 2^{-n})) > 0$, then choose $\eta \cdot 2^{-n-1} < \epsilon_n < \eta \cdot 2^{-n}$ so that $\mu(\partial B_d(x, \epsilon_n)) = 0$; such a choice is possible because the open interval $(\eta \cdot 2^{-n-1}, \eta \cdot 2^{-n})$ is uncountable and there are at most countably many $\epsilon > 0$ such that $\mu(\partial B_d(x, \epsilon)) > 0$. We then let $B_{x,n} = B_d(x, \epsilon_n)$.
The D_x thus constructed is a collection of open balls with measure 0 boundaries and radii less than η that is strictly decreasing with respect to n.

By the separability of (E, d), there exists a countable set $E_0 \subset E$ sucht that $\overline{E_0} = E$, and define

$$\mathbb{B}' = \bigcup_{x \in E_0} D_x.$$

It can be shown that $\mathbb{B}'$ is a countable base on E that generates the topology τ.

Clearly, $\mathbb{B}'$ is a countable collection of open sets on E, since each D_x is made up of countably many open balls, and E_0 is countable. Furthermore, $\mathbb{B}'$ covers E; to see this, note that, for any $y \in E$, by the denseness of E_0 on E, there exists an $x \in E_0$ such that $d(x, y) < \frac{\eta}{4}$. By the construction of D_x, there exists a ball B centered around x in D_x such that $B_d(x, \eta/4) \subset B \subset B_d(x, \eta/2)$; therefore,

$$y \in B_d(x, \eta/4) \subset B.$$

Since B is an element of $\mathbb{B}'$, $\mathbb{B}'$ covers E.

Now choose any $A \in \tau$ and $y \in A$. Because τ is the metric topology induced by d, there exists an $\epsilon \in (0, \eta)$ such that $y \in B_d(y, \epsilon) \subset A$. Let $N \in N_+$ be chosen so that $\eta \cdot 2^{-N} < \frac{\epsilon}{2}$. Since E_0 is dense in E, there exists an $x \in E_0$ such that $d(x, y) < \eta \cdot 2^{-N-1}$, and there then exists a $\eta \cdot 2^{-N-1} \leq \delta \leq \eta \cdot 2^{-N}$ such that $B_d(x, \delta) \in D_x$. It now follows that $y \in B_d(x, \eta \cdot 2^{-N-1}) \subset B_d(x, \delta)$, and for any $z \in B_d(x, \delta)$,

$$d(y, z) \leq d(x, z) + d(x, y) < \delta + \eta \cdot 2^{-N-1} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

so that $B_d(x, \delta) \subset B_d(y, \epsilon)$. Therefore,

$$y \in B_d(x, \delta) \subset B_d(y, \epsilon) \subset A,$$

and because $B_d(x, \delta) \in \mathbb{B}'$, we can see that $\mathbb{B}'$ is a base on E (let A be the non-empty intersection of two elements of $\mathbb{B}'$), and that $\mathbb{B}'$ generates τ.

Q.E.D.

The representation theorem is stated below:

Theorem 4.6 (Skorokhod's Reprsentation Theorem)
Let (E,d) be a separable metric space, τ the metric topology induced by d, and $\mathcal{E}$ the Borel σ-algebra on E generated by τ. Let $\{\mu_n\}_{n\in N_+}$ be a sequence of probability measures on $(E,\mathcal{E})$ weakly converging to a probability measure μ on $(E,\mathcal{E})$.
There exists a probability space and random variables $\{X_n\}_{n\in N_+}$ and X taking values in $(E,\mathcal{E})$ defined on that probability space such that

i) X_n has distribution μ_n for any $n \in N_+$ and X has distribution μ.

ii) X_n converges almost surely to X.

Proof) The proof is long and technical, but the main idea is simple. Namely, the idea is to partition the space E into smaller and smaller sets on which μ is concentrated (which is possible due to the separability of (E,d)), and define X_n in a manner so that it takes values on the same small set as X. This ensures that X_n converges to X on each of these sets, and thus almost surely.
The proof therefore proceeds by first constructing these small sets, and then constructing $\{X_n\}_{n\in N_+}$ so that each X_n almost surely lies in the same small set as X.

Step 1: Existence and Construction of Small Sets

We first construct, for any $\epsilon > 0$, a finite measurable partition $\{B_0, B_1, \cdots, B_k\}$ of E such that:

$$\begin{cases} \mu(B_0) < \epsilon & \\ \mu(B_i) > 0 & \text{for } 1 \le i \le k \\ \mu(\partial B_i) = 0 & \text{for } 0 \le i \le k \\ \text{diam}(B_i) \le \epsilon & \text{for } 1 \le i \le k, \end{cases}$$

where we define $\text{diam}A = \sup_{x,y\in A} d(x,y)$ for any set $A \subset E$. Note that the diameter of an open ball is equal to 2 times its radius. It is in the sense that the diameter of each B_i is smaller than ϵ that $\{B_0, \cdots, B_k\}$ is a partition of E of "small sets".

By lemma 4.5, the separability of (E,d) implies that, for any $\epsilon > 0$, there exists a countable base $\mathbb{B}' = \{A_n\}_{n\in N_+}$ of open balls with radii less than $\frac{\epsilon}{2}$ that generates τ such that $\mu(\partial A_n) = 0$ for any $n \in N_+$. Since $\mathbb{B}'$ covers E (by the definition of a base), $\bigcup_{i=1}^n A_i \nearrow E$, and by sequential continuity

$$\lim_{n\to\infty} \mu\left(\bigcup_{i=1}^n A_i\right) = \mu(E) = 1.$$

Define $B_1 = A_1$ and

$$B_n = A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i \right)$$

for any $n \geq 2$. Then, $\{B_n\}_{n \in N_+}$ is a disjoint sequence of measurable sets on E such that $\bigcup_{i=1}^{n} B_i \subset \bigcup_{i=1}^{n} A_i$ for any $n \in N_+$, so we have

$$\lim_{n \to \infty} \mu \left(\bigcup_{i=1}^{n} B_i \right) = \lim_{n \to \infty} \mu \left(\bigcup_{i=1}^{n} A_i \right) = 1.$$

For any $\epsilon > 0$, there exists an $N \in N_+$ such that

$$1 - \mu \left(\bigcup_{i=1}^{N} B_i \right) < \epsilon.$$

Let $B'_1, \cdots, B'_k$ be the subcollection of $\{B_1, \cdots, B_N\}$ of sets of positive μ-measure, and let B_0 be the union of $\left(\bigcup_{i=1}^{N} B_i \right)^c$ and the sets in $\{B_1, \cdots, B_N\}$ of measure 0 under μ. Then, $\{B_0, B'_1, \cdots, B'_k\}$ satisfies the desired properties:

- Because the sets in $\{B_0, B'_1, \cdots, B'_k\}$ are all disjoint and has union E, it is a finite measurable partition of E.

- Since the sets of measure 0 in $\{B_1, \cdots, B_N\}$ and $\left(\bigcup_{i=1}^{N} B_i \right)^c$ are disjoint,

$$\mu(B_0) = \mu \left(\left(\bigcup_{i=1}^{N} B_i \right)^c \right) = 1 - \mu \left(\bigcup_{i=1}^{N} B_i \right) < \epsilon.$$

- By design, $\mu(B_i) > 0$ for $1 \leq i \leq k$.

- For any $1 \leq i \leq k$, because $B_i \subset A$ for some $A \in \mathbb{B}'$, where $\mu(\partial A) = 0$ and A has radius less than $\frac{\epsilon}{2}$, it follows that $\mu(\partial B_i) = 0$ and B_i has diameter less than or equal to ϵ.

- Finally, each B_i^c has boundary of measure 0 under μ (since $\partial B_i = \overline{B_i} \cap \overline{B_i^c} = \partial B_i^c$), so that $\mu(\partial B_0) = 0$ as well.

The preceding result now implies that, for any $m \in N_+$, there exists a finite measurable

partition $\{B_0^m, \cdots, B_{k_m}^m\}$ of E such that

$$\begin{cases} \mu(B_0^m) < 2^{-m} & \\ \mu(B_i^m) > 0 & \text{for } 1 \le i \le k_m \\ \mu(\partial B_i^m) = 0 & \text{for } 0 \le i \le k_m \\ \operatorname{diam}(B_i^m) \le 2^{-m} & \text{for } 1 \le i \le k_m, \end{cases}.$$

Since $\mu_n \to \mu$ weakly as $n \to \infty$, and the boundary of each B_i^m has measure 0 under μ, by the Portmanteau theorem we can see that

$$\lim_{n\to\infty} \mu_n(B_i^m) = \mu(B_i^m).$$

Thus, if $\mu(B_i^m) > 0$, then there exists an $N_{m,i} \in N_+$ such that

$$|\mu(B_i^m) - \mu_n(B_i^m)| < 2^{-m} \cdot \mu(B_i^m)$$

for any $n \ge N_{m,i}$, and in particular,

$$\mu_n(B_i^m) > (1 - 2^{-m})\mu(B_i^m)$$

for any $n \ge N_{m,i}$. If $\mu(B_i^m) = 0$, then

$$\mu_n(B_i^m) \ge 0 = (1 - 2^{-m})\mu(B_i^m)$$

for any $n \in N_+$, so we can take $N_{m,i}$ as any natural number in this case. Defining $N_m = \max(N_{m,1}, \cdots, N_{m,k_m}) \in N_+$, it follows that

$$\mu_n(B_i^m) \ge (1 - 2^{-m})\mu(B_i^m)$$

for any $n \ge N_m$ and $0 \le i \le k_m$. Finally, we can choose the sequence $\{N_m\}_{m\in N_+}$ to be a subsequence of N_+.

For each $0 \le i \le k_m$ and $n \in N_+$, define the probability measure $\mu_{n,i}^m$ on $(E, \mathcal{E})$ as

$$\mu_{n,i}^m(A) = \begin{cases} \frac{\mu_n(A \cap B_i^m)}{\mu_n(B_i^m)} & \text{if } \mu_n(B_i^m) > 0 \\ \mu_n(A) & \text{if } \mu_n(B_i^m) = 0 \end{cases}$$

for any $A \in \mathcal{E}$. It is easy to show that $\mu_{n,i}^m$ is indeed a probability measure on $(E, \mathcal{E})$.

Step 2: Constructing the Sequence $\{X_n\}_{n\in N_+}$

Now we are ready to construct the random variables corresponding to $\{\mu_n\}_{n\in N_+}$ and μ.

By Ionescu-Tulcea's theorem for independent random variables, there exists a probability space $(\Omega', \mathcal{H}', \mathbb{P}')$ such that the random variables $X, \zeta, \{Y_n\}_{n\in N_+}, \{Y_{ni}\}_{n,i\in\mathbb{N}}, \{Z_n\}_{n\in N_+}$ are mutually independent and have the following distributions:

- X has distribution μ.
- ζ is uniformly distributed on $[0,1]$.
- For any $n \in N_+$, Y_n has distribution μ_n.
- If $N_m \le n < N_{m+1}$, then for any $0 \le i \le k_m$, Y_{ni} has distribution $\mu^m_{n,i}$.
- For any $n \in N_+$, if $N_m \le n < N_{m+1}$, then

$$\mathbb{P}'(Z_n \in A) = 2^m \sum_{i=0}^{k_m} \mu^m_{n,i}(A)\left[\mu_n(B^m_i) - (1-2^{-m})\mu(B^m_i)\right]$$

  for any $A \in \mathcal{E}$.
  Letting the distribution of Z_n be denoted v_n, it is clear that v_n is a probability measure on $(E, \mathcal{E})$ because it is the linear combination of measures $\mu^m_{n,0}, \cdots, \mu^m_{n,k_m}$ with non-negative coefficients, and

$$v_n(E) = 2^m \sum_{i=0}^{k_m} \left[\mu_n(B^m_i) - (1-2^{-m})\mu(B^m_i)\right] = 2^m\mu_n(E) - 2^m(1-2^{-m})\mu(E) = 1$$

  by finite additivity ($B^m_0, \cdots, B^m_{k_m}$ are disjoint and have union E).

Define the sequence $\{X_n\}_{n\in N_+}$ as follows:

i) $n < N_1$
   In this case, let $X_n = Y_n$. Clearly, X_n has distribution μ_n.

ii) $N_m \le n < N_{m+1}$ **for some** $m \in N_+$
   In this case, we let

$$X_n = \begin{cases} Y_{ni} & \text{if } \zeta < 1-2^{-m} \text{ and } X \in B^m_i \\ Z_n & \text{if } \zeta \ge 1-2^{-m} \end{cases}.$$

   Then, for any $A \in \mathcal{E}$,

$$\{X_n \in A\} = \left(\bigcup_{i=1}^{k_m} \left[\{Y_{ni} \in A\} \cap \{X \in B^m_i\} \cap \{\zeta < 1-2^{-m}\}\right]\right) \cup \left(\{\zeta \ge 1-2^{-m}\} \cap \{Z_n \in A\}\right)$$

so by finite additivity and the independence of all these random variables,

$$\begin{aligned}\mathbb{P}'(X_n \in A) &= (1-2^{-m})\sum_{i=0}^{k_m}\mathbb{P}'(Y_{ni} \in A)\,\mathbb{P}'(X \in B_i^m) + 2^{-m}\cdot\mathbb{P}'(Z_n \in A)\\ &= \sum_{i=0}^{k_m}\Big((1-2^{-m})\mu_{n,i}^m(A)\mu(B_i^m) + \mu_{n,i}^m(A)\left[\mu_n(B_i^m) - (1-2^{-m})\mu(B_i^m)\right]\Big)\\ &= \sum_{i=0}^{k_m}\mu_{n,i}^m(A)\mu_n(B_i^m).\end{aligned}$$

For any $m \in N_+$ and $0 \le i \le k_m$, we can consider two cases: if $\mu_n(B_i^m) = 0$, then

$$\mu_{n,i}^m(A)\mu_n(B_i^m) = 0 = \mu_n(A \cap B_i^m),$$

while if $\mu_n(B_i^m) > 0$, then by definition

$$\mu_{n,i}^m(A)\mu_n(B_i^m) = \mu_n(A \cap B_i^m).$$

Therefore,

$$\mathbb{P}'(X_n \in A) = \sum_{i=0}^{k_m}\mu_{n,i}^m(A)\mu_n(B_i^m) = \sum_{i=0}^{k_m}\mu_n(A \cap B_i^m) = \mu_n(A)$$

by finite additivity and the fact that $E = \bigcup_{i=0}^{k_m} B_i^m$. It follows that each X_n has distribution μ_n.

Step 3: Showing that X_n Converges to X

It remains to show that $\{X_n\}_{n\in N_+}$ converges almost surely to X. To this end, first note that, for any $n \ge N_m$ and $1 \le i \le k_m$, the fact that $\mu_n(B_i^m) \ge (1-2^{-m})\mu(B_i^m)$ and $\mu(B_i^m) > 0$ implies that $\mu_n(B_i^m) > 0$. Thus,

$$\mathbb{P}'(Y_{ni} \in B_i^m) = \frac{\mu_n(B_i^m)}{\mu_n(B_i^m)} = 1.$$

It follows that the set

$$\Omega_0 = \bigcap_m \bigcap_{n=N_m}^{N_{m+1}-1} \bigcap_{i=1}^{k_m}\{Y_{ni} \in B_i^m\}$$

has measure 1 under $\mathbb{P}'$.

Define

$$A_m = \{X \notin B_0^m\} \cap \{\zeta < 1-2^{-m}\} \cap \Omega_0.$$

for any $m \in N_+$, and let

$$A = \liminf_{m\to\infty} A_m = \bigcup_k \left(\bigcap_{m\geq k} A_m \right).$$

For any $n \in N_+$ such that $N_m \leq n < N_{m+1}$, then for any $\omega \in A_m$, $X(\omega) \in B_i^m$ for some $1 \leq i \leq k_m$, and because $\zeta < 1 - 2^{-m}$, it follows that $X_n(\omega) = Y_{ni}(\omega)$. Since $\omega \in \Omega_0$, it follows that $Y_{ni}(\omega) \in B_i^m$. Each B_i^m has diameter less than or equal to 2^{-m}, so we have

$$d(X(\omega), X_n(\omega)) \leq \sup_{x,y\in B_i^m} d(x,y) = \mathrm{diam} B_i^m \leq 2^{-m}.$$

For any $\omega \in A$, there exists a $k \in N_+$ such that $\omega \in \bigcap_{m\geq k} A_m$. For any $\epsilon > 0$, choosing $M \in N_+$ such that $M > k$ and $2^{-M} < \epsilon$, since $\omega \in A_m$ for any $m \geq M$, we can see that

$$d(X_n(\omega), X(\omega)) < \epsilon$$

for any $n \geq N_M$. This holds for any $\omega \in A$ and $\epsilon > 0$, so X_n converges to X pointwise on A.

It remains to be seen that A is a set of measure 1 under $\mathbb{P}$. This can be accomplished by noting that

$$\mathbb{P}'(A_m^c) \leq \mathbb{P}'(X \in B_0^m) + \mathbb{P}'(\zeta \geq 1 - 2^{-m}) + \mathbb{P}'(\Omega_0^c) = 2^{-m+1}.$$

Since

$$A^c = \bigcap_k \left(\bigcup_{m\geq k} A_m^c \right),$$

it follows that, for any $k \in N_+$,

$$\mathbb{P}'(A^c) \leq \mathbb{P}'\left(\bigcup_{m\geq k} A_m^c \right) \leq \sum_{m=k}^{\infty} \mathbb{P}'(A_m^c) = \sum_{m=k}^{\infty} 2^{-m+1}$$

by countable subadditivity. Therefore,

$$\mathbb{P}'(A^c) = \lim_{n\to\infty} \mathbb{P}'(A^c) \leq \lim_{k\to\infty} \sum_{m=k}^{\infty} 2^{-m+1} = 0.$$

Q.E.D.

4.2.1 Application: The Continuous Mapping Theorem

Skorokhod's representation theorem is powerful because it shows us that, for a suitable choice of probability space, weak convergence is as strong as almost sure convergence. We illustrate this power in the next result, which is the continuous mapping theorem for weak convergence:

Theorem 4.7 (Continuous Mapping Theorem)
Let (E,d) be a separable metric space, with corresponding Borel σ-algebra $\mathcal{E}$. Let $\{X_n\}_{n\in N_+}$ be a sequence of random variables taking values in $(E,\mathcal{E})$ that converge in distribution to some random variable X taking values in $(E,\mathcal{E})$.

Let (F,ρ) be another separable metric space with corresponding Borel σ-algebra $\mathcal{F}$. If $f: E \to F$ is a Borel measurable function on E that is continuous μ-a.e., then

$$f \circ X_n \xrightarrow{d} f \circ X.$$

Proof) We show the result by relying on the continuous mapping theorem for almost surely convergent sequences of random variables. First, let μ_n be the distribution of each X_n, and μ that of X. By assumption, $\{\mu_n\}_{n\in N_+}$ converges weakly to μ.

By Skorokhod's representation theorem, there exists a probability space $(\Omega',\mathcal{H}',\mathbb{P}')$, a sequence $\{Y_n\}_{n\in N_+}$ of random variables defined on Ω' taking values in $(E,\mathcal{E})$, and a random variable Y defined on Ω' taking values in $(E,\mathcal{E})$ such that

- Y_n has the distribution μ_n for any $n \in N_+$,
- Y has the distribution μ, and
- $Y_n \xrightarrow{a.s.} Y$.

Denote the set of discontinuity points of f by D; by assumption, $\mu(D) = 0$. Let $\Omega_0 \in \mathcal{H}'$ be an almost sure set on which $Y_n \to Y$ pointwise, and define

$$\Omega_1 = Y^{-1}(D^c) \in \mathcal{H}',$$

where the inclusion holds because $D \in \mathcal{E}$. Because $\mu(D) = 0$, it follows that

$$\mathbb{P}'(\Omega_1) = \mathbb{P}'(Y \in D^c) = \mu(D^c) = 1,$$

and as such, $\mathbb{P}'(\Omega_0 \cap \Omega_1) = 1$. Since f is continuous at all points on D^c, it follows that, for any $\omega \in \Omega_1 \cap \Omega_0$,

$$\lim_{n\to\infty} f(Y_n(\omega)) = f(Y(\omega)),$$

since $Y(\omega) \in D^c$. By definition, $f \circ Y_n \xrightarrow{a.s.} f \circ Y$, and because almost sure convergence implies convergence in probabilty, which in turn implies convergence in distribution

(due to the completeness and separability of (F, ρ)), it follows that

$$f \circ Y_n \xrightarrow{d} f \circ Y.$$

The distribution of each $f \circ Y_n$ is the pushforward measure $\mu_n \circ f^{-1}$, and likewise, the distribution of $f \circ Y$ is $\mu \circ f^{-1}$; we have seen that

$$\mu_n \circ f^{-1} \to \mu \circ f^{-1}$$

weakly. Because the distribution of each $f \circ X_n$ is $\mu_n \circ f^{-1}$ and that of $f \circ X$ is $\mu \circ f^{-1}$, this implies that

$$f \circ X_n \xrightarrow{d} f \circ X.$$

Q.E.D.

4.3 Slutsky's Theorem

This section concerns another important connection between convergence in probability and distribution. In a sense, it shows that convergence in distribution dominates convergence in probability; if two variables converge, one in distribution and one in probabilty, then any continuous function of the two variables converges in distribution.
We first state the first variant of the theorem:

Theorem 4.8 (Slutsky's Thereom I)
Let (E,d) be a separable metric space, with corresponding Borel σ-algebra $\mathcal{E}$. Let $\{X_n\}_{n\in N_+}$, $\{Y_n\}_{n\in N_+}$ be sequences of random variables taking values in $(E,\mathcal{E})$.

Suppose that $\{X_n\}_{n\in N_+}$ converges in distribution to some random variable X taking values in $(E,\mathcal{E})$, and that

$$d(X_n,Y_n)\xrightarrow{p}0.$$

Then, $\{Y_n\}_{n\in N_+}$ also converges in distribution to X.

Proof) Choose any $f\in C_b(E,\mathbb{R})$ that is Lipschitz continuous with Lipschitz constant equal to $L>0$, that is,

$$|f(x)-f(y)|\le L\cdot d(x,y)$$

for any $x,y\in E$, and suppose that f is bounded above by $M\in(0,+\infty)$. For any $\epsilon>0$, by Lipschitz continuity

$$|f(x)-f(y)|<\epsilon$$

for any $x,y\in E$ such that $d(x,y)\le\frac{\epsilon}{L}=\delta$.

For any $n\in N_+$, we now have

$$\begin{aligned}|\mathbb{E}[f\circ X_n]-\mathbb{E}[f\circ Y_n]|&\le\mathbb{E}|f\circ X_n-f\circ Y_n|\\&\le\mathbb{E}\left[|f\circ X_n-f\circ Y_n|\cdot I_{\{d(X_n,Y_n)\le\delta\}}\right]+\mathbb{E}\left[|f\circ X_n-f\circ Y_n|\cdot I_{\{d(X_n,Y_n)>\delta\}}\right]\\&\le\epsilon\mathbb{P}(d(X_n,Y_n)\le\delta)+2M\cdot\mathbb{P}(d(X_n,Y_n)>\delta)\\&\le\epsilon+2M\cdot\mathbb{P}(d(X_n,Y_n)>\delta).\end{aligned}$$

Therefore,

$$|\mathbb{E}[f\circ Y_n]-\mathbb{E}[f\circ X]|\le|\mathbb{E}[f\circ X_n]-\mathbb{E}[f\circ X]|+\epsilon+2M\cdot\mathbb{P}(d(X_n,Y_n)>\delta).$$

By definition,

$$\lim_{n\to\infty} \mathbb{P}\left(d(X_n, Y_n) > \delta\right) = 0,$$

and because $X_n \xrightarrow{d} X$,

$$\lim_{n\to\infty} \left|\mathbb{E}\left[f \circ X_n\right] - \mathbb{E}\left[f \circ X\right]\right| = 0.$$

Taking $n \to \infty$ on both sides thus yields

$$\limsup_{n\to\infty} \left|\mathbb{E}\left[f \circ Y_n\right] - \mathbb{E}\left[f \circ X\right]\right| \leq \epsilon.$$

This holds for any $\epsilon > 0$, so

$$\lim_{n\to\infty} \mathbb{E}\left[f \circ Y_n\right] = \mathbb{E}\left[f \circ X\right].$$

This in turn holds for any Lipschitz continuous $f \in C_b(E, \mathbb{R})$, so by the Portmanteau theorem,

$$Y_n \xrightarrow{d} X.$$

Q.E.D.

The next result applies the previous theorem to make a claim about a pair of random variables.

Theorem 4.9 (Slutsky's Thereom II)
Let (E, d) and (F, ρ) be separable metric spaces with corresponding Borel σ-algebras $\mathcal{E}$ and $\mathcal{F}$. Let $\{X_n\}_{n\in N_+}$ and $\{Y_n\}_{n\in N_+}$ be sequences of random variables taking values in $(E, \mathcal{E})$ and $(F, \mathcal{F})$, respectively.

Suppose that $\{X_n\}_{n\in N_+}$ converges in distribution to some random variable X in $(E, \mathcal{E})$, and that $\{Y_n\}_{n\in N_+}$ converges in distribution to some random variable Y in $(F, \mathcal{F})$.

If Y is a degenerate random variable, then the sequence $\{(X_n, Y_n)\}_{n\in N_+}$ of random variables in $(E \times F, \mathcal{E} \bigotimes \mathcal{F})$ satisfies

$$(X_n, Y_n) \xrightarrow{d} (X, Y).$$

Moreover, for any continuous function g on $E \times F$,

$$g(X_n, Y_n) \xrightarrow{d} g(X, Y).$$

Proof) Because (E, d) and (F, ρ) are separable, letting τ and s be their metric topologies induced by d and ρ, the product topology $\tau \times s$ generates the product σ-algebra $\mathcal{E} \bigotimes \mathcal{F}$.

Furthermore, we also showed that the product metric $d \times \rho$ metrizes $\tau \times s$, so that $\mathcal{E} \bigotimes \mathcal{F}$ is precisely the Borel σ-algebra associated with the metric space $(E \times F, d \times \rho)$. Because the individual metric spaces (E, d) and (F, ρ) are separable, so is $(E \times F, d \times \rho)$; this means that the theorems concerning convergence in distribution we have studied so far also apply to sequences of probability measures on the product space $(E \times F, \mathcal{E} \bigotimes \mathcal{F})$.

Note first that, for any $n \in N_+$,

$$(d \times \rho)\left[(X_n, Y), (X_n, Y)\right] = \max(d(X_n, X_n), \rho(Y_n, Y)) = \max(0, \rho(Y_n, Y)) = \rho(Y_n, Y),$$

and because

$$\rho(Y_n, Y) \xrightarrow{p} 0$$

by the assumption that $Y_n \xrightarrow{p} Y$, it follows that

$$(d \times \rho)\left[(X_n, Y), (X_n, Y)\right] \xrightarrow{p} 0$$

as well.

Because Y is a degenerate random variable, $Y = k$ almost surely for some $k \in F$. It then follows that, for any $f \in C_b(E \times F, \mathbb{R})$,

$$\mathbb{E}\left[f \circ (X_n, Y)\right] = \mathbb{E}\left[f \circ (X_n, k)\right]$$

for any $n \in N_+$. Defining the function $h : E \to \mathbb{R}$ as $h(x) = f(x, k)$ for any $x \in E$, h is continuous because the sections of continuous functions are continuous, and it is bounded because f is. Therefore, by the definition of convergence in distribution,

$$\begin{aligned}\lim_{n \to \infty} \mathbb{E}\left[f \circ (X_n, Y)\right] &= \lim_{n \to \infty} \mathbb{E}\left[f \circ (X_n, k)\right] \\ &= \lim_{n \to \infty} \mathbb{E}\left[h \circ X_n\right] = \mathbb{E}\left[h \circ X\right] = \mathbb{E}\left[f \circ (X, k)\right] = \mathbb{E}\left[f \circ (X, Y)\right].\end{aligned}$$

This holds for any $f \in C_b(E \times F, \mathbb{R})$, so by definition

$$(X_n, Y) \xrightarrow{d} (X, Y).$$

Finally, by the first version of Slutsky's theorem, we can now see that

$$(X_n, Y_n) \xrightarrow{d} (X, Y).$$

Finally, letting g be a function on $E \times F$ that is continuous with respect to $d \times \rho$, by the continuous mapping theorem we have

$$g(X_n, Y_n) \xrightarrow{d} (X, Y).$$

Q.E.D.

4.3.1 Application: Convergence of Random Vectors and Matrices

The following are some applications of Slutsky's theorem to random vectors on euclidean spaces. Let $\{X_n\}_{n\in N_+}$, $\{B_n\}_{n\in N_+}$ be sequences of k-dimensional random vectors, and $\{A_n\}_{n\in N_+}$ a sequence of $m\times k$ random matrices. Suppose that $X_n \xrightarrow{d} X$ for some k-dimensional random vector X, and that $A_n \xrightarrow{p} A$, $B_n \xrightarrow{p} B$ for some non-random $A\in\mathbb{R}^{m\times k}$ and $B\in\mathbb{R}^k$.

Convergence of Sums and Products

Denoting by d_k the euclidean metric on $\mathbb{R}^k$ and $\rho_{m\times k}$ the metric induced by the trace norm on $\mathbb{R}^{m\times k}$, define the functions $f:\mathbb{R}^k\times\mathbb{R}^k\to\mathbb{R}^k$ and $g:\mathbb{R}^{m\times k}\times\mathbb{R}^k\to\mathbb{R}^k$ as

$$f(x,y)=x+y \quad \text{and} \quad g(M,x)=Mx$$

for any $x,y\in\mathbb{R}^k$ and $M\in\mathbb{R}^{m\times k}$. Both these functions are continuous with respect to the product metrics on their respective domains: starting with f, for any $x,y,z,w\in\mathbb{R}^k$,

$$\begin{aligned}|f(x,y)-f(z,w)| &= |x+y-(z+w)| \le |x-z|+|y-w| \\ &\le 2\cdot\max(|x-z|,|y-w|) = 2\cdot(d_k\times d_k)((x,y),(z,w)),\end{aligned}$$

so that f is actually Lipschitz continuous with Lipschitz constant equal to 2. Similarly, for any $M,S\in\mathbb{R}^{m\times k}$ and $x,y\in\mathbb{R}^k$,

$$\begin{aligned}|g(M,x)-g(S,y)| &= |Mx-Sy| = |Mx-My+My-Sy| \\ &\le \|M\||x-y|+\|M-S\|\cdot|y|.\end{aligned}$$

For any $\epsilon>0$, define

$$\delta=\min\left(\frac{\epsilon}{\|M\|+|x|+1},1\right)>0.$$

For any $(S,y)\in\mathbb{R}^{m\times k}\times\mathbb{R}^k$ such that

$$(\rho_{m\times k}\times d_k)((M,x),(S,y))=\max(\|M-S\|,|x-y|)<\delta,$$

we can see that

$$|y|\le|x-y|+|x|<1+|x|$$

and

$$|g(M,x)-g(S,y)|\le\|M\||x-y|+\|M-S\|\cdot|y|\le\delta\left[\|M\|+|x|+1\right]<\epsilon.$$

This holds for any $\epsilon > 0$, so g is continuous at (M, x), and indeed on $\mathbb{R}^{m\times k} \times \mathbb{R}^k$.

In light of the continuity results above, by Slutsky's theorem,

$$A_n X_n + B_n \xrightarrow{d} AX + B.$$

Convergence of Quadratic Forms

Another popular application concerns quadratic forms. Suppose $m = k$, and note that the mapping $h : \mathbb{R}^{k\times k} \times \mathbb{R}^k \to \mathbb{R}$ defined as

$$h(M, x) = x'Mx$$

for any $M \in \mathbb{R}^{k\times k}$ and $x \in \mathbb{R}^k$ is continuous with respect to the product metric $\rho_{k\times k} \times d_k$. To see this, choose any $M \in \mathbb{R}^{k\times k}$ and $x \in \mathbb{R}^k$. For any $\epsilon > 0$, define

$$\delta = \min\left(\frac{\epsilon}{|x|\|M\| + |x|(|x|+1) + (\|M\|+1)(|x|+1)}, 1\right) > 0.$$

Then, for any $S \in \mathbb{R}^{k\times k}$ and $y \in \mathbb{R}^k$ such that

$$(\rho_{k\times k} \times d_k)((M, x), (S, y)) = \max(\|M - S\|, |x - y|) < \delta,$$

we can see that

$$\begin{aligned} |y| &\leq |x - y| + |x| < 1 + |x| \\ \|S\| &\leq \|M - S\| + \|M\| < 1 + \|M\| \end{aligned}$$

and

$$\begin{aligned} |h(M, x) - h(S, y)| = |x'Mx - y'Sy| &= |x'Mx - x'My + x'My - x'Sy + x'Sy - y'Sy| \\ &\leq |x|\|M\||x - y| + |x|\|M - S\||y| + |x - y|\|S\||y| \\ &\leq [|x|\|M\| + |x|(|x|+1) + (\|M\|+1)(|x|+1)]\,\delta < \epsilon, \end{aligned}$$

where the last inequality follows by design. This holds for any $\epsilon > 0$, so by definition h is continuous at (M, x) and indeed on $\mathbb{R}^{k\times k} \times \mathbb{R}^k$. By Slutsky's theorem, we can now see that

$$X_n' A_n X_n \xrightarrow{d} X'AX.$$

4.4 Tightness and Prohorov's Theorem

In this section we state and prove a theorem that allows us to determine whether a given sequence or collection of probability measures converges weakly. This theorem forms the cornerstone in both the usual central limit theorem (via the continuity theorem, proved below) and the functional central limit theorem.

A central concept is the relative compactness of probability measures. Let (E,d) be an arbitrary metric space and $\mathcal{E}$ the associated Borel σ-algebra. A collection Π of probability measures on $(E,\mathcal{E})$ is said to be relatively compact if, for any sequence $\{\mu_n\}_{n\in N_+}$ in Π there exists a subsequence $\{\mu_{n_k}\}_{k\in N_+}$ of $\{\mu_n\}_{n\in N_+}$ that converges weakly to some probability measure μ on $(E,\mathcal{E})$, not necessarily in Π. This is similar to the characterization of relative compactness in a metric space, in which a set is relatively compact if and only if any sequence in that set has a convergent subsequence. However, the relative compactness of measures is not quite the same as the topological concept of the same name, since we have not defined any metric or topology on the space of probability measures.
Relative compactness comes in handy because of the following result:

Lemma 4.10 Let (E,d) be a metric space with associated Borel σ-algebra $\mathcal{E}$. Suppose $\{\mu_n\}_{n\in N_+}$ is a relatively compact sequence of probability measures on $(E,\mathcal{E})$. In addition, assume that every weakly convergent subsequence of $\{\mu_n\}_{n\in N_+}$ is has the same weak limit μ. Then, $\{\mu_n\}_{n\in N_+}$ itself converges weakly to μ.

Proof) Suppose that $\{\mu_n\}_{n\in N_+}$ does not converge weakly to μ. Then, there exists a bounded and continuous real valued function f such that $\{\int_E f d\mu_n\}_{n\in N_+}$ does not converge to $\int_E f d\mu$. By definition, there exists an $\epsilon > 0$ such that, for any $n \in N_+$ there exists an $N \geq n$ such that

$$\left| \int_E f d\mu_N - \int_E f d\mu \right| \geq \epsilon.$$

As usual, we can construct a subsequence $\{\int_E f d\mu_{n_k}\}_{k\in N_+}$ of $\{\int_E f d\mu_n\}_{n\in N_+}$ such that

$$\left| \int_E f d\mu_{n_k} - \int_E f d\mu \right| \geq \epsilon$$

for any $k \in N_+$. Since $\{\mu_{n_k}\}_{k\in N_+}$ is a sequence in the set $\Pi = \{\mu_n \mid n \in N_+\}$, which is relatively compact, by definition there exists a subsequence $\{\mu_{n_{k_m}}\}_{m\in N_+}$ of $\{\mu_{n_k}\}_{k\in N_+}$ that is weakly convergent. $\{\mu_{n_{k_m}}\}_{m\in N_+}$ is itself a subsequence of $\{\mu_n\}_{n\in N_+}$, so by assumption, it converges weakly to μ. As such,

$$\lim_{m\to\infty} \int_E f d\mu_{n_{k_m}} = \int_E f d\mu.$$

However, the fact that

$$\left| \int_E f d\mu_{n_{k_m}} - \int_E f d\mu \right| \geq \epsilon$$

for any $m \in N_+$ contradicts the above convergence result. It follows that $\{\mu_n\}_{n \in N_+}$ should converge weakly to μ.

Q.E.D.

Therefore, if we can establish the two results above, we can easily show that a given sequence of probability measures converges weakly. It is for this reason that we wish to find sufficient conditions that ensure relative compactness. The main concern here is finding, for any sequence in a collection Π of probability measures, a subsequence that converges weakly to a **probability** measure. In order to ensure that the weak limit is a probability measure, we require a condition that precludes any mass from escaping the probability measures in the limit. This condition is called tightness, and the formal definition is given below.

A collection of probability measures Π on $(E, \mathcal{E})$ is said to be tight if, for any $\epsilon > 0$, there exists a compact set K in E such that

$$\mu(K^c) < \epsilon$$

for any $\mu \in \Pi$. In other words, there exists a compact set such that the mass of each measure in Π on the set can be made arbitrarily close to 1. This is the sense in which mass is precluded from escaping in the limit.

An easy result that follows from the definition is that a probability measure on a metric space that is both complete and separable is itself tight. We can further extend this so that any relatively compact collection of probability measures on a complete and separable metric space is tight; it turns out that this is the necessity part of Prohorov's theorem.

Theorem 3.10 (Prohorov's Theorem: Necessity)
Let (E, d) be a complete and separable metric space, τ the topology induced by d and $\mathcal{E}$ the Borel σ-algebra generated by τ. For any collection of probability measures Π on $(E, \mathcal{E})$, Π is tight if it is relatively compact.

Proof) The three properties of relative compactness, separability and completeness come into play one by one, and the proof can be divided into steps in which each property is used.

Step 1: Relative Compactness

Let $\{G_n\}_{n \in N_+}$ be a sequence of open sets increasing to E. The relative compactness of Π implies that, for any $\epsilon > 0$, there exists an $N \in N_+$ such that $\mu(G_N) \geq 1 - \epsilon$ for any $\mu \in \Pi$.

To see this, suppose that, for any $n \in N_+$, there exists a $\mu_n \in \Pi$ such that $\mu_n(G_n) < 1-\epsilon$. Then, $\{\mu_n\}_{n\in N_+}$ is a sequence in Π and, by relative compactness, it has a weakly convergent subsequence $\{\mu_{n_k}\}_{k\in N_+}$. Letting v be the weak limit of this subsequence, by the Portmanteau theorem we have

$$v(G_n) \leq \liminf_{k\to\infty} \mu_{n_k}(G_n)$$

for any $n \in N_+$. For $m \in N_+$ such that $n_m > n$, we have

$$\mu_{n_m}(G_n) \leq \mu_{n_m}(G_{n_m})$$

because $G_n \subset G_{n_m}$, so that

$$\inf_{m\geq k} \mu_{n_m}(G_n) \leq \inf_{m\geq k} \mu_{n_m}(G_{n_m})$$

for any $k \in N_+$ and thus

$$\liminf_{k\to\infty} \mu_{n_k}(G_n) = \sup_{k\in N_+} \left(\inf_{m\geq k} \mu_{n_m}(G_n) \right) \leq \liminf_{k\to\infty} \mu_{n_k}(G_{n_k}).$$

Finally, $\mu_{n_k}(G_{n_k}) < 1-\epsilon$ for any $k \in N_+$, so

$$v(G_n) \leq \liminf_{k\to\infty} \mu_{n_k}(G_{n_k}) \leq 1-\epsilon.$$

This holds for any $n \in N_+$, so by sequential continuity

$$1 = v(E) = \lim_{n\to\infty} v(G_n) \leq 1-\epsilon,$$

which is a contradiction. Therefore, there must exist an $N \in N_+$ such that

$$\mu(G_N) \geq 1-\epsilon \quad \text{for any } \mu \in \Pi.$$

Indeed, because $\{G_n\}_{n\in N_+}$ is an increasing sequence of open sets,

$$\mu(G_n) \geq 1-\epsilon \quad \text{for any } n \geq N \text{ and } \mu \in \Pi.$$

Step 2: Separability

Now we will construct a sequence $\{G_n\}_{n\in N_+}$ of open sets increasing to E using the separability of (E,d). Throughout, we fix $\epsilon > 0$.

Since (E,d) is separable, there exists a countable set E^0 that is dense in E. For any $k \in N_+$, letting $\mathbb{B}_k = \{A_{kn}\}_{n\in N_+}$ be the collection of all open balls centered at some

point in E^0 with radius $\frac{1}{k}$, $\mathbb{B}_k$ covers E. To see this, let $x \in E$ and note that, by the denseness of E^0 in E, there exists some $y \in E^0$ such that $d(x,y) < \frac{1}{k}$, which implies that $x \in B_d(y, 1/k)$, where $B_d(y,1/k) \in \mathbb{B}_k$.

Defining

$$G_n = \bigcup_{i=1}^{n} A_{ki}$$

for any $n \in N_+$, $\{G_n\}_{n \in N_+}$ is now a sequence of open sets in E that increases to E. By the preceding result, there exists an $N_k \in N_+$ such that

$$\mu\left(\bigcup_{i=1}^{N_k} A_{ki}\right) > 1 - \frac{\epsilon}{2^k},$$

or equivalently,

$$\mu\left(\bigcap_{i=1}^{N_k} A_{ki}^c\right) < \frac{\epsilon}{2^k}$$

for any $\mu \in \Pi$.

Now define

$$B_k = \bigcap_{i=1}^{N_k} A_{ki}^c \text{ for any } k \in N_+ \quad \text{and} \quad B = \bigcup_k B_k.$$

Then, we can see that

$$\mu(B) \leq \sum_{k=1}^{\infty} \mu(B_k) \leq \epsilon$$

by countable subadditivity.

Finally, defining $K = \overline{B^c}$, because $K^c \subset B$, we have $\mu(K^c) \leq \epsilon$. It will now follow that Π is tight if we can show that K is compact. To do so, we show that B^c is relatively compact (in the topological sense); by definition, this means that K, the closure of B^c, is a compact set.

Step 3: Completeness

Recall that B^c is relatively compact (in the topological sense) if any sequence in B^c has convergent subsequence. Let $\{x_n\}_{n \in N_+}$ be a sequence in

$$B^c = \bigcap_k \bigcup_{i=1}^{N_k} A_{ki}.$$

Put $\{x_{n,0}\}_{n\in N_+} = \{x_n\}_{n\in N_+}$, and suppose that we have constructed a subsequence $\{x_{n,k-1}\}_{n\in N_+}$ of $\{x_n\}_{n\in N_+}$ for some $k \geq 1$. Then, since $\{x_{n,k-1}\}_{n\in N_+}$ lies in the set $\bigcup_{i=1}^{N_k} A_{ki}$, there must exist an $1 \leq i \leq N_k$ such that A_{ki} contains infinitely many elements of $\{x_{n,k-1}\}_{n\in N_+}$. Let $\{x_{n,k}\}_{n\in N_+}$ collect these elements of $\{x_{n,k-1}\}_{n\in N_+}$; $\{x_{n,k}\}_{n\in N_+}$ is a subsequence of $\{x_{n,k-1}\}_{n\in N_+}$ and by extension $\{x_n\}_{n\in N_+}$. Moreover, because each $x_{n,k} \in A_{ki}$ and A_{ki} has radius smaller than $\frac{1}{k}$, we have

$$d(x_{n,k}, x_{m,k}) < \frac{1}{k}$$

for any $n, m \in N_+$.
Now consider the sequence $\{x_{k,k}\}_{k\in N_+}$ of elements of E. We can immediately tell this is a subsequence of $\{x_n\}_{n\in N_+}$, since, for any $k \in N_+$, $\{x_{n,k+1}\}_{n\in N_+}$ is a subsequence of $\{x_{n,k}\}_{n\in N_+}$ and thus $x_{k+1,k+1}$ is at least as far along $\{x_n\}_{n\in N_+}$ as $x_{k+1,k}$, whcih tells us that $x_{k+1,k+1}$ is further along $\{x_n\}_{n\in N_+}$ than $x_{k,k}$.
In addition, for any $k, m \in N_+$, we saw above that

$$d(x_{k,k}, x_{m,m}) < \frac{1}{\min(k,m)}.$$

Thus, as $k, m \to \infty$, the quantity on the left hand side converges to 0.
We have seen that $\{x_{k,k}\}_{k\in N_+}$ is a subsequence of $\{x_n\}_{n\in N_+}$ that is Cauchy with respect to the metric d. By the completeness of (E, d), it converges to some $x \in E$, which completes our proof.

Q.E.D.

Note that the above theorem holds when $\Pi = \{\mu\}$ for some probability measure μ on $(E, \mathcal{E})$ becauase Π is trivially relatively compact in this case. Therefore, the theorem shows that every singleton is tight if the underlying metric space is complete and separable.

Prohorov's theorem furnishes sufficient and necessary conditions for a collection of probability measures to be relatively compact, and it is the main result of this section; we proved the necessity part above. The surprising result is that tightness by itself is sufficient and (under some additional conditions) necessary for relative compactness. First, we present a preliminary result concerning the convergence of functions defined on a countable space.

Lemma 4.12 Let E be a countable set and $\{f_n\}_{n\in N_+}$ a sequence of real vector or complex valued pointwise bounded functions on E. Then, there exists a subsequence $\{f_{n_k}\}_{k\in N_+}$ of $\{f_n\}_{n\in N_+}$ that converges pointwise.

Proof) Let $F = \mathbb{R}^n$ or $\mathbb{C}$.
Let $\{x_n\}_{n\in N_+}$ be the elements of the countable set E arranged into a sequence. Because $\{f_n(x_1)\}_{n\in N_+}$ is a bounded sequence, by the Bolzano-Weierstrass theorem there exists

a subsequence $\{f_{n,1}\}_{n\in N_+}$ of $\{f_n\}_{n\in N_+}$ such that $\{f_{n,1}(x_1)\}_{n\in N_+}$ converges pointwise to some $f_{x_1} \in F$.
Suppose, for some $k \geq 1$, that we have found a subsequence $\{f_{n,k}\}_{n\in N_+}$ of $\{f_n\}_{n\in N_+}$ such that $\{f_{n,k}(x_i)\}_{n\in N_+}$ converges to some $f_{x_i} \in F$ for $1 \leq i \leq k$. Again, $\{f_{n,k}(x_{k+1})\}_{n\in N_+}$ is a bounded sequence, so there exists a subsequence $\{f_{n,k+1}\}_{n\in N_+}$ of $\{f_{n,k}\}_{n\in N_+}$ and by extension $\{f_n\}_{n\in N_+}$ such that

$$f_{n,k+1}(x_{k+1}) \to f_{x_{k+1}}$$

as $n \to \infty$ for some $f_{x_{k+1}} \in F$.
Note that, for any $k \in N_+$ and $1 \leq i \leq k$,

$$\lim_{m\to\infty} f_{m,k}(x_i) = f_{x_i}$$

and $\{f_{m,k}(x_i)\}_{m\in N_+}$ is a subsequence of $\{f_{m,j}(x_i)\}_{m\in N_+}$ for any $1 \leq j \leq k$.
Now consider the sequence $\{f_{k,k}\}_{k\in N_+}$ of functions. $\{f_{k,k}\}_{k\in N_+}$ is a subsequence of $\{f_n\}_{n\in N_+}$ because, for any $k \in N_+$, $\{f_{n,k+1}\}_{n\in N_+}$ is a subsequence of $\{f_{n,k}\}_{n\in N_+}$ and thus $f_{k+1,k+1}$ is at least as far along $\{f_n\}_{n\in N_+}$ as $f_{k+1,k}$, which implies that $f_{k+1,k+1}$ is further along $\{f_n\}_{n\in N_+}$ than $f_{k,k}$.

It remains to see whether $\{f_{k,k}\}_{k\in N_+}$ converges pointwise. For any $\epsilon > 0$ and $i \in N_+$,

$$\lim_{m\to\infty} f_{m,i}(x_i) = f_{x_i},$$

so there exists an $N \in N_+$ such that $N \geq i$ and

$$|f_{m,i}(x_i) - f_{x_i}| < \epsilon$$

for any $m \geq N$.
Choose any $k \geq N$. Because $\{f_{m,k}(x_i)\}_{m\in N_+}$ is a subsequence of $\{f_{m,i}(x_i)\}_{m\in N_+}$, there exists an $m \in N_+$ such that $f_{k,k}(x_i) = f_{m,i}(x_i)$. $f_{k,k}(x_i)$ is the kth element of $\{f_{m,k}(x_i)\}_{m\in N_+}$, so $m \geq k \geq N$, which implies that

$$|f_{k,k}(x_i) - f_{x_i}| = |f_{m,i}(x_i) - f_{x_i}| < \epsilon$$

by the above result.
This holds for any $\epsilon > 0$, so by definition the sequence $\{f_{k,k}(x_i)\}_{k\in N_+}$ converges to f_{x_i}.
Define $f : E \to F$ as

$$f(x_i) = f_{x_i}$$

for any $i \in N_+$. Then, the above result shows that $\{f_{k,k}\}_{k\in N_+}$ is a subsequence of $\{f_n\}_{n\in N_+}$ that converges pointwise to f.

Q.E.D.

We now state and prove the sufficiency part of Prohorov's theorem below:

Theorem 4.13 (Prohorov's Theorem: Sufficiency)
Let (E,d) be a metric space, τ the metric topology induced by d, and $\mathcal{E}$ the Borel σ-algebra on E generated by τ. Let Π be a collection of probability measures on $(E,\mathcal{E})$.
If Π is tight, then it is relatively compact.

Proof) Interestingly, the sufficiency part holds for metric spaces that are not necessarily complete or separable; it is shown below that, instead of separability on the metric space, we can make use of the separability of compact sets to obtain the desired result.

The proof proceeds roughly as follows. Choosing an arbitrary sequence in Π, we first construct a well-behaved countable collection of Borel sets $\mathcal{F}$ and find a subsequence of that sequence that converges pointwise on that countable collection. Using the pointwise limit above, we construct an outer measure on E in a similar manner to the Riesz representation theorem. It can then be shown that $\mathcal{E}$ is exactly the collection of all outer measurable sets, and as such, by Caratheodory's restriction theorem, the restriction of that outer measure to $\mathcal{E}$ defines a measure μ on $(E,\mathcal{E})$. Finally, the choice of $\mathcal{F}$ allows us to conclude that the pointwise convergent subsequence found above converges weakly to μ, and that μ is a probability measure.

We now detail each of the above procedures. Initially, let $\{\mu_n\}_{n\in N_+}$ be a sequence in the tight collection Π of probability measures on $(E,\mathcal{E})$.

Step 1: Constructing $\mathcal{F}$

By the tightness of Π, for any $n \in N_+$ there exists a compact set K_n such that

$$\mu(K_n) > 1 - \frac{1}{n}$$

for any $\mu \in \Pi$; without loss of generality, we can choose K_n so that $\{K_n\}_{n\in N_+}$ is an increasing sequence of compact subsets of E. Define $K = \bigcup_n K_n \subset E$. We show that K is covered by a countable collection of open balls.

For any $q \in \mathbb{Q}$ and $n \in N_+$, note that the collection $\{B_d(x,q)\}_{x\in K_n}$ is an open cover of K_n; by the compactness of K_n, there exists a finite collection $E_{n,q}$ of points in K_n such that

$$K_n \subset \bigcup_{x\in E_{n,q}} B_d(x,q).$$

This holds for any $n \in N_+$, so defining $E_q = \bigcup_n E_{n,q}$,

$$K = \bigcup_n K_n \subset \bigcup_{x \in E_q} B_d(x, q).$$

Defining

$$\mathcal{A} = \{B_d(x, q) \mid q \in \mathbb{Q}, x \in E_q\}.$$

$\mathcal{A}$ is a countable collection of open sets in E such that

$$K \subset \bigcup_{A \in \mathcal{A}} A.$$

Furthermore, for any open G and $x \in K \cap G$, because G is open there exists an $\epsilon > 0$ such that $x \in B_d(x, \epsilon) \subset G$. Choosing any $q \in \mathbb{Q}$ such that $q < \frac{\epsilon}{4}$, because $K \subset \bigcup_{y \in E_q} B_d(y, q)$ and $x \in K$, there exists a $y \in E_q$ such that $x \in B_d(y, q)$. It is also the case that $B_d(y, q)$ is contained in G; to see this, note that, for any $z \in B_d(y, q)$,

$$d(x, z) \leq d(x, y) + d(y, z) < 2q < \frac{\epsilon}{2},$$

which implies that

$$B_d(y, q) \subset B_d(x, \epsilon/2).$$

Denoting $A = B_d(y, q)$, by construction $A \in \mathcal{A}$ and

$$\overline{A} \subset \{z \in E \mid d(x, z) \leq \epsilon/2\} \subset B_d(x, \epsilon) \subset G,$$

so that

$$x \in A \subset \overline{A} \subset G.$$

Summarizing our construction so far, we have found a countable collection $\mathcal{A}$ of open balls in E such that:

- $\mathcal{A}$ covers $K = \bigcup_n K_n$
- For any open set G and $x \in K \cap G$, there exists an $A \in \mathcal{A}$ such that $x \in A \subset \overline{A} \subset G$.

This suggests that $\mathcal{A}$ is actually a countable base for the subspace topology τ_K on K induced by τ. Despite the fact that we have not assumed the separability (and thus second countability) of (E, τ), the second countability of (K, τ_K) follows from the fact that K is the countable union of compact sets.

From $\mathcal{A}$, we can construct the countable collection of interest. Define $\mathcal{F}$ as the collection of all subset of E that can be expressed as the finite union of sets of the form $\overline{A} \cap K_n$,

where $A \in \mathcal{A}$ and $n \in N_+$; formally,

$$\mathcal{F} = \Big\{ \bigcup_{j \in J} (\overline{A_j} \cap K_{n_j}) \mid J \text{ is finite, } \forall j \in J, A_j \in \mathcal{A} \text{ and } n_j \in N_+ \Big\} \cup \{\emptyset\}.$$

Step 2: The Properties of $\mathcal{F}$

We list below some properties of $\mathcal{F}$:

- **$\mathcal{F}$ is a countable collection of subsets of E**
  This is clear because $\mathcal{F}$ consists of finite unions of sets from a countable collection of subsets of E.

- **Any $H \in \mathcal{F}$ is compact and contained in some K_m**
  Let $H \in \mathcal{F}$ be nonempty. Then, there exist $A_1, \cdots, A_n \in \mathcal{A}$ and an $m_1, \cdots, m_n \in N_+$ such that

  $$H = \bigcup_{i=1}^{n} \left(\overline{A_i} \cap K_{m_i} \right).$$

  Because H is the union of closed sets, it is itself closed. Furthermore, it is contained in the compact set $\bigcup_{i=1}^{n} K_{m_i} = K_m$, where $m = \max(m_1, \cdots, m_n)$, so H is itself compact.
  If H is the empty set, then it is trivially compact.
  By implication, H is Borel measurable and $\mathcal{F} \subset \mathcal{E}$.

- **Each K_n is contained in $\mathcal{F}$**
  Because $E_{n,q}$ is finite for each $n \in N_+$ and $q \in \mathbb{Q}$, we have

  $$K_n = \bigcup_{x \in E_{n,q}} \left(\overline{B_d(x,q)} \cap K_n \right) \in \mathcal{F},$$

  since $B_d(x,q) \in \mathcal{A}$ for any $x \in E_{n,q}$.

- **$\mathcal{F}$ is closed under finite unions**

This is obvious, since $\mathcal{F}$ consists of finite unions of sets in the countable collection

$$\{A \cap K_n \mid A \in \mathcal{A}, n \in N_+\}$$

along with the empty set.

Heuristically, $\mathcal{F}$ consists of compact sets that separate closed and open sets. Doing so allows us to formulate results on open sets in terms of approximating sets from below that belong in $\mathcal{F}$.
The formal result can be stated as follows:

Let F and G be closed and open sets in E such that $F \subset G$, and suppose that F is contained in some $H \in \mathcal{F}$. Then, there exists an $H_0 \in \mathcal{F}$ such that $F \subset H_0 \subset G$.

To prove the above statement, let F, G be closed and open sets such taht $F \subset G$ and $F \subset H$ for some $H \in \mathcal{F}$.
If $H = \emptyset$, then $F = \emptyset$ and thus we can take $H_0 = \emptyset$.
Suppose now that H is nonempty. Choose any $x \in F$; because H is a compact set contained in K, $x \in K$. Furthermore, $x \in F \subset G$, which implies that $x \in G$ and thus $x \in K \cap G$. There then exists an $A_x \in \mathcal{A}$ such that

$$x \in A_x \subset \overline{A_x} \subset G.$$

$\{A_x\}_{x \in F}$ is thus an open cover of F whose union is contained in G, and because F, being a closed subset of a compact set H, is itself compact, there exists a finite collection $x_1, \cdots, x_n \in F$ such that

$$F \subset \bigcup_{i=1}^{n} \overline{A_{x_i}}.$$

Lastly, because H is contained in K_m for some $m \in N_+$, $F \subset H \subset K_m$, and defining

$$H_0 = \bigcup_{i=1}^{n} \left(\overline{A_{x_i}} \cap K_m \right) \in \mathcal{F},$$

we thus have

$$F \subset H_0 \subset \bigcup_{i=1}^{n} \overline{A_{x_i}} \subset G.$$

We can also show, as an application of the last result, the following:

Let G_1, G_2 be open sets such that $H \subset G_1 \cup G_2$ for some $H \in \mathcal{F}$. Then, there exist $H_1, H_2 \in \mathcal{F}$ such that $H_1 \subset G_1$, $H_2 \subset G_2$ and $H \subset H_1 \cup H_2$.

To prove this, choose any open sets G_1, G_2 and let $H \subset G_1 \cup G_2$ for some $H \in \mathcal{F}$. Define

$$F_1 = \{x \in H \mid d(x, G_1^c) \geq d(x, G_2^c)\} \quad \text{and} \quad F_2 = \{x \in H \mid d(x, G_1^c) \leq d(x, G_2^c)\}.$$

Since G_1^c and G_2^c are closed, the function $h = d(\cdot, G_1^c) - d(\cdot, G_2^c)$ is continuous on E; we can write

$$F_1 = h^{-1}([0, +\infty)) \cap H \quad \text{and} \quad F_2 = h^{-1}((-\infty, 0]) \cap H,$$

which tells us that F_1 and F_2 are closed sets (H is closed, and the inverse images of h are as well by the closedness of $[0, +\infty)$, $(-\infty, 0]$ and the continuity of h). Furthermore, for any $x \in F_1$, suppose $x \notin G_1$; then, since $x \in H \subset G_1 \cup G_2$, this means that $x \in G_2$. However, this means that

$$d(x, G_2^c) > 0 = d(x, G_1^c),$$

which contradicts $x \in F_1$. Therefore, $x \in G_1$, so that $F_1 \subset G_1$. Likewise, $F_2 \subset G_2$.
We have thus seen that F_1, F_2 are closed sets contained in G_1, G_2 and also in the set $H \in \mathcal{F}$; therefore, by the previous result, there exist $H_1, H_2 \in \mathcal{F}$ such that

$$F_i \subset H_i \subset G_i$$

for $i = 1, 2$.
Finally, since $H = F_1 \cup F_2$, it follows that

$$H \subset H_1 \cup H_2.$$

Step 3: Choosing a Pointwise Convergent Subsequence

The chosen sequence $\{\mu_n\}_{n\in N_+}$ can be viewed as a sequence of functions on the countable collection $\mathcal{F}$ of subsets of E that are pointwise bounded (in fact, they are uniformly bounded since they, being probability meausres, all take values in $[0,1]$). By lemma 4.12, then, there exists a subsequence $\{\mu_{n_k}\}_{k\in N_+}$ of $\{\mu_n\}_{n\in N_+}$ and a function $\alpha : \mathcal{F} \to \mathbb{R}$ such that

$$\lim_{k\to\infty} \mu_{n_k}(H) = \alpha(H)$$

for any $H \in \mathcal{F}$. Because each μ_{n_k} takes values in $[0,1]$, so does α.
The function α possesses the following properties:

- $\alpha(\emptyset) = 0$ trivially, since $\mu_{n_k}(\emptyset) = 0$ for any $k \in N_+$.

- For any disjoint $H_1, H_2 \in \mathcal{F}$,

  $$\mu_{n_k}(H_1 \cup H_2) = \mu_{n_k}(H_1) + \mu_{n_k}(H_2)$$

  for any $k \in N_+$ by finite additivity, so

  $$\alpha(H_1 \cup H_2) = \alpha(H_1) + \alpha(H_2)$$

  as well; α is finitely additive.

- For any $H_1, H_2 \in \mathcal{F}$ that are not necessarily disjoint,

  $$\begin{aligned}\alpha(H_1 \cup H_2) &= \alpha(H_1 \setminus H_2) + \alpha(H_2) \\ &= \alpha(H_1) - \alpha(H_1 \cap H_2) + \alpha(H_2) \leq \alpha(H_1) + \alpha(H_2)\end{aligned}$$

  by finite additivity, so that α is finitely subadditive.

- For any $H_1, H_2 \in \mathcal{F}$ such that $H_1 \subset H_2$,

  $$\alpha(H_2) = \alpha(H_1) + \alpha(H_2 \setminus H_1) \geq \alpha(H_1)$$

  by finite additivity, since $H_1 \cap H_2 = H_1$. Thus, α is monotonic.

Suppose μ is a probability measure on $(E, \mathcal{E})$ such that

$$\mu(G) = \sup\{\alpha(H) \mid H \in \mathcal{F}, H \subset G\}.$$

for any open set G. Then, for any $k \in N_+$ and $H \in \mathcal{F}$ such that $H \subset G$,

$$\mu_{n_k}(G) \geq \mu_{n_k}(H)$$

holds, and taking $k \to \infty$ on both sides yields

$$\liminf_{k\to\infty} \mu_{n_k}(G) \geq \alpha(H).$$

Finally, this holds for any $H \in \mathcal{F}$ such that $H \subset G$, so

$$\liminf_{k\to\infty} \mu_{n_k}(G) \geq \sup\{\alpha(H) \mid H \in \mathcal{F}, H \subset G\}. = \mu(G),$$

and by the Portmanteau theorem, $\mu_{n_k} \to \mu$ weakly.

Therefore, our goal is now to construct some probability measure μ on $(E, \mathcal{E})$ such that

$$\mu(G) = \sup\{\alpha(H) \mid H \in \mathcal{F}, H \subset G\}.$$

It is during this process that Caratheodory's restriction theorem comes into play. Specifically, we first construct an outer measure on E that satisfies the above property for any subset of E, and then show that its restriction to $\mathcal{E}$ is a probability measure.

Step 4: Constructing an Outer Measure on E

As in the Riesz representation theorem, define the function $\beta : \tau \to [0,1]$ as

$$\beta(G) = \sup\{\alpha(H) \mid H \in \mathcal{F}, H \subset G\}$$

for any $G \in \tau$. Note that the set $\{\alpha(H) \mid H \in \mathcal{F}, H \subset G\}$ is nonempty; this is because, for any $G \in \tau$, $\emptyset \subset G$ and $\emptyset \in \mathcal{F}$. so that $\alpha(\emptyset) = 0$ is contained in the above set. It follows that β is well-defined because the set $\{\alpha(H) \mid H \in \mathcal{F}, H \subset G\}$ is nonempty and a subset of $[0,1]$, so that the least upper bound property of the real line ensures β takes values in $[0,1]$ as well.

It is easy to show that β possesses the following properties:

- $\beta(\emptyset) = 0$ because $\emptyset$ is the only set in $\mathcal{F}$ contained in $\emptyset$.

- For any $G_1, G_2 \in \tau$ such that $G_1 \subset G_2$,

  $$\{\alpha(H) \mid H \in \mathcal{F}, H \subset G_1\} \subset \{\alpha(H) \mid H \in \mathcal{F}, H \subset G_2\},$$

  it follows that $\beta(G_1) \leq \beta(G_2)$. Thus, β is monotonic on τ.

We can show that β is countably subadditive on the collection of open sets. We first show that β is finitely subadditive, and then move onto countable subadditivity.

- **Finite Subadditivty**
  Choose any open G_1, G_2 with union G. Then, for any $H \in \mathcal{F}$ such that $H \subset G$, there exist $H_1, H_2 \in \mathcal{F}$ such that $H \subset H_1 \cup H_2$ and $H_1 \subset G_1$ and $H_2 \subset G_2$. By implication,

  $$\begin{aligned} \alpha(H) &\leq \alpha(H_1) + \alpha(H_2) && \text{(Monotonicity and Subadditivity of } \alpha) \\ &\leq \beta(G_1) + \beta(G_2) && \text{(Definition of } \beta) \end{aligned}$$

  This holds for any $H \in \mathcal{F}$ such that $H \subset G$, so

  $$\beta(G) \leq \beta(G_1) + \beta(G_2),$$

  which shows that β is finitely subadditive on τ.

- **Countable Subadditivity**
  Now let $\{G_n\}_{n \in N_+}$ be a sequence of open sets with union G (that is also open).

Choose any $H \in \mathcal{F}$ such that $H \subset G$. This makes $\{G_n\}_{n \in N_+}$ an open cover of the compact set H, and as such there exists an $N \in N_+$ such that

$$H \subset \bigcup_{i=1}^{N} G_i.$$

Then, by definition and finite subadditivity,

$$\alpha(H) \leq \beta\left(\bigcup_{i=1}^{N} G_i\right) \leq \sum_{i=1}^{N} \beta(G_i) \leq \sum_{n=1}^{\infty} \beta(G_n).$$

This holds for any $H \in \mathcal{F}$ such that $H \subset G$, so

$$\beta(G) \leq \sum_{n=1}^{\infty} \beta(G_n).$$

Now we define the function $\mu^* : 2^E \to [0,1]$ as

$$\mu^*(A) = \inf\{\beta(G) \mid A \subset G, G \in \tau\}$$

for any $A \subset E$. Again, $\{\beta(G) \mid A \subset G, G \in \tau\}$ is nonempty (each $A \subset E$ is contained in the open set E) and a subset of $[0,1]$ (β takes values in $[0,1]$), so that $\mu^*(A)$ is well-defined and takes values in $[0,1]$ for any $A \subset E$.
It is also the case that μ^* and β agree on the collection of all open sets. To see this, choose any $G \in \tau$; since G is itself an open set containing G,

$$\mu^*(G) \leq \beta(G).$$

On the other hand, for any $\epsilon > 0$, by the definition of the infimum there exists an open set $V \in \tau$ such that $G \subset V$ and

$$\mu^*(G) \leq \beta(V) < \mu^*(G) + \epsilon.$$

By the monotonicity of β, it follows that

$$\beta(G) \leq \beta(V) < \mu^*(G) + \epsilon,$$

and since this holds for any $\epsilon > 0$, we have $\beta(G) \leq \mu^*(G)$ and thus $\mu^*(G) = \beta(G)$.

We can now show that μ^* is an outer measure on E, similarly to the Riesz representation theorem. We verify each property one by one.

1) Because $\emptyset$ is an open set and μ^* and β agree on τ,

$$\mu^*(\emptyset) = \beta(\emptyset) = 0.$$

2) For any $A_1, A_2 \in 2^E$ such that $A_1 \subset A_2$, because

$$\{\beta(G) \mid A_2 \subset G, G \in \tau\} \subset \{\beta(G) \mid A_1 \subset G, G \in \tau\},$$

it follows that

$$\mu^*(A_1) \leq \mu^*(A_2).$$

3) It remains to verify that μ^* is countably subadditive. Choose any sequence $\{A_n\}_{n \in N_+}$ of subsets of E with union A. By the definition of the infimum, for any $\epsilon > 0$ and $n \in N_+$, there exists an open set G_n such that $A_n \subset G_n$ and

$$\mu^*(A_n) \leq \beta(G_n) < \mu^*(A_n) + \frac{\epsilon}{2^n}.$$

Then, $G = \bigcup_n G_n$ is an open set that contains A; by definition and the countable subadditivity of β,

$$\mu^*(A) \leq \beta(G) \leq \sum_{n=1}^{\infty} \beta(G_n) \leq \sum_{n=1}^{\infty} \mu^*(A_n) + \epsilon.$$

This holds for any $\epsilon > 0$, so

$$\mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(A_n),$$

which shows that μ^* is countably subadditive.

Thereofre, μ^* is an outer measure on E; define the set of all μ^*-measurable sets as $\mathcal{M}$, and recall that

$$\mathcal{M} = \{A \subset E \mid \mu^*(B) = \mu^*(B \cap A) + \mu^*(B \cap A^c) \text{ for all } B \subset E\}.$$

We will now show that $\mathcal{M}$ contains $\mathcal{E}$.

Step 5: $\mathcal{M}$ contains $\mathcal{E}$

Because $\mathcal{M}$ is a σ-algebra on E, it suffices to show that $\mathcal{M}$ contains every closed set; then, because $\mathcal{M}$ is closed under complements, it contains every open set, and thus contains $\mathcal{E}$ because the latter is by definition the smallest σ-algebra on E containing every open set.

Let F be a closed set. Then,

$$\mu^*(B) \leq \mu^*(B \cap F) + \mu^*(B \cap F^c)$$

by the suabdditivity of μ^*, so it remains to see that the reverse inequality holds.

Let G be an open set. Because $F^c \cap G$ is an open set, by the definition of the supremum there exists for any $\epsilon > 0$ an $H_1 \subset F^c \cap G$ such that

$$\beta(F^c \cap G) - \epsilon < \alpha(H_1) \leq \beta(F^c \cap G).$$

Since H_1 is again compact and thus closed, $H_1^c \cap G$ is open, and by the same line of reasoning as above, there exists an $H_2 \subset H_1^c \cap G$ such that

$$\beta(H_1^c \cap G) - \epsilon < \alpha(H_2) \leq \beta(H_1^c \cap G).$$

Because H_1 and H_2 are disjoint, by the finite additivity of α we have

$$\begin{aligned}\alpha(H_1 \cup H_2) = \alpha(H_1) + \alpha(H_2) &> \beta(F^c \cap G) + \beta(H_1^c \cap G) - 2\epsilon \\ &= \mu^*(F^c \cap G) + \mu^*(H_1^c \cap G) - 2\epsilon,\end{aligned}$$

where the last equality follows becausae μ^* and β agree on the open sets.

Furthermore, because $H_1 \subset F^c$, we have $F \subset H_1^c$ and

$$\alpha(H_1 \cup H_2) > \mu^*(F^c \cap G) + \mu^*(F \cap G) - 2\epsilon$$

by the monotonicity of μ^*.

Finally, since $H_1 \cup H_2 \subset G$, by the definition of β we have

$$\mu^*(G) = \beta(G) \geq \alpha(H_1 \cup H_2) > \mu^*(F^c \cap G) + \mu^*(F \cap G) - 2\epsilon.$$

This holds for any $\epsilon > 0$, so

$$\mu^*(G) = \beta(G) \geq \mu^*(F^c \cap G) + \mu^*(F \cap G).$$

Now let $B \subset E$ arbitrarily. For any open G such that $B \subset G$,

$$\beta(G) \geq \mu^*(F^c \cap G) + \mu^*(F \cap G) \geq \mu^*(F^c \cap B) + \mu^*(F \cap B),$$

where the last inequality follows by the monotonicity of μ^*. Therefore,

$$\begin{aligned} \mu^*(B) &= \inf\{\beta(G) \mid G \in \tau, B \subset G\} \\ &\geq \mu^*(F^c \cap B) + \mu^*(F \cap B). \end{aligned}$$

By definition, $F \in \mathcal{M}$, and thus $\mathcal{M} \subset \mathcal{E}$.

Step 6: Constructing the Probability Measure μ

By Caratheodory's restriction theorem, letting μ_0 be the restriction of μ^* to $\mathcal{M}$, $(E, \mathcal{M}, \mu_0)$ is a complete measure space such that

$$\mu_0(G) = \mu^*(G) = \beta(G)$$

for any open set G. Furthermore, since μ^* takes values in $[0,1]$, so does μ_0.

As such, letting $\mu : \mathcal{E} \to [0,1]$ be defined as

$$\mu(A) = \mu_0(A)$$

for any $A \in \mathcal{E}$, so that μ is the restriction of μ_0 to $\mathcal{E}$, the triple $(E, \mathcal{E}, \mu)$ forms a measure space. In particular, for any open set G,

$$\mu(G) = \mu_0(G) = \beta(G) = \sup\{\alpha(H) \mid H \in \mathcal{F}, H \subset G\}.$$

By the remark at the end of step 3, the proof is complete if we can show that μ is a probability measure on $(E, \mathcal{E})$.

To see this, note that, for any $m \in N_+$,

$$1 \geq \mu(E) = \beta(E) \geq \alpha(K_m) = \lim_{k \to \infty} \mu_{n_k}(K_m),$$

where the second inequality follows because $K_m \in \mathcal{F}$, E is open and $K_m \subset E$. By how we chose K_m,

$$\mu_{n_k}(K_m) > 1 - \frac{1}{m}$$

for any $k \in N_+$, which tells us that

$$1 \geq \mu(E) \geq \alpha(K_m) \geq 1 - \frac{1}{m}.$$

This holds for any $m \in N_+$, so taking $m \to \infty$ on both sides yields

$$\mu(E) = 1.$$

Therefore, μ is a probability measure on $(E, \mathcal{E})$, and the proof is complete.

Q.E.D.

4.4.1 Application: Lévy's Continuity Theorem

Prohorov's theorem can be used to prove, among other things, Lévy's continuity theorem, which shows the equivalence of the convergence in distribution of sequences of real valued random vectors and the pointwise convergence of the corresponding characteristic functions.
The formal statement and proof are stated below:

Theorem 4.14 (Lévy's Continuity Theorem)
Let $\{X_n\}_{n\in N_+}$ be a sequence of k-dimensional random vectors with corresponding characteristic functions $\{\varphi_n\}_{n\in N_+}$. For any k-dimensional random vector X with characteristic function φ, $X_n \xrightarrow{d} X$ if and only if

$$\lim_{n\to\infty} \varphi_n(t) = \varphi(t)$$

for any $t \in \mathbb{R}^k$.

Proof) Let $\{\mu_n\}_{n\in N_+}$ be the distributions of $\{X_n\}_{n\in N_+}$ and μ that of X.

Necessity follows easily. Suppose $X_n \xrightarrow{d} X$. For any non-zero $t \in \mathbb{R}^k$, becuase the sine and cosine functions are bounded, real valued and continuous functions on $\mathbb{R}^k$,

$$\lim_{n\to\infty} \int_{\mathbb{R}^k} \cos(t'x) d\mu_n(x) = \int_{\mathbb{R}^k} \cos(t'x) d\mu(x) \quad \text{and}$$
$$\lim_{n\to\infty} \int_{\mathbb{R}^k} \sin(t'x) d\mu_n(x) = \int_{\mathbb{R}^k} \sin(t'x) d\mu(x)$$

by the definition of weak convergence. Therefore,

$$\begin{aligned}
\lim_{n\to\infty} \varphi_n(t) &= \lim_{n\to\infty} \int_{\mathbb{R}^k} \exp(it'x) d\mu_n(x) \\
&= \lim_{n\to\infty} \int_{\mathbb{R}^k} \cos(t'x) d\mu_n(x) + i \cdot \left(\lim_{n\to\infty} \int_{\mathbb{R}^k} \sin(t'x) d\mu_n(x) \right) \\
&= \int_{\mathbb{R}^k} \cos(t'x) d\mu(x) + i \left(\int_{\mathbb{R}^k} \sin(t'x) d\mu(x) \right) = \int_{\mathbb{R}^k} \exp(it'x) d\mu = \varphi(t).
\end{aligned}$$

$\varphi_n(\mathbf{0}) = 1 = \varphi(\mathbf{0})$ for any $n \in N_+$, so it follows that

$$\varphi_n \to \varphi$$

pointwise on $\mathbb{R}^k$.

Moving onto sufficiency, suppose that $\varphi_n \to \varphi$ pointwise on $\mathbb{R}^k$. Then, if $\{\mu_{n_k}\}_{k\in N_+}$ is a subsequence of $\{\mu_n\}_{n\in N_+}$ that converges weakly to some probability measure v on

$(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$, then $v = \mu$. To see this, note that, for any $t \in \mathbb{R}^k$,

$$\begin{aligned}
\int_{\mathbb{R}^k} \exp(it'x) dv(x) &= \lim_{n\to\infty} \int_{\mathbb{R}^k} \exp(it'x) d\mu_{n_k}(x) && (\mu_{n_k} \to v \text{ weakly}) \\
&= \lim_{n\to\infty} \varphi_{n_k}(t) \\
&= \varphi(t). && (\text{By assumption})
\end{aligned}$$

Therefore, v and μ have the same characteristic functions; as such, $v = \mu$ on $\mathcal{B}(\mathbb{R}^k)$, and $\mu_{n_k} \to \mu$ weakly.

As such, in light of lemma 4.10, we can show that $\mu_n \to \mu$ weakly if $\Pi = \{\mu_n\}_{n\in N_+}$ is relatively compact. Going one step further, it is sufficient, in light of Prohorov's theorem, to show that Π is tight to conclude that Π is relatively compact. This is precisely what we set out to prove below.

We want to prove that, for any $\epsilon > 0$, there exists a compact set K such that

$$\mu_n(K) > 1 - \epsilon$$

for any $n \in N_+$.
The proof proceeds roughly as follows.

Since φ, being a characteristic function, is continuous at $\mathbf{0}$, for sufficiently small $m > 0$ it holds that the integral

$$\frac{1}{(2m)^k} \int_{[-m,m]^k} (1 - \varphi(t)) dt$$

is also sufficiently small (note that $\varphi(\mathbf{0}) = 1$). Furthermore, because $\varphi_n \to \varphi$ pointwise, for large enough n the above tells us that

$$\frac{1}{(2m)^k} \int_{[-m,m]^k} (1 - \varphi_n(t)) dt$$

is also small. It can be shown that the integral above is bounded below by the measure of the complement of $[-m^{-1}, m^{-1}]^k$ under μ_n. As such, for large enough n, the value of μ_n outside some compact k-cell can be made arbitrarily small. We can also find such compact k-cells for μ_n with small n, since there are only finitely many μ_n with small n, so taking the (finite) union of these compact k-cells yields the desired K.
The details of the above proof are given below:

Step 1: A Positive Lower Bound for $\frac{1}{(2m)^k}\int_{[-m,m]^k}(1-\varphi_n(t))dt$

For any $n \in N_+$ and $m > 0$, the following holds:

$$\begin{aligned}\frac{1}{(2m)^k}\int_{[-m,m]^k}(1-\varphi_n(t))dt &= 1-\frac{1}{(2m)^k}\int_{[-m,m]^k}\int_{\mathbb{R}^k}\exp(it'x)d\mu_n(x)dt \\ &= 1-\int_{\mathbb{R}^k}\left(\frac{1}{2m}\prod_{j=1}^{k}\int_{-m}^{m}\exp(it_jx_j)dt_j\right)d\mu_n(x) \\ &\qquad\qquad\text{(Fubini's theorem)} \\ &= \int_{\mathbb{R}^k}\left(1-\frac{1}{2m}\prod_{j=1}^{k}\int_{-m}^{m}\exp(it_jx_j)dt_j\right)d\mu_n(x). \\ &\qquad\qquad\text{(μ_n is a probability measure)}\end{aligned}$$

For any $1 \le j \le k$, assuming that $x_j \neq 0$,

$$\begin{aligned}\frac{1}{2m}\int_{-m}^{m}\exp(it_jx_j)dt_j &= \frac{1}{2m}\int_{-m}^{m}\cos(t_jx_j)dt_j + i\cdot\left(\frac{1}{2m}\int_{-m}^{m}\sin(t_jx_j)dt_j\right) \\ &= \frac{1}{m}\int_{0}^{m}\cos(t_jx_j)dt_j = \frac{\sin(mx_j)}{mx_j}.\end{aligned}$$

On the other hand, if $x_j = 0$, then

$$\frac{1}{2m}\int_{-m}^{m}\exp(it_jx_j)dt_j = 1 = \lim_{x_j\to 0}\frac{\sin(mx_j)}{mx_j}.$$

Thus,

$$\frac{1}{(2m)^k}\int_{[-m,m]^k}(1-\varphi_n(t))dt = \int_{\mathbb{R}^k}\left(1-\prod_{j=1}^{k}\frac{\sin(mx_j)}{mx_j}\right)d\mu_n(x),$$

where the value of $\frac{\sin(mx_j)}{mx_j}$ at $x_j = 0$ is taken to be its limit 1.
For any $x \in \mathbb{R}$ such that $|x| > \frac{\pi}{2}$, we have $\left|\frac{\sin(x)}{x}\right| \le \frac{1}{|x|} \le \frac{2}{\pi} < 1$, so if $|mx_j| > \frac{\pi}{2}$ for any $1 \le j \le k$, then

$$1-\prod_{j=1}^{k}\frac{\sin(mx_j)}{mx_j} \ge 1-\left(\frac{2}{\pi}\right)^k = c > 0.$$

Therefore,

$$\begin{aligned}\frac{1}{(2m)^k}\int_{[-m,m]^k}(1-\varphi_n(t))dt &= \int_{\mathbb{R}^k}\left(1-\prod_{j=1}^{k}\frac{\sin(mx_j)}{mx_j}\right)d\mu_n(x) \\ &\ge \int_{\left(\left[\frac{\pi}{2m},\frac{\pi}{2m}\right]^k\right)^c}\left(1-\prod_{j=1}^{k}\frac{\sin(mx_j)}{mx_j}\right)d\mu_n(x) \\ &\ge c\left(1-\mu_n\left(\left[\frac{\pi}{2m},\frac{\pi}{2m}\right]^k\right)\right) \ge 0.\end{aligned}$$

Step 2: A Small Upper Bound for $\frac{1}{(2m)^k}\int_{[-m,m]^k}(1-\varphi_n(t))dt$

Recall that φ, being a characteristic function, is uniformly continuous, and in particular, continuous at $\mathbf{0}$. Therefore, for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|1-\varphi(t)| < \frac{c\cdot\epsilon}{2}$$

for any $t \in \mathbb{R}^k$ such that $t \in [-\delta,\delta]^k$. Therefore,

$$\left|\frac{1}{(2m)^k}\int_{[-m,m]^k}(1-\varphi(t))dt\right| \leq \frac{1}{(2m)^k}\int_{[-m,m]^k}|1-\varphi(t)|dt < \frac{c\cdot\epsilon}{2}$$

for any $m \leq \delta$. In addition, because $\varphi_n \to \varphi$ pointwise and $|\varphi_n| \leq 1$ for any $n \in N_+$, by the BCT we have

$$\lim_{n\to\infty}\frac{1}{(2\delta)^k}\int_{[-\delta,\delta]^k}(1-\varphi_n(t))dt = \frac{1}{(2\delta)^k}\int_{[-\delta,\delta]^k}(1-\varphi(t))dt,$$

which implies that there exists an $N \in N_+$ such that

$$\left|\frac{1}{(2\delta)^k}\int_{[-\delta,\delta]^k}(1-\varphi_n(t))dt - \frac{1}{(2\delta)^k}\int_{[-\delta,\delta]^k}(1-\varphi(t))dt\right| < \frac{c\cdot\epsilon}{2}$$

for any $n \geq N$, where N depends only on δ and thus only on ϵ.

Step 3: Constructing K

From the preceding results, we can see that, for any $n \geq N$,

$$\begin{aligned}\frac{1}{(2\delta)^k}\int_{[-\delta,\delta]^k}(1-\varphi_n(t))dt &\leq \left|\frac{1}{(2\delta)^k}\int_{[-\delta,\delta]^k}(1-\varphi_n(t))dt - \frac{1}{(2\delta)^k}\int_{[-\delta,\delta]^k}(1-\varphi(t))dt\right| \\ &\quad + \left|\frac{1}{(2\delta)^k}\int_{[-\delta,\delta]^k}(1-\varphi(t))dt\right| \\ &< \frac{c\cdot\epsilon}{2} + \frac{c\cdot\epsilon}{2} = c\cdot\epsilon\end{aligned}$$

which implies that

$$1-\mu_n\left(\left[\frac{\pi}{2\delta},\frac{\pi}{2\delta}\right]^k\right) \leq \frac{1}{c}\frac{1}{(2\delta)^k}\int_{[-\delta,\delta]^k}(1-\varphi_n(t))dt < \epsilon.$$

In other words,

$$\mu_n\left(\left[\frac{\pi}{2\delta},\frac{\pi}{2\delta}\right]^k\right) > 1-\epsilon$$

for any $n \geq N$.

Finally, since, for any $1 \leq i \leq N$,

$$\lim_{m \to \infty} \mu_i([-m, m]^k) = 1$$

by sequential continuity, there exists an $M \in N_+$ such that

$$\mu_i([-M, M]^k) > 1 - \epsilon$$

for any $1 \leq i \leq N$. Defining

$$K = \left[-\max\left(M, \frac{\pi}{2\delta}\right), \max\left(M, \frac{\pi}{2\delta}\right) \right],$$

K is a compact subset of E such that

$$\mu_n(K) > 1 - \epsilon$$

for any $n \in N_+$.

It follows by definition that Π is a tight collection of probability measures, so by Prohorov's theorem, $\{\mu_n\}_{n \in N_+}$ is relatively compact. Then, by lemma 4.10, we can conclude that $\mu_n \to \mu$ weakly, or in other words, $X_n \xrightarrow{d} X$.

Q.E.D.

4.4.2 Application: Returning to the Cramer-Wold Device

Like how the continuity theorem above extended the inversion formula so that characteristic functions characterized not only the distribution of a random vector but also the weak convergence of a sequence of random vectors, we can generalize the Cramer-Wold device studied earlier to account for weak convergence as well. Specifically, we can show that a given sequence $\{X_n\}_{n\in N_+}$ of random vectors converges in distribution to some random vector X if and only if $\{r'X_n\}_{n\in N_+}$, converges in distribution to $r'X$ for any nonrandom vector r. This allows us to reduce multidimensional problems to univariate problems, and as such is very useful in proving multidimensional analogues of the central limit theorem. The result is important enough to warrant its own section.

Theorem 4.15 (Cramer-Wold Device II)
Let $\{X_n\}_{n\in N_+}$ be a sequence of k-dimensional random vectors and X a k-dimensional random vector. Then, $X_n \xrightarrow{d} X$ if and only if $r'X_n \xrightarrow{d} r'X$ for any nonzero $r \in \mathbb{R}^k$.

Proof) Let $\{\varphi_n\}_{n\in N_+}$ and $\{\mu_n\}_{n\in N_+}$ be the characteristic functions and distributions corresponding to the sequence $\{X_n\}_{n\in N_+}$, and φ, μ the characteristic function and distribution of X.

Suppose that $X_n \xrightarrow{d} X$. Then, for any non-zero $r \in \mathbb{R}^k$, the function $f : \mathbb{R}^k \to \mathbb{R}$ defined as

$$f(x) = r'x$$

for any $x \in \mathbb{R}^k$ is continuous on $\mathbb{R}^k$. By the continuous mapping theorem,

$$r'X_n = f \circ X_n \xrightarrow{d} f \circ X = r'X.$$

Conversely, suppose that $r'X_n \xrightarrow{d} r'X$ for any non-zero $r \in \mathbb{R}^k$. Choose any $r \in \mathbb{R}^k$ and define f as above. Let c_n be the characteristic functions of $r'X_n$ and c that of $r'X$. The distribution of each $r'X_n$ is the pushforward measure $\mu \circ f^{-1}$, and likewise, the distribution of X is $\mu \circ f^{-1}$. Defining $h : \mathbb{R} \to \mathbb{C}$ as

$$h(x) = \exp(ix)$$

for any $x \in \mathbb{R}$, by lemma 1.1

$$\begin{aligned}\varphi_n(r) = \int_{\mathbb{R}^k} \exp(ir'x) d\mu_n(x) &= \int_{\mathbb{R}^k} (h \circ f) d\mu_n \\ &= \int_{\mathbb{R}} h d(\mu_n \circ f^{-1}) = \int_{\mathbb{R}} \exp(ix) d(\mu_n \circ f^{-1})(x) = c_n(1),\end{aligned}$$

and likewise,

$$\varphi(r) = \int_{\mathbb{R}} \exp(ix) d(\mu \circ f^{-1})(x) = c(1).$$

Because $\mu_n \circ f^{-1} \to \mu \circ f^{-1}$ weakly as $n \to \infty$, by the continuity theorem the characteristic functions $\{c_n\}_{n \in N_+}$ converge pointwise to c. Therefore,

$$\lim_{n \to \infty} \varphi_n(r) = \varphi(r).$$

This holds for any non-zero $r \in \mathbb{R}^k$, and it is trivial when $r = \mathbf{0}$, so by the continuity theorem, $X_n \xrightarrow{d} X$.

Q.E.D.

4.5 Metric for Weak Convergence: The Prohorov Metric

We develop here a metric for weak convergence, similarly to how we developed a metric for convergence in probability. There, given an underlying metric space (E,d), we defined the metric d_{prob} on the space of all random variables taking values in E so that convergence in d_{prob} is equivalent to convergence in probability. Likewise, our goal in this section is to develop a metric π on the space of all probability measures, so that convergence in π is equivalent to weak convergence. This metric, called the Prohorov metric, is useful because it allows us to easily work with functions of probability measures, for instance by letting us to verify continuity via results concerning weak convergence such as the Portmanteau theorem. We formally develop the Prohorov metric and derive relevant proprties below.

Let (E,d) be a metric space, τ the metric topology induced by d, and $\mathcal{E} = \mathcal{B}(E,\tau)$ the Borel σ-algebra generated by τ. For any $A \subset E$ and $\epsilon > 0$, the ϵ-neighborhood A^ϵ of A is defined as the union of all open balls that are centered in A and have radius ϵ; formally,

$$A^\epsilon = \bigcup_{x \in A} B_d(x,\epsilon).$$

Clearly, A^ϵ contains A. Furthermore, being the union of open sets, A^ϵ is open and thus always Borel-measurable.
The following are properties of ϵ-neighborhoods:

Lemma 4.16 Let (E,d) be a metric space, τ the metric topology induced by d, and $\mathcal{E} = \mathcal{B}(E,\tau)$ the Borel σ-algebra generated by τ. The following hold true:

i) For any closed set F, and a sequence $\{\epsilon_n\}_{n \in N_+}$ of positive real numbers that decrease to 0, $\{F^{\epsilon_n}\}_{n \in N_+}$ is a decreasing sequence of open sets such that

$$F = \bigcap_n F^{\epsilon_n}.$$

ii) For any $\epsilon_1, \epsilon_2 > 0$ and $A \subset E$,

$$(A^{\epsilon_1})^{\epsilon_2} \subset A^{\epsilon_1 + \epsilon_2}.$$

Proof) i) It is clear that $\{F^{\epsilon_n}\}_{n \in N_+}$ is a decreasing sequence of sets, since

$$F^{\epsilon_{n+1}} = \bigcup_{x \in F} B_d(x,\epsilon_{n+1}) \subset \bigcup_{x \in F} B_d(x,\epsilon_n) = F^{\epsilon_n}$$

for any $n \in N_+$. Because $F \subset F^{\epsilon_n}$ for any $n \in N_+$, we have

$$F \subset \bigcap_n F^{\epsilon_n}.$$

Conversely, for any $x \in \bigcap_n F^{\epsilon_n}$, there exists for any $n \in N_+$ a point $x_n \in F$ such that $x \in B_d(x_n, \epsilon_n)$. Therefore, $\{x_n\}_{n \in N_+}$ is a sequence in F converging to x, and because F is closed, $x \in F$. This implies that the reverse inclusion holds, and as such

$$F = \bigcap_n F^{\epsilon_n},$$

that is, $\{F^{\epsilon_n}\}_{n \in N_+}$ is a sequence in $\mathcal{B}(E, \tau)$ decreasing to F.

ii) Choose any $x \in (A^{\epsilon_1})^{\epsilon_2}$. Then, because

$$x \in \bigcup_{y \in A^{\epsilon_1}} B_d(y, \epsilon_2),$$

there exists a $y \in A^{\epsilon_1}$ such that $x \in B_d(y, \epsilon_2)$. Similarly, because

$$y \in \bigcup_{z \in A} B_d(z, \epsilon_1),$$

there exists a $z \in A$ such that $y \in B_d(z, \epsilon_1)$. It now follows that $x \in B_d(z, \epsilon_1 + \epsilon_2)$, since

$$d(x, z) \leq d(x, y) + d(y, z) < \epsilon_1 + \epsilon_2.$$

Therefore,

$$x \in \bigcup_{z \in A} B_d(z, \epsilon_1 + \epsilon_2) = A^{\epsilon_1 + \epsilon_2},$$

and because this holds for any $x \in (A^{\epsilon_1})^{\epsilon_2}$, we have the inequality

$$(A^{\epsilon_1})^{\epsilon_2} \subset A^{\epsilon_1 + \epsilon_2}.$$

Q.E.D.

Denote by $\mathcal{M}_p(\mathcal{E})$ the collection of all probability measures on $(E, \mathcal{E})$. Define the function $\pi : \mathcal{M}_p(\mathcal{E})^2 \to [0, +\infty)$ as

$$\pi(\mu, v) = \inf\{\epsilon > 0 \mid \mu(A) \leq v(A^\epsilon) + \epsilon \quad \forall A \in \mathcal{B}(E, \tau)\}$$

for any $\mu, v \in \mathcal{M}_p(\mathcal{E})$. Note that $\pi(\mu, v)$ is well-defined and takes values in $[0, +\infty)$ because the

set

$$\Pi_{\mu,v} = \{\epsilon > 0 \mid \mu(A) \leq v(A^\epsilon) + \epsilon \quad \forall A \in \mathcal{B}(E,\tau)\}$$

is non-empty (the inequality is satisfied for any $A \in \mathcal{B}(E,\tau)$ for $\epsilon = 1$) and bounded below by 0.

We first establish the following important property:

Lemma 4.17 Let (E,d) be a metric space, τ the metric topology induced by d, and $\mathcal{E} = \mathcal{B}(E,\tau)$ the Borel σ-algebra generated by τ. For any $\mu, v \in \mathcal{M}_p(\mathcal{E})$ and $\epsilon > 0$, if

$$\mu(A) \leq v(A^\epsilon) + \epsilon,$$

for any $A \in \mathcal{B}(E,\tau)$, then

$$v(A) \leq \mu(A^\epsilon) + \epsilon$$

for any $A \in \mathcal{B}(E,\tau)$ as well.

Proof) Let $A, B \subset E$, and choose any $\epsilon > 0$. Then, $A \subset (B^\epsilon)^c$ if and only if $B \subset (A^\epsilon)^c$. To show this, suppose that $A \subset (B^\epsilon)^c$. Then, $A \cap B^\epsilon = \emptyset$. Assume, for the sake of contradiction, that $x \in B \cap A^\epsilon$. By definition, $x \in B$ and there exists a $y \in A$ such that $x \in B_d(y,\epsilon)$. The last inclusion can be written as $y \in B_d(x,\epsilon)$, and since $x \in B$, we have $y \in B^\epsilon$. In other words, $y \in A \cap B^\epsilon$, a contradiction. It follows that $B \cap A^\epsilon = \emptyset$. The converse holds by a symmetric argument.

Suppose, for any $\mu, v \in \mathcal{M}_p(\mathcal{E})$ and $\epsilon > 0$, that

$$\mu(A) \leq v(A^\epsilon) + \epsilon.$$

for any $A \in \mathcal{B}(E,\tau)$. Choose any $A \in \mathcal{B}(E,\tau)$ and define

$$B = (A^\epsilon)^c.$$

Then, by definition $B \subset (A^\epsilon)^c$, so the preceding result tells us that $A \subset (B^\epsilon)^c$, and by the monotonicity of measures,

$$v(A) \leq \mu((B^\epsilon)^c) = 1 - \mu(B^\epsilon).$$

Because $(A^\epsilon)^c \in \mathcal{B}(E,\tau)$ as well, we have

$$\mu(B) = 1 - \mu(A^\epsilon) \leq v(B^\epsilon) + \epsilon$$

by assumption. Therefore,

$$v(A) \leq 1 - \mu(B^\epsilon) \leq \mu(A^\epsilon) + \epsilon.$$

Q.E.D.

The above result tells us that, for any $\mu, v \in \mathcal{M}_p(\mathcal{E})$, we can re-express $\Pi_{\mu,v}$ as

$$\Pi_{\mu,v} = \{\epsilon > 0 \mid \mu(A) \leq v(A^\epsilon) + \epsilon \text{ and } v(A) \leq \mu(A^\epsilon) + \epsilon \quad \forall A \in \mathcal{B}(E,\tau)\}.$$

We now show that the function π defined above is a metric on the space $\mathcal{M}_p(\mathcal{E})$:

Theorem 4.18 Let (E,d) be a metric space, τ the metric topology induced by d, and $\mathcal{E} = \mathcal{B}(E,\tau)$ the Borel σ-algebra generated by τ. Then, $\pi : \mathcal{M}_p(\mathcal{E})^2 \to [0,+\infty)$ defined as

$$\pi(\mu,v) = \inf\{\epsilon > 0 \mid \mu(A) \leq v(A^\epsilon) + \epsilon \quad \forall A \in \mathcal{B}(E,\tau)\}$$

for any $\mu, v \in \mathcal{M}_p(\mathcal{E})$ is a metric on $\mathcal{M}_p(\mathcal{E})$.

Proof) By lemma 4.17, we immediately have the reflexivity property

$$\pi(\mu,v) = \pi(v,\mu)$$

for any $\mu, v \in \mathcal{M}_p(\mathcal{E})$.

Suppose that

$$\pi(\mu,v) = 0$$

for some $\mu, v \in \mathcal{M}_p(\mathcal{E})$. Then, for any closed $F \subset E$ and $n \in N_+$, there exists an $0 < \epsilon_n < \min\left(\frac{1}{n}, \epsilon_{n-1}\right)$ (where we define $\epsilon_0 = 1$) such that

$$\mu(F) \leq v\left(F^{\epsilon_n}\right) + \epsilon_n \quad \text{and} \quad v(F) \leq \mu\left(F^{\epsilon_n}\right) + \epsilon_n.$$

By lemma 4.16, the sequence $\{F^{\epsilon_n}\}_{n \in N_+}$ is a sequence of open sets decreasing to F, so by sequential continuity, taking $n \to \infty$ on both sides yields

$$\mu(F) = v(F).$$

This holds for any closed F, and thus μ and v agree for every open set. Since they are probability measures and τ is a π-system generating $\mathcal{B}(E,\tau)$, we can see that $\mu = v$ on $\mathcal{B}(E,\tau)$.

On the other hand, for any $\mu \in \mathcal{M}_p(\mathcal{E})$, it is clear that

$$\pi(\mu, \mu) = 0.$$

It remains to establish the triangle inequality. For any $\mu, w, v \in \mathcal{M}_p(\mathcal{E})$, note that, for any $\epsilon_1 \in \Pi_{\mu,w}$ and $\epsilon_2 \in \Pi_{w,v}$, we have $\epsilon_1 + \epsilon_2 \in \Pi_{\mu,v}$. This can be seen with the help of lemma 4.16: for any $A \in \mathcal{B}(E, \tau)$,

$$\mu(A) \leq w(A^{\epsilon_1}) + \epsilon_1 \quad \text{and} \quad w(A^{\epsilon_1}) \leq v((A^{\epsilon_1})^{\epsilon_2}) + \epsilon_2,$$

and because

$$(A^{\epsilon_1})^{\epsilon_2} \subset A^{\epsilon_1 + \epsilon_2},$$

we have

$$\mu(A) \leq v((A^{\epsilon_1})^{\epsilon_2}) + \epsilon_1 + \epsilon_2 \leq v(A^{\epsilon_1 + \epsilon_2}) + \epsilon_1 + \epsilon_2.$$

Therefore,

$$\pi(\mu, v) = \inf \Pi_{\mu,v} \leq \epsilon_1 + \epsilon_2,$$

and since this holds for any $\epsilon_1 \in \Pi_{\mu,w}$ and $\epsilon_2 \in \Pi_{w,v}$,

$$\pi(\mu, v) \leq \inf \Pi_{\mu,w} + \inf \Pi_{w,v} = \pi(\mu, w) + \pi(w, v).$$

Q.E.D.

The metric π is called the Prohorov metric on $\mathcal{M}_p(\mathcal{E})$.

4.5.1 Convergence in the Prohorov Metric

We now show the most important property of the metric π, namely that a sequence $\{\mu_n\}_{n\in N_+}$ in $\mathcal{M}_p(\mathcal{E})$ converges in the metric π if and only if it converges weakly, given that the underlying space (E,d) is separable.

Theorem 4.19 (Convergence in Prohorov Metric is Weak Convergence)
Let (E,d) be a metric space, τ the metric topology induced by d, and $\mathcal{E}=\mathcal{B}(E,\tau)$ the Borel σ-algebra generated by τ. Let π be the Prohorov metric on $\mathcal{M}_p(\mathcal{E})$, and $\{\mu_n\}_{n\in N_+}$, μ measures belonging to $\mathcal{M}_p(\mathcal{E})$. The following hold true:

i) If $\pi(\mu_n,\mu)\to 0$ as $n\to\infty$, then $\mu_n\to\mu$ weakly.

ii) If (E,d) is separable and $\mu_n\to\mu$ weakly, then $\pi(\mu_n,\mu)\to 0$ as $n\to\infty$.

Proof) **Sufficiency**

Suppose that $\pi(\mu_n,\mu)\to 0$ as $n\to\infty$. Then, for any $n\in N_+$ define

$$\epsilon_n=\min\left(\pi(\mu_n,\mu)+\frac{1}{n},\epsilon_{n-1}\right),$$

where $\epsilon_0=1$; $\{\epsilon_n\}_{n\in N_+}$ is a sequence of positive reals that decreases to 0 such that $\pi(\mu_n,\mu)<\epsilon_n$ for any $n\in N_+$. Choose any closed subset F of E; by lemma 4.16, we immediately know that $\{F^{\epsilon_n}\}_{n\in N_+}$ is a sequence of open sets that sdecreases to F, and that

$$\lim_{n\to\infty}\mu(F^{\epsilon_n})=\mu(F)$$

by sequential continuity.

For any $n\in N_+$, because $\pi(\mu_n,\mu)<\epsilon_n$, there exist a $\delta>0$ such that $\delta<\epsilon_n$ and

$$\mu_n(F)\le\mu(F^{\delta})+\delta.$$

It follows that

$$\mu_n(F)\le\mu(F^{\epsilon_n})+\epsilon_n,$$

and since this holds for any $n\in N_+$,

$$\limsup_{n\to\infty}\mu_n(F)\le\limsup_{n\to\infty}(\mu(F^{\epsilon_n})+\epsilon_n)=\mu(F).$$

This in turn holds for any closed set F, so by the Portmanteau theorem, $\mu_n\to\mu$ weakly.

Necessity

Suppose now that (E,d) is separable, and that $\mu_n \to \mu$ weakly. By the separability of E, there exists a countable subset $E_0 = \{x_k\}_{k\in N_+}$ of E that is dense in E. For any $\epsilon > 0$, this means that the collection $\{B_k\}_{k\in N_+} \subset \mathcal{B}(E,\tau)$ defined as

$$B_k = B_d(x_k, \epsilon/2)$$

for any $k \in N_+$ covers E, that is, $E = \bigcup_k B_k$. By design, each B_k has a diameter of ϵ. Now define $A_1 = B_1$ and

$$A_k = B_k \setminus \left(\bigcup_{i=1}^{k-1} B_i \right)$$

for any $k \geq 2$. $\{A_k\}_{k\in N_+}$ is now a disjoint sequence of sets such whose union is $E = \bigcup_k B_k$, and the diameter of each A_k is less than or equal to ϵ because each A_k is contained in B_k, which has a diameter of ϵ.

By sequential continuity,

$$\lim_{k\to\infty} \mu\left(\bigcup_{i=1}^{k} A_i \right) = \mu(E) = 1,$$

so there exists an $N \in N_+$ such that

$$1 - \mu\left(\bigcup_{i=1}^{N} A_i \right) = \mu\left(\bigcup_{k>N} A_k \right) < \epsilon.$$

We can now define the finite collection of open sets $\mathcal{G}$ as

$$\mathcal{G} = \left\{ (A_{i_1} \cup \cdots \cup A_{i_m})^\epsilon \mid 1 \leq i_1 < \cdots < i_m \leq N \right\} \cup \{\emptyset\}.$$

For any $A \in \mathcal{B}(E,\tau)$, let $\mathcal{A} \subset \{1,\cdots,N\}$ be defined as

$$\mathcal{A} = \{1 \leq i \leq N \mid A_i \cap A \neq \emptyset\},$$

and let the open set A_0 be defined as

$$A_0 = \bigcup_{i\in\mathcal{A}} A_i,$$

where $A_0 = \emptyset$ if $\mathcal{A} = \emptyset$.

If $A_0 \neq \emptyset$, then by how we defined $\mathcal{G}$, it can be seen that $A_0^\epsilon \in \mathcal{G}$. In this case, $A_0^\epsilon \subset A^{2\epsilon}$, since for any $x \in A_0^\epsilon$, there exists a $y \in A_0$ such that $x \in B_d(y,\epsilon)$, and in turn, there exists a $i \in \mathcal{A}$ such that $y \in A_i$. Since $A_i \cap A \neq \emptyset$ by definition, we can choose a $z \in A \cap A_i$; A_i has

diameter ϵ, and $z, y \in A_i$, so we have $d(z,y) \leq \epsilon$. Finally, we can see that $x \in B_d(z, 2\epsilon)$, since

$$d(x,z) \leq d(x,y) + d(y,z) < 2\epsilon.$$

This holds for any $x \in A_0^\epsilon$, so

$$A_0^\epsilon \subset \bigcup_{z \in A} B_d(z, 2\epsilon) = A^{2\epsilon}.$$

If $A_0 = \emptyset$, then the inclusion $A_0^\epsilon \subset A^{2\epsilon}$ follows trivially, since $A_0^\epsilon = \emptyset$.

Now we can see that, because $\mathcal{G}$ is a finite collection of open sets, $\mu_n \to \mu$ weakly, and by the Portmanteau theorem,

$$\liminf_{n \to \infty} \mu_n(G) \geq \mu(G) > \mu(G) - \epsilon$$

for any $G \in \mathcal{G}$. This means that there exists an $N_0 \in N_+$ such that, for any $n \geq N_0$,

$$\inf_{k \geq n} \mu_k(G) > \mu(G) - \epsilon,$$

and in particular,

$$\mu_n(G) > \mu(G) - \epsilon$$

for any $G \in \mathcal{G}$. Note that the choice of N_0 depends only on $\mathcal{G}$ and ϵ, and because $\mathcal{G}$ is defined only on the basis of ϵ, N_0 in turn depends only on ϵ.

It can now be seen that, for any $n \geq N_0$, because $A_0^\epsilon \in \mathcal{G}$,

$$\begin{aligned}
\mu(A) &= \mu\left(A \cap \left(\bigcup_{i=1}^{N} A_i\right)\right) + \mu\left(A \cap \left(\bigcup_{i>N} A_i\right)\right) \\
&\leq \mu\left(A \cap \left(\bigcup_{i \in \mathcal{A}} A_i\right)\right) + \mu\left(\bigcup_{i>N} A_i\right) \\
&\leq \mu(A_0) + \epsilon \\
&\leq \mu(A_0^\epsilon) + \epsilon \\
&\leq \mu_n(A_0^\epsilon) + 2\epsilon \\
&\leq \mu_n(A^{2\epsilon}) + 2\epsilon.
\end{aligned}$$

This holds for any $A \in \mathcal{B}(E, \tau)$, so by definition, for any $n \geq N_0$ we have $2\epsilon \in \Pi_{\mu,\mu_n}$ and thus

$$\pi(\mu, \mu_n) \leq 2\epsilon.$$

Such an N_0 exists for any $\epsilon > 0$, so we can conclude that

$$\lim_{n \to \infty} \pi(\mu_n, \mu) = 0.$$

Q.E.D.

4.5.2 Separability of Spaces of Probability Measures

Another important property of the Prohorov metric is that the space of probability measures $\mathcal{M}_p(\mathcal{E})$ is separable under the Prohorov metric, with the countable dense set being a collection of probability measures with finite support. Specifically, under the Prohorov metric π, any probability measure μ on $(E,\mathcal{E})$ is the π-limit of a sequence $\{\mu_n\}_{n\in N_+}$ of a sequence of probability measures with finite support, that is, a probability measure that assigns positive probability only to a finite set.

Any probability measure v on $(E,\mathcal{E})$ with finite support $\{x_1,\cdots,x_n\}\subset E$ can be written as

$$v=\sum_{i=1}^{n} a_i\cdot\delta_{x_i},$$

where $a_1,\cdots,a_n\in[0,1]$ sum to 1 and δ_x is the Dirac delta measure sitting at x. In light of this formulation, the separability result can be viewed as an analogue, for the space of probability measures, of the fact that any measurable real function can be approximated by a sequence of simple functions.

Theorem 4.20 (Separability under the Prohorov Metric)
Let (E,d) be a separable metric space, τ the metric topology induced by d, and $\mathcal{E}=\mathcal{B}(E,\tau)$ the Borel σ-algebra generated by τ. Let π be the Prohorov metric on $\mathcal{M}_p(\mathcal{E})$ Then, there exists a countable subset E_0 of E such that the collection $\mathbb{B}$ defined as

$$\mathbb{B}=\Big\{\sum_{i=1}^{n} r_i\cdot\delta_{x_i}\mid r_1,\cdots,r_n\in\mathbb{Q}_+,\quad \sum_{i=1}^{n} r_i=1,\quad x_1,\cdots,x_n\in E_0\Big\}$$

is a countable subset of $\mathcal{M}_p(\mathcal{E})$ that is dense in $\mathcal{M}_p(\mathcal{E})$.

Proof) We proceed in steps.

Step 1: Constructing a Countable Partition of E

Choose any $\epsilon>0$.
Let $\{x_k\}_{k\in N_+}$ be a countable subset of E that is dense in E, which exists by the separability of (E,d). The collection $\{B_k\}_{k\in N_+}\subset\mathcal{B}(E,\tau)$ defined as

$$B_k=B_d(x_k,\epsilon/4)$$

for any $k\in N_+$ covers E, that is, $E=\bigcup_k B_k$. Each B_k has a diameter of $\frac{\epsilon}{2}$.
Now define $A_1=B_1$ and

$$A_k=B_k\setminus\left(\bigcup_{i=1}^{k-1}B_i\right)$$

for any $k\geq 2$. $\{A_k\}_{k\in N_+}$ is now a disjoint sequence of sets such whose union is $E=$

$\bigcup_k B_k$, and the diameter of each A_k is less than or equal to $\frac{\epsilon}{2}$ because each A_k is contained in B_k, which has a diameter of $\frac{\epsilon}{2}$. For any $k \in N_+$, choose any $y_k \in B_k$ if B_k is non-empty, and let y_k be any point in E if B_k is empty; define the countable subset E_ϵ of E as

$$E_\epsilon = \{y_k \mid k \in N_+\},$$

and the countable set of probability measures $\mathbb{B}_\epsilon$ as

$$\mathbb{B}_\epsilon = \Big\{ \sum_{i=1}^{n} r_i \cdot \delta_{x_i} \mid r_1, \cdots, r_n \in \mathbb{Q}_+, \quad \sum_{i=1}^{n} r_i = 1, \quad x_1, \cdots, x_n \in E_\epsilon \Big\}.$$

We will now show that within the ϵ-ball of any probability measure in $\mathcal{M}_p(\mathcal{E})$ lies a probability measure v in $\mathbb{B}_\epsilon$.

Step 2: Constructing the Measure v in $\mathbb{B}_\epsilon$

Choose any $\mu \in \mathcal{M}_p(\mathcal{E})$. By sequential continuity,

$$\lim_{k \to \infty} \mu \left(\bigcup_{i=1}^{k} A_i \right) = \mu(E) = 1,$$

so there exists an $N \in N_+$ such that

$$1 - \mu \left(\bigcup_{i=1}^{N} A_i \right) = \mu \left(\bigcup_{k>N} A_k \right) < \frac{\epsilon}{8}.$$

Now choose $r_1, \cdots, r_{N-1} \in \mathbb{Q}$ so that, for any $1 \leq i \leq N-1$,

$$|r_i - \mu(A_i)| = \mu(A_i) - r_i < \frac{\epsilon}{8N}$$

if $\mu(A_i) \geq \frac{\epsilon}{8N}$, and $r_i = 0$ otherwise. This means that $r_i \in [0, \mu(A_i)]$ and

$$|r_i - \mu(A_i)| < \frac{\epsilon}{8N}$$

for any $1 \leq i \leq N-1$.
Defining $r_N = 1 - \sum_{i=1}^{N-1} r_i \in \mathbb{Q}$, since

$$\begin{aligned} \mu(A_N) &= \mu \left(\bigcup_{i=1}^{N} A_i \right) - \mu \left(\bigcup_{i=1}^{N-1} A_i \right) \\ &= 1 - \mu \left(\bigcup_{i>N} A_i \right) - \sum_{i=1}^{N-1} \mu(A_i), \end{aligned}$$

we have

$$\begin{aligned} r_N - \mu(A_N) &= 1 - \sum_{i=1}^{N-1} r_i - \left(1 - \mu\left(\bigcup_{i>N} A_i\right) - \sum_{i=1}^{N-1} \mu(A_i)\right) \\ &= \mu\left(\bigcup_{i>N} A_i\right) + \sum_{i=1}^{N-1} (\mu(A_i) - r_i) \geq 0, \end{aligned}$$

so that $r_N \geq \mu(A_N)$ and

$$|r_N - \mu(A_N)| = r_N - \mu(A_N) < \frac{\epsilon}{8} + \frac{N-1}{N}\frac{\epsilon}{8} < \frac{\epsilon}{8} + \frac{\epsilon}{8} = \frac{\epsilon}{4}.$$

Therefore, $r_1, \cdots, r_N$ are non-negative rational numbers that sum to 1 and satisfy

$$\begin{aligned} \sum_{i=1}^{N} |r_i - \mu(A_i)| &= \sum_{i=1}^{N-1} (\mu(A_i) - r_i) + (r_N - \mu(A_N)) \\ &< \frac{N-1}{N}\frac{\epsilon}{8} + \frac{\epsilon}{4} < \frac{\epsilon}{2}. \end{aligned}$$

By implication, for any subset $\mathcal{A}$ of $\{1, \cdots, N\}$,

$$\left|\sum_{i\in\mathcal{A}} \mu(A_i) - \sum_{i\in\mathcal{A}} r_i\right| \leq \sum_{i\in\mathcal{A}} |r_i - \mu(A_i)| \leq \sum_{i=1}^{N} |r_i - \mu(A_i)| < \frac{\epsilon}{2},$$

so that

$$\sum_{i\in\mathcal{A}} \mu(A_i) < \sum_{i\in\mathcal{A}} r_i + \frac{\epsilon}{2}.$$

Define

$$v = \sum_{i=1}^{N} r_i \cdot \delta_{y_i} \in \mathbb{B}_\epsilon,$$

where $y_i \in B_i$ if B_i is non-empty by construction.

In what follows, we will show that the distance between μ and v is less than ϵ in the Prohorov metric.

Step 3: The Distance between μ and v

Choose any $A \in \mathcal{B}(E,\tau)$, and define

$$\mathcal{A} = \{1 \leq i \leq N \mid A_i \cap A \neq \emptyset\},$$

and

$$A_0 = \bigcup_{i \in \mathcal{A}} A_i,$$

where $A_0 = \emptyset$ if $\mathcal{A} = \emptyset$. Furthermore, $A_0 \subset A^\epsilon$; this is trivial if $A_0 = \emptyset$. If $A_0 \neq \emptyset$, then for any $x \in A_0$, there exists an $i \in \mathcal{A}$ such that $x \in A_i$, and letting $z \in A_i \cap A$, it follows from the fact that A_i has diameter less than or equal to $\frac{\epsilon}{2}$ that

$$d(x,z) \leq \frac{\epsilon}{2} < \epsilon$$

and $x \in B_d(z,\epsilon)$, from which it follows that $A_0 \subset A^\epsilon$.
Furthermore, because

$$y_i \in A_0 = \bigcup_{i \in \mathcal{A}} A_i$$

if and only if $i \in \mathcal{A}$, we have

$$v(A_0) = \sum_{i=1}^{N} r_i \cdot \delta_{y_i}(A_0) = \sum_{i \in \mathcal{A}} r_i.$$

We now have

$$\begin{aligned}
\mu(A) &= \mu\left(A \cap \left(\bigcup_{i=1}^{N} A_i\right)\right) + \mu\left(A \cap \left(\bigcup_{i>N} A_i\right)\right) \\
&\leq \sum_{i \in \mathcal{A}} \mu(A \cap A_i) + \mu\left(\bigcup_{i>N} A_i\right) \\
&\leq \sum_{i \in \mathcal{A}} \mu(A_i) + \frac{\epsilon}{8} \\
&< \sum_{i \in \mathcal{A}} r_i + \frac{\epsilon}{2} + \frac{\epsilon}{8} \\
&= v(A_0) + \epsilon \\
&\leq v(A^\epsilon) + \epsilon.
\end{aligned}$$

This holds for any $A \in \mathcal{B}(E,\tau)$, so

$$\pi(\mu, v) \leq \epsilon.$$

Step 4: The Separability of $\mathcal{M}_p(\mathcal{E})$

Now define the countable subset E_0 of E as

$$E_0 = \bigcup_{n \in N_+} E_{\frac{1}{n}},$$

from which we can deduce that

$$\mathbb{B} = \bigcup_{n} \mathbb{B}_{\frac{1}{n}}.$$

We have seen that, for any $\mu \in \mathcal{M}_p(\mathcal{E})$ and $n \in N_+$, there exists a probability measure v in $\mathbb{B}_{\frac{1}{n}}$ such that $\pi(v, \mu) \leq \frac{1}{n}$. Therefore, for any $\epsilon > 0$, there exists a $v \in \mathbb{B}$ such that $\pi(v, \mu) \leq \epsilon$, and we can see that $\mathbb{B}$ is dense in $\mathcal{M}_p(\mathcal{E})$.

Q.E.D.

4.5.3 Continuous Functions on Spaces of Probability Measures

The fact that any probability measure in $\mathcal{M}_p(\mathcal{E})$ can be approximated with a sequence of probability measures with finite support yields the following representation theorem for continuous functions on $\mathcal{M}_p(\mathcal{E})$. This is a particularly useful result in the study of expected utility in microeconomics:

Theorem 4.21 (Representation of Continuous Functions on $\mathcal{M}_p(\mathcal{E})$)
Let (E,d) be a separable metric space, τ the metric topology induced by d, and $\mathcal{E}=\mathcal{B}(E,\tau)$ the Borel σ-algebra generated by τ. Let π be the Prohorov metric on $\mathcal{M}_p(\mathcal{E})$, and $\mathcal{P}$ a convex subset of $\mathcal{M}_p(\mathcal{E})$ that contains the Dirac delta measures δ_x for any $x\in E$.

Suppose $U:\mathcal{P}\to\mathbb{R}$ a continuous and bounded function with respect to the Prohorov metric such that

$$U(t\mu+(1-t)v)=t\cdot U(\mu)+(1-t)\cdot U(v)$$

for any $\mu,v\in\mathcal{P}$ and $t\in[0,1]$, a property we refer to as linearity.

Then, there exists a continuous and bounded function $u:E\to\mathbb{R}$ such that

$$U(\mu)=\int_E u d\mu$$

for any $\mu\in\mathcal{P}$.

Proof) From the previous theorem, we know that there exists a countable subset E_0 of E such that the collection $\mathbb{B}$ of probability measures with finite support defined as

$$\mathbb{B}=\Big\{\sum_{i=1}^{n} r_i\cdot\delta_{x_i}\mid r_1,\cdots,r_n\in\mathbb{Q}_+,\quad \sum_{i=1}^{n} r_i=1,\quad x_1,\cdots,x_n\in E_0\Big\}$$

is dense in $\mathcal{M}_p(\mathcal{E})$.

We first define the function $u:E\to\mathbb{R}$ as

$$u(x)=U(\delta_x)$$

for any $x\in E$, where $\delta_x\in\mathcal{M}_p(\mathcal{E})$ is the Dirac delta measure sitting at x. u is then bounded because U is.

To see the continuity of u on E, choose any limit point x of E and a sequence $\{x_n\}_{n\in N_+}$ in E that converges to x. For any continuous and bounded function $f:E\to\mathbb{R}$,

$$\int_E f d\delta_{x_n}=f(x_n)\to f(x)=\int_E f d\delta_x$$

as $n\to\infty$ by the continuity of f, so by the definition of weak convergence, $\delta_{x_n}\to\delta_x$

weakly as $n \to \infty$. By the equivalence of weak convergence and convergence in the Prohorov metric π, we can see that

$$\lim_{n\to\infty} \pi(\delta_{x_n}, \delta_x) = 0,$$

and because U is continuous on $\mathcal{P}$, which includes the Dirac delta measures, this implies that

$$\lim_{n\to\infty} u(x_n) = \lim_{n\to\infty} U(\delta_{x_n}) = U(\delta_x) = u(x).$$

Therefore, u is continuous at x, and by extension on E.

u is now trivially $\mathcal{E}$-measurable by continuity, and since it is bounded, it is μ-integrable for any $\mu \in \mathcal{M}_p(\mathcal{E})$. It remains to show that

$$U(\mu) = \int_E u d\mu$$

for any $\mu \in \mathcal{P}$. We proceed in steps.

Step 1: Probability Measures with Finite Support

Choose any $\mu \in \mathcal{M}_p(\mathcal{E})$ with finite support; then, it has the representation

$$\mu = \sum_{i=1}^{n} a_i \cdot \delta_{x_i}$$

for some $a_1, \cdots, a_n \in [0,1]$ that sum to 1, and $x_1, \cdots, x_n \in E$. The convexity of $\mathcal{P}$ then tells us that $\mu \in \mathcal{P}$, and by the linearity of U, we have

$$U(\mu) = \sum_{i=1}^{n} a_i \cdot U(\delta_{x_i}) = \sum_{i=1}^{n} a_i \cdot u(x_i).$$

To evaluate the sum on the right hand side, we turn to a familiar stepwise construction:

1) **Simple Functions**

   For any simple functions f on E with canonical form

   $$f = \sum_{i=1}^{m} b_i \cdot I_{A_i},$$

   we can see that

   $$\begin{aligned}\sum_{i=1}^{n} a_i \cdot f(x_i) &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \cdot \delta_{x_i}(A_i) \\ &= \sum_{j=1}^{m} b_j \cdot \left(\sum_{i=1}^{n} a_i \cdot \delta_{x_i}(A_i)\right) = \sum_{j=1}^{m} b_j \cdot \mu(A_i) = \int_E f d\mu.\end{aligned}$$

2) **Non-negative Functions**

Letting $\{f_k\}_{k\in N_+}$ be a sequence of simple functions on E that increases pointwise to the positive part u^+ of u, this tells us that

$$\begin{aligned}\sum_{i=1}^{n} a_i \cdot u^+(x_i) &= \lim_{k\to\infty}\left[\sum_{i=1}^{n} a_i \cdot f_k(x_i)\right] \\ &= \lim_{k\to\infty}\int_E f_k d\mu = \int_E u^+ d\mu.\end{aligned}$$

where the last equality is justified by the MCT. Likewise,

$$\sum_{i=1}^{n} a_i \cdot u^-(x_i) = \int_E u^- d\mu.$$

3) **Arbitrary Integrable Functions**

By the linearity of integration,

$$\sum_{i=1}^{n} a_i \cdot u(x_i) = \sum_{i=1}^{n} a_i \cdot u^+(x_i) - \sum_{i=1}^{n} a_i \cdot u^-(x_i) = \int_E u^+ d\mu - \int_E u^- d\mu = \int_E u d\mu.$$

Therefore, we can conclude that

$$U(\mu) = \sum_{i=1}^{n} a_i \cdot u(x_i) = \int_E u d\mu$$

when μ has finite support.

Step 2: Arbitrary Probability Measures in $\mathcal{P}$

Now let $\mu \in \mathcal{P}$ in general. By the separability result, there exists a sequence $\{\mu_n\}_{n\in N_+}$ in $\mathbb{B}$ that converges to μ in π, and, in light of the equivalence of convergence in π and weak convergence, weakly as well. Since each μ_n has finite support, it is contained in $\mathcal{P}$, and the preceding result tells us that

$$U(\mu_n) = \int_E u d\mu_n$$

for any $n \in N_+$. u is a continuous and bounded function, so by the definition of weak convergence,

$$\lim_{n\to\infty}\int_E u d\mu_n = \int_E u d\mu.$$

Moreover, because U is continuous on $\mathcal{P}$ and $\{\mu_n\}_{n\in N_+}$ converges to μ in the Prohorov metric,

$$\lim_{n\to\infty} U(\mu_n) = U(\mu).$$

Finally, the uniqueness of limits in $\mathbb{R}$ tells us that

$$U(\mu) = \int_E u d\mu.$$

Q.E.D.

Chapter 5

The Law of Large Numbers and Central Limit Theorems

The various modes of convegence for random variables and weak convergence are primarily used in probability theory to derive laws of large numbers (LLNs) and central limit theorems (CLTs). The former deal with convergence almost surely and in probability, while the latter involve convergence in distribution, or weak convergence. If the convergence is almost sure, then the LLN is question is called a strong LLN (SLLN), while it is called a weak LLN (WLLN) if the convergence is in probability. While a stronger result, SLLNs are relatively rare and exceedingly difficult to prove, so here we focus only on WLLNs.

We present in this chapter various WLLNs and CLTs. We initially start with uncorrelated and i.i.d. sequences of real random vectors, and then consider the more general case of dependent processes.

5.1 WLLN and CLTs for I.I.D. Sequences

5.1.1 WLLN for I.I.D. Sequences

The most well-known version of the WLLN is Chebyshev's WLLN, which holds for pairwise uncorrelated sequences of random variables with finite second moments. On the other hand, Khinchin's WLLN holds for sequences of independent random variables, which is more general in the sense that we need only assume finite first moments.

We often use Chebyshev's WLLN because of how simple it is to prove. Below, we state and prove Chebyshev's WLLN for doubly infinite sequences of random variables, that is, random variables of the form $\{Y_t\}_{t\in\mathbb{Z}}$, since this is the form assumed by most time series. We continue abiding by this convention when discussing law of large numbers and central limit theorems.

Theorem 5.1 (Chebyshev's WLLN)
Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a sequence of pairwise uncorrelated L^2-bounded n-dimensional real random vectors with common mean $\mu \in \mathbb{R}^n$. Then,

$$\frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} \mu$$

as $T \to \infty$.

Proof) Let

$$M = \sup_{t\in\mathbb{Z}} \mathbb{E}|Y_t|^2 < +\infty.$$

For any $t \in \mathbb{Z}$, since

$$\begin{aligned}
\operatorname{tr}(\operatorname{Var}[Y_t]) = \mathbb{E}|Y_t - \mu|^2 &= \mathbb{E}\left[(Y_t-\mu)'(Y_t-\mu)\right] \\
&= \mathbb{E}\left[Y_t'Y_t\right] - \mu'\mathbb{E}\left[Y_t\right] - \mathbb{E}\left[Y_t'\right]\mu + \mu'\mu \\
&= \mathbb{E}|Y_t|^2 - \mu'\mu \\
&\le \mathbb{E}|Y_t|^2
\end{aligned}$$

by the linearity of integration, we can see that

$$\sup_{t\in\mathbb{Z}} \operatorname{tr}(\operatorname{Var}[Y_t]) \le M < +\infty$$

as well.

Since $\mathbb{R}^n$ is a separable metric space given the euclidean metric on $\mathbb{R}^n$, to show that

$$\frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} \mu$$

we must show that

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} Y_t - \mu\right| > \delta\right) = \mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} (Y_t - \mu)\right| > \delta\right) \to 0$$

as $T \to \infty$ for any $\delta > 0$. Choose any $\delta > 0$ and $T \in N_+$. Then,

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} (Y_t - \mu)\right| > \delta\right) &\le \frac{1}{\delta^2}\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T} (Y_t - \mu)\right|^2 \\
&= \frac{1}{T^2\delta^2}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathbb{E}\left[(Y_t-\mu)'(Y_s-\mu)\right] \\
&= \frac{1}{T^2\delta^2}\sum_{t=1}^{T}\sum_{s=1}^{T} \operatorname{tr}(\operatorname{Cov}[Y_t, Y_s])
\end{aligned}$$

by Markov's inequality. Since $\{Y_t\}_{t\in\mathbb{Z}}$ is pairwise uncorrelated,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}(Y_t-\mu)\right|>\delta\right)\leq\frac{1}{T^2\delta^2}\sum_{t=1}^{T}\mathrm{tr}\left(\mathrm{Var}\left[Y_t\right]\right)\leq\frac{M}{T\delta^2},$$

and taking $T\to\infty$ on both sides shows us that

$$\lim_{T\to\infty}\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}(Y_t-\mu)\right|>\delta\right)=0.$$

This holds for any $\delta>0$, so by definition

$$\frac{1}{T}\sum_{t=1}^{T}Y_t\xrightarrow{p}\mu.$$

Q.E.D.

Below we state and prove a version of the WLLN that requires independence instead of the L^2-boundedness assumption. It relies heavily on the Taylor expansion of the characteristic function, so we first state two related results below:

Lemma 5.2 (Taylor Expansion of Characteristic Function)
Let X be an n-dimensional real random vector, and $\varphi:\mathbb{R}^n\to\mathbb{C}$ its characteristic function. Then, the following hold true:

i) If X has finite first moments with mean $\mu\in\mathbb{R}^n$, then

$$\varphi(h)=1+ih'\mu+o(1),\quad |h|\to 0$$

for any $h\in\mathbb{R}^n$.

ii) If X has finite second moments with mean $\mu\in\mathbb{R}^n$ and covariance matrix $\Sigma\in\mathbb{R}^{n\times n}$, then

$$\varphi(h)=1+ih'\mu-\frac{1}{2}h'\left(\Sigma+\mu\mu'\right)h+o(|h|),\quad |h|\to 0.$$

Proof) Suppose that X has a finite first moment with mean $\mathbb{E}[X]=\mu\in\mathbb{R}^n$. For any $h\in\mathbb{R}^n$, the first-order Taylor expansion of the mapping $x\mapsto\exp(ih'x)$ around the zero vector $\mathbf{0}$ tells us that

$$\exp(ih'x)=1+\exp(ir'x_0)\cdot ih'x,$$

for any $x \in \mathbb{R}^n$, where $x_0 = tx$ for some $0 \leq t \leq 1$. It follows that

$$\left|\exp(ih'x) - 1 - ih'x\right| = \left|(\exp(ih'x_0) - 1)ih'x\right| \leq 2 \cdot \left|h'x\right| \leq 2 \cdot |h||x|,$$

where we used the Cauchy-Schwarz inequality to justify the last inequality. This holds for any $x \in \mathbb{R}^n$, so

$$\left|\exp(ih'X) - (1 + ih'X)\right| \leq 2 \cdot |h||X|$$

on Ω. As such,

$$\begin{aligned}\left|\varphi(h) - (1 + ih'\mu)\right| &= \left|\mathbb{E}\left[\exp(ih'X) - (1 + ih'X)\right]\right| \\ &\leq \mathbb{E}\left|\exp(ih'X) - (1 + ih'X)\right| \leq |h| \cdot (2\mathbb{E}|X|).\end{aligned}$$

$\mathbb{E}|X| < +\infty$ by the assumption of finite first moments, so it follows that

$$\varphi(h) = 1 + ih'\mu + r_1(h),$$

where

$$r_1(h) = o(1) \text{ as } |h| \to 0.$$

Now suppose that X has a finite second moment with covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ and mean $\mu \in \mathbb{R}^n$. Then, for any $h \in \mathbb{R}^n$, by the second order Taylor expansion of the mapping $x \mapsto \exp(ih'x)$ around the zero vector,

$$\exp(ih'x) = 1 + ih'x - \frac{1}{2}\exp(ih'x_0)x'hh'x$$

for any $x \in \mathbb{R}^n$, where $x_0 = tx$ for some $0 \leq t \leq 1$. It follows that

$$\begin{aligned}\left|\exp(ih'x) - 1 - ih'x + \frac{1}{2}x'hh'x\right| &= \left|\frac{1}{2}\left(\exp(ih'x_0) - 1\right)x'hh'x\right| \\ &\leq \left|x'h\right|^2 \leq |x|^2|h|^2\end{aligned}$$

for any $x \in \mathbb{R}^n$, so that

$$\left|\exp(ih'X) - (1 + ih'X - \frac{1}{2}X'hh'X)\right| \leq |X|^2|h|^2$$

on Ω. Since

$$\mathbb{E}\left[X'hh'X\right] = \operatorname{tr}\left(hh' \cdot \mathbb{E}\left[XX'\right]\right) = h'\mathbb{E}\left[XX'\right]h = h'\left(\Sigma + \mu\mu'\right)h,$$

we can see that

$$\begin{aligned}\left|\varphi(h)-\left(1+ih'\mu-\frac{1}{2}h'\left(\Sigma+\mu\mu'\right)h\right)\right| &= \left|\mathbb{E}\left[\exp(ih'X)-(1+ih'X-\frac{1}{2}X'hh'X)\right]\right| \\ &\leq \mathbb{E}\left|\exp(ih'X)-(1+ih'X-\frac{1}{2}X'hh'X)\right| \\ &\leq |h|^2\cdot\mathbb{E}|X|^2 = |h|^2\cdot\mathrm{tr}\left(\Sigma+\mu\mu'\right).\end{aligned}$$

It follows that

$$\varphi(h)=1+ih'\mu-\frac{1}{2}h'\left(\Sigma+\mu\mu'\right)h+r_2(h),$$

where

$$r_2(h)=o(|h|)\text{ as }|h|\to 0.$$

Q.E.D.

Theorem 5.3 (Khinchin's WLLN)
Let $\{Y_t\}_{t\in\mathbb{Z}}$ be a sequence of independent and identically distributed n-dimensional real random vectors with common mean $\mu\in\mathbb{R}^n$. Then,

$$\frac{1}{T}\sum_{t=1}^{T}Y_t\xrightarrow{p}\mu.$$

Proof) Since $\{Y_t\}_{t\in\mathbb{Z}}$ is identically distributed, they share a common characteristic function $\varphi:\mathbb{R}^n\to\mathbb{C}$. Defining

$$X_T=\frac{1}{T}\sum_{t=1}^{T}Y_t$$

for any $T\in N_+$, the characteristic function of X_T can be written as

$$\begin{aligned}\psi_T(h)=\mathbb{E}\left[\exp(ih'X_T)\right] &= \mathbb{E}\left[\exp\left(\sum_{t=1}^{T}\frac{1}{T}ih'Y_t\right)\right] \\ &= \mathbb{E}\left[\prod_{t=1}^{T}\exp\left(i\frac{h'}{T}Y_t\right)\right].\end{aligned}$$

Since $Y_1,\cdots,Y_T$ are independent, the characterization of independence via characteristic functions tells us that

$$\psi_T(h)=\prod_{t=1}^{T}\mathbb{E}\left[\exp\left(i\frac{h'}{T}Y_t\right)\right]=\varphi\left(\frac{h}{T}\right)^T,$$

where the last equality follows from the fact that $Y_1, \cdots, Y_T$ are identically distributed. Since each random vector in $\{Y_t\}_{t\in\mathbb{Z}}$ has the same finite first moments, by the preceding lemma we can see that

$$\varphi\left(\frac{h}{T}\right) = 1 + i\frac{h'}{T}\mu + o(1), \quad T \to \infty.$$

Therefore,

$$\psi_T(h) = \left(1 + \frac{ih'\mu}{T} + o(1)\right)^T, \quad T \to \infty,$$

so that

$$\lim_{T\to\infty} \psi_T(h) = \lim_{T\to\infty} \left(1 + \frac{ih'\mu}{T}\right)^T = \exp(ih'\mu).$$

Defining the degenerate random vector $X = \mu$, we can see that the limit on the right hand side is precisely the characteristic function ψ of X evaluated at $h \in \mathbb{R}^n$. This result can now be written as

$$\lim_{T\to\infty} \psi_T(h) = \psi(h)$$

for any $h \in \mathbb{R}^n$, and by the continuity theorem,

$$X_T \xrightarrow{d} X.$$

Since convergence in distribution to a degenerate random variable implies convergence in probability to that same random variable, this implies that

$$X_T = \frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} \mu.$$

Q.E.D.

5.1.2 The CLT for I.I.D. Random Variables

The same Taylor expansion machinery used to prove Khinchin's WLLN to prove the central limit theorem for i.i.d. sequences. This is the most well-known version of the central limit theorem, and it cements the central role played by the normal distribution in probability theory.

Theorem 5.4 (Lindeberg-Levy CLT)
Let $\{Y_t\}_{t\in\mathbb{Z}}$ be an independent and identically distributed sequence of n-dimensional real random vectors with common mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. Defining

$$\bar{Y}_T = \frac{1}{T}\sum_{t=1}^{T} Y_t$$

for any $T \in N_+$,

$$\sqrt{T}\left(\bar{Y}_T - \mu\right) \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \Sigma\right].$$

Proof) Define $Z_t = Y_t - \mu$ for any $t \in \mathbb{Z}$. Then, $\{Z_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of n-dimensional real random vectors with mean zero and common covariance matrix Σ. We need only prove that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \xrightarrow{d} \mathcal{N}\left[\mathbf{0}, \Sigma\right].$$

To this end, let φ be the common characteristic function of the Z_t, and ψ_T the characteristic function of $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t$. As in the proof of Khinchin's WLLN, the i.i.d. property of $\{Z_t\}_{t\in\mathbb{Z}}$ shows us that

$$\begin{aligned}
\psi_T(h) &= \mathbb{E}\left[\exp\left(ih'\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t\right)\right)\right] \\
&= \mathbb{E}\left[\prod_{t=1}^{T}\exp\left(i\frac{h'}{\sqrt{T}}Z_t\right)\right] \\
&= \prod_{t=1}^{T}\mathbb{E}\left[\exp\left(i\frac{h'}{\sqrt{T}}Z_t\right)\right] \\
&= \varphi\left(\frac{h}{\sqrt{T}}\right)^T.
\end{aligned}$$

By the preceding lemma,

$$\varphi\left(\frac{h}{\sqrt{T}}\right) = 1 - \frac{1}{2T}h'\Sigma h + o(1) \text{ as } T \to \infty.$$

We can now see that

$$\psi_T(h) = \varphi\left(\frac{h}{\sqrt{T}}\right)^T = \left(1 + \left(-\frac{1}{2}h'\Sigma h\right)\frac{1}{T} + o(1)\right)^T \to \exp\left(-\frac{1}{2}h'\Sigma h\right)$$

as $T \to \infty$. Letting X be an n-dimensional real random vector with distribution $\mathcal{N}[\mathbf{0}, \Sigma]$, its characteristic function is given as

$$\psi(h) = \exp\left(-\frac{1}{2}h'\Sigma h\right)$$

for any $h \in \mathbb{R}^n$, so we can see that

$$\lim_{T\to\infty} \psi_T(h) = \psi(h)$$

for any $h \in \mathbb{R}^n$. It follows from the continuity theorem that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \xrightarrow{d} X \sim \mathcal{N}[\mathbf{0}, \Sigma].$$

Q.E.D.

Note that, since $\sqrt{T}\left(\bar{Y}_T - \mu\right)$ converges in distribution as $T \to \infty$, it is an $O_p(1)$ sequence. In other words, $\bar{Y}_T$ is itself $O_p(T^{-1/2})$. Heuristically, this result can be understood to mean that the sample mean $\bar{Y}_T$ of an i.i.d. process converges to its probability limit μ, the population mean, at around the same speed as $\frac{1}{\sqrt{T}}$ converges to 0.

5.2 WLLN and CLTs for Dependent Processes

The last section mainly focused on WLLN and CLTs for i.i.d. processes. In most time series applications, however, we cannot assume that the processes of interest are independent, nor can we realistically assume that they are identically distributed. At most, we can construct them so that they are pairwise uncorrelated, but this is insufficient to ensure that they follow some form of the WLLN or CLT.

Here we focus on processes whose elements are not quite independent but unrelated to on one another in a stronger sense than pairwise uncorrelatedness. We show that these processes, called martingale difference sequences, do satisfy the WLLN and CLT given mild additional assumptions. The results presented below are the workhorse asymptotic results in time series analysis.

5.2.1 Martingale Difference Arrays

An array $\{Z_{T,t}\}_{1\leq t\leq k(T),T\in N_+}$ of real random variables is said to be a martingale difference array with respect to the filtration array $\{\mathcal{F}_{T,t}\}_{1\leq t\leq k(T),T\in N_+}$ if

- $Z_{T,t}$ is $\mathcal{F}_{T,t}$ measurable and integrable for any $T\in N_+$ and $1\leq t\leq k(T)$

- For any $T\in N_+$ and $1\leq t\leq k(T)$,

$$\mathbb{E}\left[Z_{T,t}\mid \mathcal{F}_{T,t-1}\right]=0,$$

  where $\mathcal{F}_{T,0}$ is taken to be the trivial σ-algebra.

Before presenting a CLT for martingale difference arrays, we state some preliminary results:

Lemma 5.5 The following hold true:

i) For any $x\in\mathbb{R}$,

$$\left|\exp(ix)-\left(1+ix-\frac{x^2}{2}\right)\right|\leq \min\left(|x|^3,|x|^2\right).$$

ii) For any $z\in\mathbb{C}$,

$$|\exp(z)-(1+z)|\leq |z|^2\exp(|z|).$$

Proof) i) Choose any $x \in \mathbb{R}$, and let $n \in N_+$. Since

$$\frac{\partial}{\partial s}\left(-\frac{1}{n+1}(x-s)^{n+1}\exp(is)\right) = (x-s)^n\exp(is) - \frac{i}{n+1}(x-s)^{n+1}\exp(is)$$

on $\mathbb{R}$, we can see that

$$\int_0^x (x-s)^n\exp(is)ds - \frac{i}{n+1}\int_0^x (x-s)^{n+1}\exp(is)ds = \frac{1}{n+1}x^{n+1},$$

or equivalently,

$$\int_0^x (x-s)^n\exp(is)ds = \frac{1}{n+1}x^{n+1} + \frac{i}{n+1}\int_0^x (x-s)^{n+1}\exp(is)ds.$$

Putting $n=1$ reveals that

$$i - i\cdot\exp(ix) = x + i\cdot\int_0^x (x-s)\exp(is)ds,$$

or that

$$\exp(ix) = 1 + ix - \int_0^x (x-s)\exp(is)ds.$$

The result for $n=2$ now shows us that

$$\int_0^x (x-s)\exp(is)ds = \frac{x^2}{2} + \frac{i}{2}\int_0^x (x-s)^2\exp(is)ds,$$

or that

$$\exp(ix) = 1 + ix - \frac{x^2}{2} - \frac{i}{2}\int_0^x (x-s)^2\exp(is)ds$$

We therefore have the upper bound

$$\left|\exp(ix) - \left(1 + ix - \frac{x^2}{2}\right)\right| \leq \frac{1}{2}\left|\int_0^x (x-s)^2\exp(is)ds\right|.$$

To obtain the first bound, note that, if $x \geq 0$, then

$$\left|\int_0^x (x-s)^2\exp(is)ds\right| \leq \int_0^x (x-s)^2 ds = \frac{x^3}{3}.$$

On the other hand, if $x < 0$, then

$$\left|\int_0^x (x-s)^2\exp(is)ds\right| \leq \int_x^0 (x-s)^2 ds = -\frac{x^3}{3},$$

so that

$$\left|\exp(ix) - \left(1 + ix - \frac{x^2}{2}\right)\right| \leq \frac{|x|^3}{6} \leq |x|^3.$$

It is slightly trickier to obtain the second bound. From the relationship

$$\int_0^x (x-s)\exp(is)ds = \frac{x^2}{2} + \frac{i}{2}\int_0^x (x-s)^2\exp(is)ds,$$

we can see that

$$\frac{1}{2}\left|\int_0^x (x-s)^2\exp(is)ds\right| = \left|\frac{i}{2}\int_0^x (x-s)^2\exp(is)ds\right| \leq \left|\int_0^x (x-s)\exp(is)ds\right| + \frac{x^2}{2}.$$

If $x \geq 0$, then

$$\left|\int_0^x (x-s)\exp(is)ds\right| \leq \int_0^x |x-s|ds = \int_0^x (x-s)ds = \frac{x^2}{2},$$

while if $x < 0$, then

$$\left|\int_0^x (x-s)\exp(is)ds\right| \leq \int_x^0 |x-s|ds = \int_x^0 (s-x)ds = \frac{x^2}{2}.$$

Therefore,

$$\frac{1}{2}\left|\int_0^x (x-s)^2\exp(is)ds\right| \leq \frac{x^2}{2} \leq x^2,$$

and we have

$$\left|\exp(ix) - \left(1 + ix - \frac{x^2}{2}\right)\right| \leq \min(|x|^3, |x|^2).$$

ii) Choose any $z \in \mathbb{C}$. Then,

$$\begin{aligned}\exp(z) = \sum_{n=0}^{\infty}\frac{z^n}{n!} &= 1 + z + \sum_{n=2}^{\infty}\frac{z^n}{n!} \\ &= 1 + z + \frac{z^2}{\cdot}\left(\sum_{n=2}^{\infty}\frac{z^{n-2}}{n!}\right),\end{aligned}$$

so that

$$|\exp(z) - (1+z)| \leq |z|^2 \cdot \left(\sum_{n=2}^{\infty}\frac{|z|^{n-2}}{n!}\right) \leq |z|^2 \cdot \left(\sum_{n=0}^{\infty}\frac{|z|^n}{n!}\right) = |z|^2\exp(|z|).$$

Q.E.D.

We are now ready to present a CLT for martingale difference arrays:

Theorem 5.6 (CLT for Martingale Difference Arrays)
Let $\{Z_{T,t}\}_{T\in N_+,1\leq t\leq k(T)}$ be a square integrable (or L^2) martingale difference array with respect to the filtration array $\mathcal{F} = \{\mathcal{F}_{T,t}\}_{1\leq t\leq k(T),T\in N_+}$. Define $\sigma^2_{T,t} = \mathbb{E}\left[Z^2_{T,t} \mid \mathcal{F}_{T,t-1}\right]$ for any $T \in N_+$ and $1 \leq t \leq k(T)$, and let

$$V_T = \sum_{t=1}^{k(T)} \sigma^2_{T,t}$$

for any $T \in N_+$. Assume that:

i) $V_T \xrightarrow{p} 1$ as $T \to \infty$.

ii) **(The Lindeberg Condition)** For any $\epsilon > 0$,

$$\lim_{T\to\infty} \sum_{t=1}^{k(T)} \mathbb{E}\left[\left|Z^2_{T,t}\right| \cdot I_{\{|Z_{T,t}|>\epsilon\}}\right] = 0.$$

Then, we have

$$\sum_{t=1}^{k(T)} Z_{T,t} \xrightarrow{d} N(0,1).$$

Proof) We proceed in small steps.

Part 1: Bounding V_T

We first modify the array $\{Z_{T,t}\}_{T\in N_+,1\leq t\leq k(T)}$ so that the sum of conditional variances V_T is bounded. Define $\{V_{T,t}\}_{T\in N_+,1\leq t\leq k(T)}$ as

$$V_{T,t} = \sum_{s=1}^{t} \sigma^2_{T,s}$$

for any $T \in N_+$, $1 \leq t \leq k(T)$, so that $V_{T,T} = V_T$, and let $\{Y_{T,t}\}_{T\in N_+,1\leq t\leq k(T)}$ be defined as

$$Y_{T,t} = Z_{T,t} \cdot I_{\{V_{T,t}\leq 2\}}$$

for any $T \in N_+$ and $1 \leq t \leq k(T)$. We now establish some properties of $\{Y_{T,t}\}_{T\in N_+,1\leq t\leq k(T)}$:

- **Martingale Difference Array**
  It is clear that $\{Y_{T,t}\}_{T\in N_+,1\leq t\leq k(T)}$ is a martingale difference array with respect

to $\mathcal{F}$; for any $T \in N_+$ and $1 \leq t \leq k(T)$, since $V_{T,t}$ is $\mathcal{F}_{T,t-1}$-measurable, $Y_{T,t}$ is clearly $\mathcal{F}_{T,t}$ measurable, integrable due to the integrability of $Z_{T,t}$, and

$$\mathbb{E}\left[Y_{T,t} \mid \mathcal{F}_{T,t-1}\right] = \mathbb{E}\left[Z_{T,t} \mid \mathcal{F}_{T,t-1}\right] \cdot I_{\{V_{T,t} \leq 2\}} = 0.$$

– **Square Integrable**

By the square integrability of $Z_{T,t}$, each $Y_{T,t}$ is also square integrable. Furthermore, defining

$$\Gamma_T = \sum_{t=1}^{T} \underbrace{\mathbb{E}\left[Y_{T,t}^2 \mid \mathcal{F}_{T,t-1}\right]}_{\rho_{T,t}^2},$$

since

$$\mathbb{E}\left[Y_{T,t}^2 \mid \mathcal{F}_{T,t-1}\right] = \mathbb{E}\left[Z_{T,t}^2 \mid \mathcal{F}_{T,t-1}\right] \cdot I_{\{V_{T,t} \leq 2\}} = \sigma_{T,t}^2 \cdot I_{\{V_{T,t} \leq 2\}},$$

we can see that

$$\Gamma_T = \sum_{t=1}^{T} \sigma_{T,t}^2 \cdot I_{\{V_{T,t} \leq 2\}}.$$

This implies that $\Gamma_T \leq 2$ on Ω.

Analgously to $V_{T,t}$, we define

$$\Gamma_{T,t} = \sum_{s=1}^{t} \rho_{T,s}^2$$

for any $T \in N_+$ and $1 \leq t \leq k(T)$.

– **Convergence of Γ_T**

Moreover,

$$\begin{aligned} |V_T - \Gamma_T| = V_T - \Gamma_T &= \sum_{t=1}^{T} \sigma_{T,t}^2 \cdot I_{\{V_{T,t} > 2\}} \\ &\leq \left(\sum_{t=1}^{T} \sigma_{T,t}^2 \right) \cdot I_{\{V_T > 2\}} = V_T \cdot I_{\{V_T > 2\}}. \end{aligned}$$

By assumption, $V_T \xrightarrow{p} 1$, and for any $\delta > 0$,

$$\mathbb{P}\left(V_T \cdot I_{\{V_T > 2\}} > \delta\right) \leq \mathbb{P}\left(V_T > 2\right) \leq \mathbb{P}\left(|V_T - 1| > 1\right),$$

so that

$$\lim_{T\to\infty} \mathbb{P}\Big(V_T \cdot I_{\{V_T>2\}} > \delta\Big) = 0.$$

This holds for any $\delta > 0$, so

$$V_T \cdot I_{\{V_T>2\}} \xrightarrow{p} 0$$

and we have

$$\Gamma_T - V_T \xrightarrow{p} 0$$

as $T \to \infty$ as well. Therefore, we can conclude that

$$\Gamma_T \xrightarrow{p} 1.$$

– **The Lindeberg Condition**
For any $T \in N_+$ and $\epsilon > 0$,

$$Y_{T,t} \leq Z_{T,t}$$

and thus

$$\mathbb{E}\left[Y_{T,t}^2 \cdot I_{\{|Y_{T,t}|>\epsilon\}}\right] \leq \mathbb{E}\left[Z_{T,t}^2 \cdot I_{\{|Z_{T,t}|>\epsilon\}}\right]$$

for $1 \leq t \leq k(T)$, so that

$$\sum_{t=1}^{k(T)} \mathbb{E}\left[Y_{T,t}^2 \cdot I_{\{|Y_{T,t}|>\epsilon\}}\right] \leq \sum_{t=1}^{k(T)} \mathbb{E}\left[Z_{T,t}^2 \cdot I_{\{|Z_{T,t}|>\epsilon\}}\right].$$

The right hand side goes to 0 as $T \to \infty$, so $\{Y_{T,t}\}_{T\in N_+, 1\leq t\leq k(T)}$ satisfies the Lindeberg condition

$$\lim_{T\to\infty} \sum_{t=1}^{k(T)} \mathbb{E}\left[Y_{T,t}^2 \cdot I_{\{|Y_{T,t}|>\epsilon\}}\right] = 0.$$

We have thus shown that $\{Y_{T,t}\}_{T\in N_+, 1\leq t\leq k(T)}$ possesses all the same properties as $\{Z_{T,t}\}_{T\in N_+, 1\leq t\leq k(T)}$, with the added property that

$$\sum_{t=1}^{T} \mathbb{E}\left[Y_{T,t}^2 \mid \mathcal{F}_{T,t-1}\right] \leq 2$$

for any $T \in N_+$. Furthermore, since

$$\left|\sum_{t=1}^{k(T)} Z_{T,t} - \sum_{t=1}^{k(T)} Y_{T,t}\right| = \left|\sum_{t=1}^{k(T)} Z_{T,t} \cdot I_{\{V_{T,t}>2\}}\right| \leq \left(\sum_{t=1}^{k(T)} |Z_{T,t}|\right) \cdot I_{\{V_T>2\}},$$

and

$$\mathbb{P}\left(\left(\sum_{t=1}^{k(T)} |Z_{T,t}|\right) \cdot I_{\{V_T>2\}} > \delta\right) \leq \mathbb{P}(V_T > 2)$$

for any $\delta > 0$, we can see that

$$\sum_{t=1}^{k(T)} Z_{T,t} - \sum_{t=1}^{k(T)} Y_{T,t} \xrightarrow{p} 0.$$

Therefore, if we can show that

$$\sum_{t=1}^{k(T)} Y_{T,t} \xrightarrow{d} N(0,1),$$

then by Slutsky's theorem, we can prove the claim of the theorem.

Part 2: The Characteristic Function of $\sum_{t=1}^{k(T)} Y_{T,t}$

To show that the partial sums $\sum_{t=1}^{k(T)} Y_{T,t}$ converge in distribution to the standard normal distribution, we make use of the continuity theorem and show that their characteristic functions converge to that of the desired distribution. To that end, denote by φ_T the characteristic function of

$$S_T = \sum_{t=1}^{k(T)} Y_{T,t}$$

For any $r \in \mathbb{R}$,

$$\begin{aligned}
\left|\varphi_T(r) - \exp\left(-\frac{r^2}{2}\right)\right| &= \left|\mathbb{E}\left[\exp(ir \cdot S_T)\right] - \exp\left(-\frac{r^2}{2}\right)\right| \\
&\leq \left|\mathbb{E}\left[\exp(ir \cdot S_T)\right] - \mathbb{E}\left[\exp(ir \cdot S_T)\exp\left(\frac{r^2\Gamma_T}{2}\right)\exp\left(-\frac{r^2}{2}\right)\right]\right| \\
&\quad + \left|\mathbb{E}\left[\exp(ir \cdot S_T)\exp\left(\frac{r^2\Gamma_T}{2}\right)\exp\left(-\frac{r^2}{2}\right)\right] - \exp\left(-\frac{r^2}{2}\right)\right| \\
&\leq \mathbb{E}\left|1 - \exp\left(\frac{r^2\Gamma_T}{2}\right)\exp\left(-\frac{r^2}{2}\right)\right| + \left|\mathbb{E}\left[\exp(ir \cdot S_T)\exp\left(\frac{r^2\Gamma_T}{2}\right) - 1\right]\right|.
\end{aligned}$$

Because $\Gamma_T \xrightarrow{p} 1$, by the continuous mapping theorem

$$1 - \exp\left(\frac{r^2\Gamma_T}{2}\right)\exp\left(-\frac{r^2}{2}\right) \xrightarrow{p} 0.$$

Furthermore, for any $T \in N_+$, $0 \leq \Gamma_T \leq 2$ on Ω, so that the sequence

$$\left\{1 - \exp\left(\frac{r^2\Gamma_T}{2}\right)\exp\left(-\frac{r^2}{2}\right)\right\}_{T\in N_+}$$

is L^p-bounded for any $p \in [1, +\infty)$. By implication, the sequence is uniformly integrable, which, together with the convergence in probability result above, implies that

$$1 - \exp\left(\frac{r^2\Gamma_T}{2}\right)\exp\left(-\frac{r^2}{2}\right) \xrightarrow{L^1} 0,$$

or equivalently,

$$\mathbb{E}\left|1 - \exp\left(\frac{r^2\Gamma_T}{2}\right)\exp\left(-\frac{r^2}{2}\right)\right| \to 0$$

as $T \to \infty$.

Therefore, it remains to show that

$$\left|\mathbb{E}\left[\exp(ir \cdot S_T)\exp\left(\frac{r^2\Gamma_T}{2}\right) - 1\right]\right| \to 0$$

as $T \to \infty$ for the characteristic function of S_T to converge to that of the standard normal distribution as $T \to \infty$.

Part 3: Decomposing the Second Term

We first express the term

$$\left|\mathbb{E}\left[\exp(ir \cdot S_T)\exp\left(\frac{r^2\Gamma_T}{2}\right) - 1\right]\right|$$

as a telescoping sum, that is,

$$\begin{aligned}
&\mathbb{E}\left[\exp(ir \cdot S_T)\exp\left(\frac{r^2\Gamma_T}{2}\right)\right] - 1 \\
&\qquad = \sum_{t=1}^{k(T)} \mathbb{E}\left[\exp(ir \cdot S_{T,t})\exp\left(\frac{r^2\Gamma_{T,t}}{2}\right) - \exp(ir \cdot S_{T,t-1})\exp\left(\frac{r^2\Gamma_{T,t-1}}{2}\right)\right],
\end{aligned}$$

where we define

$$S_{T,t} = \sum_{s=1}^{t} Y_{T,s}$$

for $1 \leq t \leq k(T)$. For any $1 \leq t \leq k(T)$, using the law of iterated expectations, we can see that

$$\begin{aligned}
&\mathbb{E}\left[\exp(ir \cdot S_{T,t}) \exp\left(\frac{r^2 \Gamma_{T,t}}{2}\right) - \exp(ir \cdot S_{T,t-1}) \exp\left(\frac{r^2 \Gamma_{T,t-1}}{2}\right)\right] \\
&\qquad = \mathbb{E}\left[\exp(ir \cdot S_{T,t-1}) \exp\left(\frac{r^2 \Gamma_{T,t}}{2}\right) \mathbb{E}\left[\exp(ir \cdot Y_{T,t}) - \exp\left(-\frac{r^2 \rho_{T,t}^2}{2}\right) \Big| \mathcal{F}_{T,t-1}\right]\right].
\end{aligned}$$

Therefore, using the fact that $\Gamma_{T,t}$ is bounded above by 2 on Ω, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\exp(ir \cdot S_T) \exp\left(\frac{r^2 \Gamma_T}{2}\right)\right] - 1\right| \\
&\qquad \leq \exp\left(r^2\right) \cdot \sum_{t=1}^{k(T)} \mathbb{E}\left|\mathbb{E}\left[\exp(ir \cdot Y_{T,t}) \mid \mathcal{F}_{T,t-1}\right] - \exp\left(-\frac{r^2 \rho_{T,t}^2}{2}\right)\right|.
\end{aligned}$$

Part 4: Finding an Upper Bound for the Second Term

For any $1 \leq t \leq k(T)$, the previous lemma tells us that

$$\begin{aligned}
\left|\mathbb{E}\left[\exp(ir \cdot Y_{T,t}) \mid \mathcal{F}_{T,t-1}\right] - \left(1 - \frac{r^2 \rho_{T,t}^2}{2}\right)\right| &\leq \mathbb{E}\left[\left|\exp(ir \cdot Y_{T,t}) - \left(1 + ir \cdot Y_{T,t} - \frac{r^2 Y_{T,t}^2}{2}\right)\right| \Big| \mathcal{F}_{T,t-1}\right] \\
&\leq \mathbb{E}\left[\min(|rY_{T,t}|^3, |rY_{T,t}|^2) \mid \mathcal{F}_{T,t-1}\right].
\end{aligned}$$

It can now be seen that, for any $\epsilon > 0$,

$$\begin{aligned}
&\left|\mathbb{E}\left[\exp(ir \cdot Y_{T,t}) \mid \mathcal{F}_{T,t-1}\right] - \left(1 - \frac{r^2 \sigma_{T,t}^2}{2}\right)\right| \\
&\qquad \leq \mathbb{E}\left[|rY_{T,t}|^3 \cdot I_{\{|Y_{T,t}| \leq \epsilon\}} \mid \mathcal{F}_{T,t-1}\right] + \mathbb{E}\left[|rY_{T,t}|^2 \cdot I_{\{|Y_{T,t}| > \epsilon\}} \mid \mathcal{F}_{T,t-1}\right] \\
&\qquad \leq \epsilon |r|^3 \cdot \mathbb{E}\left[Y_{T,t}^2 \cdot I_{\{|Y_{T,t}| \leq \epsilon\}} \mid \mathcal{F}_{T,t-1}\right] + r^2 \cdot \mathbb{E}\left[Y_{T,t}^2 \cdot I_{\{|Y_{T,t}| > \epsilon\}} \mid \mathcal{F}_{T,t-1}\right] \\
&\qquad \leq \epsilon |r|^3 \cdot \rho_{T,t}^2 + r^2 \cdot \mathbb{E}\left[Y_{T,t}^2 \cdot I_{\{|Y_{T,t}| > \epsilon\}} \mid \mathcal{F}_{T,t-1}\right].
\end{aligned}$$

Similarly, the second result in the previous lemma implies

$$\left|\exp\left(-\frac{r^2\rho_{T,t}^2}{2}\right)-\left(1-\frac{r^2\rho_{T,t}^2}{2}\right)\right| \leq \left|\frac{r^2\rho_{T,t}^2}{2}\right|^2 \cdot \exp\left(\frac{r^2\rho_{T,t}^2}{2}\right)$$
$$\leq \frac{r^4}{4}\exp\Big(r^2\Big)\rho_{T,t}^2 \cdot \left(\max_{1\leq s\leq k(T)}\rho_{T,s}^2\right),$$

where we used the fact that

$$\rho_{T,t}^2 \leq \Gamma_T \leq 2.$$

By implication,

$$\left|\mathbb{E}\left[\exp(ir\cdot Y_{T,t}) \mid \mathcal{F}_{T,t-1}\right]-\exp\left(-\frac{r^2\rho_{T,t}^2}{2}\right)\right|$$
$$\leq \epsilon|r|^3\cdot\rho_{T,t}^2 + r^2\cdot\mathbb{E}\left[Y_{T,t}^2\cdot I_{\{|Y_{T,t}|>\epsilon\}} \mid \mathcal{F}_{T,t-1}\right] + \frac{r^4}{4}\exp\Big(r^2\Big)\cdot\rho_{T,t}^2\cdot\left(\max_{1\leq s\leq k(T)}\rho_{T,s}^2\right),$$

and as such

$$\left|\mathbb{E}\left[\exp(ir\cdot S_T)\exp\left(\frac{r^2\Gamma_T}{2}\right)\right]-1\right|$$
$$\leq \exp\Big(r^2\Big)\epsilon|r|^3\cdot\mathbb{E}\left[\Gamma_T\right] + \exp\Big(r^2\Big)r^2\cdot\sum_{t=1}^{k(T)}\mathbb{E}\left[Y_{T,t}^2\cdot I_{\{|Y_{T,t}|>\epsilon\}}\right] + \exp\Big(2r^2\Big)\frac{r^4}{4}\mathbb{E}\left[\Gamma_T\cdot\max_{1\leq s\leq k(T)}\rho_{T,s}^2\right]$$
$$\leq \underbrace{2\exp\Big(r^2\Big)|r|^3\cdot\epsilon}_{I} + \underbrace{\exp\Big(r^2\Big)r^2\cdot\sum_{t=1}^{k(T)}\mathbb{E}\left[Y_{T,t}^2\cdot I_{\{|Y_{T,t}|>\epsilon\}}\right]}_{II} + \underbrace{\exp\Big(2r^2\Big)\frac{r^4}{2}\mathbb{E}\left[\max_{1\leq s\leq k(T)}\rho_{T,s}^2\right]}_{III}.$$

Part 5: The Convergence of the Second Term

The Lindeberg condition ensures that II converges to 0.

As for III, note that

$$\rho_{T,s}^2 = \mathbb{E}\left[Y_{T,s}^2\cdot I_{\{|Y_{T,s}|\leq\epsilon\}} \mid \mathcal{F}_{T,s-1}\right] + \mathbb{E}\left[Y_{Tst}^2\cdot I_{\{|Y_{T,s}|>\epsilon\}} \mid \mathcal{F}_{T,s-1}\right]$$
$$\leq \epsilon^2 + \mathbb{E}\left[Y_{T,s}^2\cdot I_{\{|Y_{T,s}|>\epsilon\}} \mid \mathcal{F}_{T,s-1}\right]$$
$$\leq \epsilon^2 + \sum_{t=1}^{k(T)}\mathbb{E}\left[Y_{T,t}^2\cdot I_{\{|Y_{T,t}|>\epsilon\}} \mid \mathcal{F}_{T,t-1}\right]$$

for any $1 \leq s \leq k(T)$, so

$$\max_{1\leq s\leq k(T)} \rho_{T,s}^2 \leq \epsilon^2 + \sum_{t=1}^{k(T)} \mathbb{E}\left[Y_{T,t}^2 \cdot I_{\{|Y_{T,t}|>\epsilon\}} \mid \mathcal{F}_{T,t-1}\right].$$

It follows that

$$\mathbb{E}\left[\max_{1\leq s\leq k(T)} \rho_{T,s}^2\right] \leq \epsilon^2 + \sum_{t=1}^{k(T)} \mathbb{E}\left[Y_{T,t}^2 \cdot I_{\{|Y_{T,t}|>\epsilon\}}\right],$$

and by the Lindeberg condition,

$$\limsup_{T\to\infty} \mathbb{E}\left[\max_{1\leq s\leq k(T)} \rho_{T,s}^2\right] \leq \epsilon^2.$$

Therefore, the limit supremum of the term III is bounded above by $\exp(2r^2)\frac{r^4}{2}\cdot\epsilon^2$, so that

$$\limsup_{T\to\infty} \left|\mathbb{E}\left[\exp(ir\cdot S_T)\exp\left(\frac{r^2\Gamma_T}{2}\right)\right] - 1\right| \leq \left[2\exp\Big(r^2\Big)|r|^3\cdot\epsilon + \exp\Big(2r^2\Big)\frac{r^4}{2}\cdot\epsilon^2\right].$$

Since this holds for any $\epsilon > 0$, it follows that

$$\lim_{T\to\infty} \left|\mathbb{E}\left[\exp(ir\cdot S_T)\exp\left(\frac{r^2\Gamma_T}{2}\right)\right] - 1\right| = 0.$$

We have shown that

$$\lim_{T\to\infty} \left|\varphi_T(r) - \exp\left(-\frac{r^2}{2}\right)\right| = 0,$$

and because $r \in \mathbb{R}$ was chosen arbitrarily, by the continuity theorem we may conclude that

$$S_T \xrightarrow{d} N(0,1).$$

Q.E.D.

5.2.2 Martingale Difference Sequences

We now turn our attention to martingale difference sequences instead of arrays. A sequence $\{Y_t\}_{t\in\mathbb{Z}}$ of n-dimensional random vectors is said to be an n-dimensional martingale difference sequence (MDS) with respect to the filtration $\mathcal{F} = \{\mathcal{F}_t \mid t \in \mathbb{Z}\}$ if:

- Y_t is $\mathcal{F}_t$-measurable and integrable for any $t \in \mathbb{Z}$
- $\mathbb{E}[Y_t] = \mathbf{0}$ for any $t \in \mathbb{Z}$
- For any $t \in \mathbb{Z}$,

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = \mathbf{0}.$$

Given an n-dimensional MDS $\{Y_t\}_{t\in\mathbb{Z}}$, it can easily be seen that $\{\alpha' Y_t\}_{t\in\mathbb{Z}}$ is a univariate MDS for any $\alpha \in \mathbb{R}^n$. Furthermore, given an univariate MDS $\{y_t\}_{t\in\mathbb{Z}}$ with respect to the filtration $\mathcal{F} = \{\mathcal{F}_t \mid t \in \mathbb{Z}\}$, we can always define a martingale difference array by defining

$$Z_{T,t} = y_t \quad \text{and} \quad \mathcal{F}_{T,t} = \mathcal{F}_t$$

for any $T \in N_+$ and $1 \leq t \leq T = k(T)$.

To obtain a workable version of the martingale difference array CLT for martingale difference sequences, we require the following law of large numbers, adapted from Andrews (1988).

Theorem 5.7 (A Martingale WLLN)
Let $\{Y_t\}_{t\in\mathbb{Z}}$ be an n-dimensional martingale difference sequence with respect to the filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{Z}}$ such that $\{|Y_t|^p \mid t \in \mathbb{Z}\}$ is uniformly integrable for some $1 \leq p \leq 2$. Then,

$$\frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{L^p} \mathbf{0}.$$

Proof) Choose any $\epsilon > 0$. By uniform integrability,

$$\lim_{b\to\infty} \sup_{t\in\mathbb{Z}} \mathbb{E}\left[|Y_t|^p \cdot I_{\{|Y_t|^p > b\}}\right] = 0,$$

so there exists a $B > 0$ such that

$$\sup_{t\in\mathbb{Z}} \mathbb{E}\left[|Y_t|^p \cdot I_{\{|Y_t|^p > B\}}\right] < \left(\frac{\epsilon}{4}\right)^p.$$

For any $t \in \mathbb{Z}$, define

$$e_t = Y_t \cdot I_{\{|Y_t|^p \leq B\}} \quad \text{and}$$
$$u_t = Y_t \cdot I_{\{|Y_t|^p > B\}}.$$

Then, $Y_t = e_t + u_t$, and we have

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = \mathbf{0} = \mathbb{E}[e_t \mid \mathcal{F}_{t-1}] + \mathbb{E}[u_t \mid \mathcal{F}_{t-1}].$$

Furthermore, the sequence $\{e_t - \mathbb{E}[e_t \mid \mathcal{F}_{t-1}]\}_{t\in\mathbb{Z}}$ defines an n-dimensional MDS with respect to $\mathcal{F}$, since both e_t and $\mathbb{E}[e_t \mid \mathcal{F}_{t-1}]$ are integrable random vectors and

$$\mathbb{E}[e_t - \mathbb{E}[e_t \mid \mathcal{F}_{t-1}] \mid \mathcal{F}_{t-1}] = \mathbf{0}.$$

For any $T \in N_+$, we now have

$$\begin{aligned}\left\|\frac{1}{T}\sum_{t=1}^{T} Y_t\right\|_p &\le \left\|\frac{1}{T}\sum_{t=1}^{T}(e_t - \mathbb{E}[e_t \mid \mathcal{F}_{t-1}])\right\|_p + \frac{1}{T}\sum_{t=1}^{T}\|u_t - \mathbb{E}[u_t \mid \mathcal{F}_{t-1}]\|_p \\ &\le \left\|\frac{1}{T}\sum_{t=1}^{T}(e_t - \mathbb{E}[e_t \mid \mathcal{F}_{t-1}])\right\|_p + \frac{1}{T}\sum_{t=1}^{T}\left(\|u_t\|_p + \|\mathbb{E}[u_t \mid \mathcal{F}_{t-1}]\|_p\right)\end{aligned}$$

by Minkowski's inequality. Note that, for any random vector $X \in L^p(\mathcal{H}, \mathbb{P})$, Jensen's inequality implies that

$$(\mathbb{E}|X|^p)^{\frac{2}{p}} \le \mathbb{E}|X|^2,$$

so that $\|X\|_p \le \|X\|_2$. Likewise, the conditional version of Jensen's inequality tells us that, for any $t \in \mathbb{Z}$,

$$\begin{aligned}\|\mathbb{E}[u_t \mid \mathcal{F}_{t-1}]\|_p &= (\mathbb{E}|\mathbb{E}[u_t \mid \mathcal{F}_{t-1}]|^p)^{\frac{1}{p}} \\ &\le (\mathbb{E}|u_t|^p)^{\frac{1}{p}} = \|u_t\|_p.\end{aligned}$$

It follows that

$$\left\|\frac{1}{T}\sum_{t=1}^{T} Y_t\right\|_p \le \left\|\frac{1}{T}\sum_{t=1}^{T}(e_t - \mathbb{E}[e_t \mid \mathcal{F}_{t-1}])\right\|_2 + \frac{2}{T}\sum_{t=1}^{T}\|u_t\|_p.$$

Since

$$u_t = Y_t \cdot I_{\{|Y_t|^p > B\}},$$

by assumption we have

$$\mathbb{E}|u_t|^p = \mathbb{E}\left[|Y_t|^p \cdot I_{\{|Y_t|^p > B\}}\right],$$

and as such

$$\sup_{t\in\mathbb{Z}}\|u_t\|_p \le \left(\sup_{t\in\mathbb{Z}}\mathbb{E}\left[|Y_t|^p\cdot I_{\{|Y_t|^p>B\}}\right]\right)^{\frac{1}{p}} < \frac{\epsilon}{4},$$

which implies

$$\left\|\frac{1}{T}\sum_{t=1}^{T}Y_t\right\|_p \le \left\|\frac{1}{T}\sum_{t=1}^{T}(e_t-\mathbb{E}\left[e_t\mid\mathcal{F}_{t-1}\right])\right\|_2 + \frac{\epsilon}{2}.$$

On the other hand, since martingale difference sequences are pairwise uncorrelated,

$$\begin{aligned}
\mathbb{E}\left|\frac{1}{T}\sum_{t=1}^{T}(e_t-\mathbb{E}\left[e_t\mid\mathcal{F}_{t-1}\right])\right|^2 &= \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}|e_t-\mathbb{E}\left[e_t\mid\mathcal{F}_{t-1}\right]|^2\\
&\le \frac{1}{T^2}\sum_{t=1}^{T}\mathbb{E}\left[\left(|e_t|+|\mathbb{E}\left[e_t\mid\mathcal{F}_{t-1}\right]|\right)^2\right]\\
&\le \frac{1}{T^2}\sum_{t=1}^{T}\left(\|e_t\|_2+\|\mathbb{E}\left[e_t\mid\mathcal{F}_{t-1}\right]\|_2\right)^2\\
&\qquad\qquad\text{(Minkowski's inequality)}\\
&\le \frac{4}{T^2}\sum_{t=1}^{T}\mathbb{E}|e_t|^2.\\
&\qquad\text{(Conditional version of Jensen's inequality)}
\end{aligned}$$

By definition,

$$\mathbb{E}|e_t|^2 = \mathbb{E}\left[|Y_t|^2\cdot I_{\{|Y_t|^p\le B\}}\right] \le B^{\frac{2}{p}}\mathbb{P}\left(|Y_t|^p\le B\right) \le B^{\frac{2}{p}},$$

so we have

$$\left\|\frac{1}{T}\sum_{t=1}^{T}Y_t\right\|_p \le \frac{2B^{\frac{1}{p}}}{\sqrt{T}} + \frac{\epsilon}{2}.$$

Choose $N\in N_+$ so that $\frac{2B^{\frac{1}{p}}}{\sqrt{T}} < \frac{\epsilon}{2}$ for any $T\ge N$; this N depends on B and ϵ, and because our choice of B depends only on ϵ, so does N. We can now see that, for any $T\ge N$,

$$\left\|\frac{1}{T}\sum_{t=1}^{T}Y_t\right\|_p < \epsilon.$$

This holds for any $\epsilon>0$, so by definition

$$\lim_{T\to\infty}\left\|\frac{1}{T}\sum_{t=1}^{T}Y_t\right\|_p = 0.$$

Q.E.D.

We now state and prove a CLT for (possibly multivariate) martingale difference sequences:

Theorem 5.8 (CLT for Martingale Difference Sequences)
Let $\{Y_t\}_{t\in\mathbb{Z}}$ be an n-dimensional martingale difference sequence with respect to the filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{Z}}$. Suppose $\{Y_t\}_{t\in\mathbb{Z}}$ satisfies the following properties:

i) $\{|Y_t| \mid t \in \mathbb{Z}\}$ is L^p-bounded for some $p > 2$.

ii) There exists a positive definite matrix $Q \in \mathbb{R}^{n\times n}$ such that

$$\frac{1}{T}\sum_{t=1}^{T} Y_t Y_t' \xrightarrow{p} Q.$$

Then, as $T \to \infty$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_t \xrightarrow{p} N\left[\mathbf{0}, Q\right].$$

Proof) We make use of the Cramer-Wold device to show this result. Choose any non-zero $\alpha \in \mathbb{R}^n$, and define $Z_t = \alpha' Y_t$ for any $t \in \mathbb{Z}$. As stated earlier, $\{Z_t\}_{t\in\mathbb{Z}}$ is a univariate MDS with respect to $\mathcal{F}$ satisfying

$$\sigma_T^2 = \frac{1}{T}\sum_{t=1}^{T} Z_t^2 = \alpha'\left(\frac{1}{T}\sum_{t=1}^{T} Y_t Y_t'\right)\alpha \xrightarrow{p} \alpha' Q \alpha = \sigma^2.$$

Here, $\sigma^2 > 0$ becausae Q is positive definite and α is non-zero. Furthermore, $\{Z_t\}_{t\in\mathbb{Z}}$ is L^p-bounded, since

$$\mathbb{E}|Z_t|^p \leq |\alpha|^p \cdot \mathbb{E}|Y_t|^p$$

for any $t \in \mathbb{Z}$ by the Cauchy-Schwarz inequality.

Defining

$$Z_{T,t} = \frac{Z_t}{\sigma\sqrt{T}} \quad \text{and} \quad \mathcal{F}_{T,t} = \mathcal{F}_t$$

for any $T \in N_+$ and $1 \leq t \leq T = k(T)$, we obtain the martingale difference array $\{Z_{T,t}\}_{T\in N_+, 1\leq t\leq k(T)}$ with respect to the filtration array $\{\mathcal{F}_{T,t} \mid T \in N_+, 1 \leq t \leq k(T)\}$. This martingale difference array is clearly square integrable, due to the L^p-boundedness of $\{Z_t\}_{t\in\mathbb{Z}}$ and the fact that $p > 2$. We now verify the conditions of the CLT for martingale difference arrays:

– **Convergence of Sum of Variances**

For any $T \in N_+$, define

$$V_T = \sum_{t=1}^{T} \mathbb{E}\left[Z_{T,t}^2 \mid \mathcal{F}_{T,t-1}\right] = \frac{1}{\sigma^2 \cdot T} \sum_{t=1}^{T} \mathbb{E}\left[Z_t^2 \mid \mathcal{F}_{t-1}\right].$$

We saw earlier that

$$\sigma_T^2 = \frac{1}{T} \sum_{t=1}^{T} Z_t^2 \xrightarrow{p} \sigma^2.$$

If we can show that $\frac{\sigma_T^2}{\sigma^2} - V_T \xrightarrow{p} 0$, then we will obtain the desired result $V_T \xrightarrow{p} 1$.

To this end, define

$$x_t = Z_t^2 - \mathbb{E}\left[Z_t^2 \mid \mathcal{F}_{t-1}\right]$$

for any $t \in \mathbb{Z}$. $\{x_t\}_{t\in\mathbb{Z}}$ defines a martingale difference sequence with respect to the filtration $\mathcal{F}$, since each x_t is clearly $\mathcal{F}_t$-measurable, integrable with mean 0, and

$$\mathbb{E}\left[x_t \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[Z_t^2 \mid \mathcal{F}_{t-1}\right] - \mathbb{E}\left[Z_t^2 \mid \mathcal{F}_{t-1}\right] = 0.$$

We noted above that $\{Z_t\}_{t\in\mathbb{Z}}$ was L^p-bounded; because $p > 2$, we can see that

$$\mathbb{E}|Z_t|^p = \mathbb{E}\left|Z_t^2\right|^{\frac{p}{2}}$$

for any $t \in \mathbb{Z}$, which tells us that $\{Z_t^2\}_{t\in\mathbb{Z}}$ is $L^{\frac{p}{2}}$-bounded, where $\frac{p}{2} > 1$. By implication, it is uniformly integrable, which implies that $\{x_t\}_{t\in\mathbb{Z}}$ is also a uniformly integrable martingale difference sequence. By the martingale WLLN proved earlier,

$$\frac{1}{T} \sum_{t=1}^{T} x_t \xrightarrow{L^1} 0,$$

from which it can be inferred that

$$\sigma_T^2 - \sigma^2 V_T = \frac{1}{T} \sum_{t=1}^{T} \left(Z_t - \mathbb{E}\left[Z_t^2 \mid \mathcal{F}_{t-1}\right]\right) = \frac{1}{T} \sum_{t=1}^{T} x_t \xrightarrow{p} 0.$$

Therefore,

$$V_T \xrightarrow{p} 1.$$

– **The Lindeberg Condition**

We can also show that $\{Z_{T,t}\}_{T\in N_+, 1\le t\le k(T)}$ satisfies the Lindeberg condition. By the L^p-boundedness of $\{Z_t\}_{t\in\mathbb{Z}}$, there exists an $M < +\infty$ such that

$$\mathbb{E}|Z_t|^p < M$$

for any $t \in \mathbb{Z}$, which implies that

$$\sum_{t=1}^{k(T)} \mathbb{E}|Z_{T,t}|^p = \frac{1}{\sigma^p T^{\frac{p}{2}}} \sum_{t=1}^{k(T)} \mathbb{E}|Z_t|^p \le \sigma^{-p} T^{1-\frac{p}{2}}.$$

Since $\frac{p}{2} > 1$, taking $T \to \infty$ on both sides yields

$$\lim_{T\to\infty} \sum_{t=1}^{k(T)} \mathbb{E}|Z_{T,t}|^p = 0,$$

which is actually equivalent to Lyapunov's condition.

It now remains to show that Lyapunov's condition implies Lindeberg's. For any $\epsilon > 0$, if $|Z_{T,t}| > \epsilon$, then

$$|Z_{T,t}|^p = |Z_{T,t}|^2 \cdot |Z_{T,t}|^{p-2} > |Z_{T,t}|^2 \cdot \epsilon^{p-2},$$

since $p - 2 > 0$; this means that

$$\epsilon^{p-2} \cdot |Z_{T,t}|^2 \cdot I_{\{|Z_{T,t}|>\epsilon\}} \le |Z_{T,t}|^p \cdot I_{\{|Z_{T,t}|>\epsilon\}} \le |Z_{T,t}|^p.$$

Therefore,

$$\sum_{t=1}^{k(T)} \mathbb{E}\left[|Z_{T,t}|^2 \cdot I_{\{|Z_{T,t}|>\epsilon\}}\right] \le \epsilon^{2-p} \cdot \sum_{t=1}^{k(T)} \mathbb{E}|Z_{T,t}|^p,$$

so taking $T \to \infty$ on both sides yields

$$\lim_{T\to\infty} \sum_{t=1}^{k(T)} \mathbb{E}\left[|Z_{T,t}|^2 \cdot I_{\{|Z_{T,t}|>\epsilon\}}\right] = 0.$$

We have thus seen that the two conditions in the CLT for martingale difference arrays are satisfied. As per that theorem, then, we can conclude that

$$\frac{1}{\sigma\sqrt{T}} \sum_{t=1}^{T} Z_t = \sum_{t=1}^{k(T)} Z_{T,t} \xrightarrow{d} N(0,1).$$

By Slutsky's theorem, this then implies that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \xrightarrow{d} N(0,\sigma^2),$$

or equivalently, for some n-dimensional normally distributed random vector Z with variance Q,

$$\alpha'\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_t\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} Z_t \xrightarrow{d} \alpha' Z.$$

Thish holds for any non-zero $\alpha \in \mathbb{R}^n$, so by the Cramer-Wold device,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_t \xrightarrow{d} Z \sim N\left[\mathbf{0}, Q\right].$$

Q.E.D.

Chapter 6

Continuous Function Spaces

In this chapter we apply the convergence results shown in chapter 4 to the space of continuous functions. These spaces are of great interest because, given an intuitive and tractable metric, they become complete metric spaces. In fact, it can be shown that spaces collecting continuous functions defined on a compact set are also separable. We eventually build up to a central limit theorem on these spaces, from which the existence of the Wiener process and the usual central limit theorems follow, making it a singularly powerful result.

We will work with the following basic framework.
Let (E,τ) be a topological space and $F=\mathbb{R}^n$ or $\mathbb{C}$. Then, we let $B(E,F)$ denote the set of all bounded functions from E to F, and $\mathcal{C}_b(E,F)$ the set of all continuous and bounded functions from E to F. This notation has been encountered before; recall that the definition of the weak convergence of the sequence $\{\mu_n\}_{n\in N_+}$ of probability measures on E to the probability measure μ on E is given as

$$\lim_{n\to\infty}\int_E f d\mu_n = \int_E f d\mu$$

for any $f\in\mathcal{C}_b(E,\mathbb{R})$. From this it can be shown that the above equation holds for any $F=\mathbb{R}^n$ or $\mathbb{C}$ and $f\in\mathcal{C}_b(E,F)$.

A special case arises when E is compact (indeed, this is the case of primary interest). Since real continuous functions defined on a compact set are bounded by the extreme value theorem, the space $\mathcal{C}_b(E,F)$ in this case coincides with $\mathcal{C}(E,F)$, the collection of all continuous functions from E to F. As such, we can jettison the boundedness condition here (strictly speaking, the boundedness condition is not relaxed, but rather follows from continuity and thus does not need to be explicitly stated).

Returning to the general case, recall that the collection of all functions from E to F constitutes a vector space over the complex field, with additive identity equal to the zero function on E (denote the zero element of F by 0_F). $B(E,F)$ and $\mathcal{C}_b(E,F)$, being subsets of this collection that contain the zero function and which are closed under linear combinations, is also a vector

space over the complex field. We can also show that they are normed vector spaces under the same norm.

Define the function $\|\cdot\|_{\mathcal{C}} : B(E,F) \to [0,+\infty)$ as

$$\|f\|_{\mathcal{C}} = \sup_{x\in E} |f(x)|$$

for any $f \in B(E,F)$. $\|\cdot\|_{\mathcal{C}}$ is well-defined and finite because f is bounded. We can show that $\|\cdot\|_{\mathcal{C}}$ defines a norm on $V(E,F)$:

- If $\|f\|_{\mathcal{C}} = 0$, then $f(x) = 0$ for any $x \in E$, so that f is the zero function. Conversely, if $f = 0$ on E, then $\|f\|_{\mathcal{C}} = 0$ trivially.
- For any $z \in \mathbb{C}$,

$$\|zf\|_{\mathcal{C}} = \sup_{x\in E} |zf(x)| = |z| \cdot \left(\sup_{x\in E} |f(x)| \right) = |z| \cdot \|f\|_{\mathcal{C}}.$$

- Because the eucliden norm satisfies the triangle inequality,

$$|f(x) + g(x)| \leq |f(x)| + |g(x)| \leq \|f\|_{\mathcal{C}} + \|g\|_{\mathcal{C}}$$

  for any $x \in E$. Thus, it holds that

$$\|f + g\|_{\mathcal{C}} \leq \|f\|_{\mathcal{C}} + \|g\|_{\mathcal{C}}.$$

The pair $(B(E,F), \|\cdot\|_{\mathcal{C}})$ is a normed vector space over the complex field; $\|\cdot\|_{\mathcal{C}}$ is called the supremum norm on $B(E,F)$. Let $d : B(E,F) \times B(E,F) \to [0,+\infty)$ be the metric induced by the supremum norm, defined as

$$d(f,g) = \sup_{x\in E} |f(x) - g(x)| = \|f - g\|_{\mathcal{C}}$$

for any $f, g \in B(E,F)$. We call d the supremum metric on $B(E,F)$. We first study the space of all bounded functions on E, and then move onto the space of continuous functions, which is our main area of interest.

6.1 The Properties of Continuous Function Spaces

Note that a sequence $\{f_n\}_{n\in N_+} \subset B(E,F)$ converges to some $f \in B(E,F)$ if and only if $f_n \to f$ uniformly on E; this is easy to see.
Suppose that $f_n \to f$ uniformly on E. Then, for any $\epsilon > 0$ there exists an $N \in N_+$ such that

$$|f_n(x) - f(x)| < \epsilon$$

for any $n \geq N$. By implication, for any $n \geq N$,

$$d(f_n, f) = \sup_{x\in E} |f_n(x) - f(x)| \leq \epsilon,$$

and since this holds for any $\epsilon > 0$, $f_n \to f$ in the metric d.
Conversely, if $f_n \to f$ in the metric d, for any $\epsilon > 0$ there exists an $N \in N_+$ such that, for any $n \geq N$,

$$|f_n(x) - f(x)| \leq d(f_n, f) < \epsilon$$

for any $x \in E$. By definition, $f_n \to f$ uniformly on E.

6.1.1 Completeness of Continuous Function Spaces

This equivalence between uniform convergence and convergence in d can be exploited to show that $(B(E,F), d)$ is a complete metric space. By implication, $(B(E,F), \|\cdot\|_{\mathcal{C}})$ is a Banach space over the complex field.

Theorem 6.1 (Completeness of Bounded Function Spaces)
Let E be an arbitrary set, and $F = \mathbb{R}^n$ or $\mathbb{C}$. Suppose $\{f_n\}_{n\in N_+}$ is a sequence in $B(E,F)$ that is Cauchy with respect to the metric d. Then, there exists an $f \in B(E,F)$ such that $f_n \to f$ in d.

Proof) Let $\{f_n\}_{n\in N_+} \subset B(E,F)$ be a Cauchy sequence with respect to d. By definition,

$$\lim_{n,m\to\infty} d(f_n, f_m) = 0,$$

and because $|f_n(x) - f_m(x)| \leq d(f_n, f_m)$ for any $x \in E$, for any $x \in E$ the sequence $\{f_n(x)\}_{n\in N_+}$ is a Cauchy sequence with respect to the euclidean metric on F. Since $(F, |\cdot|)$ is a complete metric space, there exists an $f_x \in F$ such that $f_n(x) \to f_x$ as $n \to \infty$.

Defining $f : E \to F$ as $f(x) = f_x$ for any $x \in E$, by construction $f_n \to f$ pointwise on E. It remains to verify that f is a bounded function on E and that $f_n \to f$ uniformly on E.

Uniform Convergence

We first show that $f_n \to f$ uniformly on E.
For any $\epsilon > 0$, by design there exists an $N \in N_+$ such that

$$d(f_n, f_m) < \frac{\epsilon}{2}$$

for any $n, m \geq N$, which implies that, if $n, m \geq N$, then

$$|f_n(x) - f_m(x)| < \frac{\epsilon}{2}$$

for any $x \in E$. Fix $x \in E$. Then, becauase $f_m(x) \to f(x)$ as $m \to \infty$, there exists an $N_1 \in N_+$ such that $N_1 > N$ and

$$|f_m(x) - f(x)| < \frac{\epsilon}{2}.$$

It follows that, for any $n \geq N$,

$$|f_n(x) - f(x)| \leq |f_n(x) - f_{N_1}(x)| + |f_{N_1}(x) - f(x)| < \epsilon.$$

Our choice of $x \in E$ above was arbitrary, so if $n \geq N$,

$$|f_n(x) - f(x)| < \epsilon$$

for any $x \in E$. This holds for any $\epsilon > 0$, so by definition, $f_n \to f$ uniformly.

Boundedness of f

It follows almost immediately that f is bounded.
From uniform convergence, there exists an $N \in N_+$ such that $n \geq N$ implies

$$|f_n(x) - f(x)| < 1$$

for any $x \in E$. Thus,

$$|f(x)| < 1 + |f_N(x)|$$

for any $x \in E$, and becauase f_N is bounded, so is f.

Q.E.D.

The completeness of the subspace $\mathcal{C}_b(E,F)$ of $B(E,F)$ can be seen as an extension of the completeness of $B(E,F)$. Specifically, given a Cauchy sequence in $\mathcal{C}_b(E,F)$, because this is also a Cauchy sequence in $B(E,F)$, the theorem above shows us that it converges to some $f \in B(E,F)$. The completeness of $\mathcal{C}_b(E,F)$ will be established if we can only show that $\mathcal{C}_b(E,F)$ is a closed subset of $B(E,F)$, that is, if we can show that f is a continuous function.

Corollary to Theorem 6.1 (Completeness of Continuous Function Spaces)
Let (E,τ) be a topologcal space, and $F = \mathbb{R}^n$ or $\mathbb{C}$. Suppose $\{f_n\}_{n\in N_+}$ is a sequence in $\mathcal{C}_b(E,F)$ that is Cauchy with respect to the metric d. Then, there exists an $f \in \mathcal{C}_b(E,F)$ such that $f_n \to f$ in d.

Proof) Let $\{f_n\}_{n\in N_+}$ be a sequence in $\mathcal{C}_b(E,F)$ that is Cauchy in the uniform metric d. By the completeness of $B(E,F)$, there exists some $f \in B(E,F)$ such that $f_n \to f$ in d. We show below that f is a continuous function on E; this completes the proof.

Choose any open subset V of F and $x \in f^{-1}(V)$. This means that $f(x) \in V$, and because $(F,|\cdot|)$ is a metric space, there exists an $\epsilon > 0$ such that $V_\epsilon = B_{|\cdot|}(f(x),\epsilon) \subset V$.
$f_m \to f$ uniformly as $m \to \infty$, so there exists an $N \in N_+$ such that, for any $m \geq N$,

$$|f_m(x) - f(x)| < \frac{\epsilon}{3}$$

for any $x \in E$.
Furthermore, since f_N is continuous and the open ball

$$B = B_{|\cdot|}\left(f_N(x), \frac{\epsilon}{3}\right)$$

is an open subset of F, the inverse image $U = f_N^{-1}(B) \in \tau$. By design, $f_N(x) \in B$, so $x \in f_N^{-1}(B) = U$, so that U is a neighborhood around x, and for any $y \in U$, $f_N(y) \in B$, or equivalently,

$$|f_N(y) - f_N(x)| < \frac{\epsilon}{3}.$$

Note that the choice of U depends on ϵ and N, but because N depends only on ϵ, U is chosen on the basis of ϵ alone.
Now let $y \in U$. Then, by the results above,

$$\begin{aligned} |f(y) - f(x)| &\leq |f(y) - f_N(y)| + |f_N(y) - f_N(x)| + |f_N(x) - f(x)| \\ &< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \end{aligned}$$

As such,

$$U \subset f^{-1}(V_\epsilon) \subset f^{-1}(V).$$

We have so far shown that, for any $x \in f^{-1}(V)$, there exists a neighborhood $U_x \in \tau$ of x such that $U_x \subset f^{-1}(V)$. Then, defining

$$U = \bigcup_{x \in f^{-1}(V)} U_x,$$

$U \in \tau$ because arbitrary unions of open sets are open sets, and

$$U = f^{-1}(V).$$

In other words, $f^{-1}(V)$ is an open set in E, and because this holds for any open subset V of F, by definition f is a continuous function on E.

Q.E.D.

6.1.2 The Modulus of Continuity

Here we introduce a useful characterization of the uniform continuity of functions in $\mathcal{C}_b(E,F)$ when (E,τ) is metrizable by some metric ρ on E. Define the function $w:\mathcal{C}_b(E,F)\times(0,+\infty)\to[0,+\infty)$ as

$$w(f,\delta)=\sup\{|f(x)-f(y)|\mid x,y\in E,\rho(x,y)<\delta\}$$

for any $f\in\mathcal{C}_b(E,F)$ and $\delta>0$. This quantity is well-defined and finite by the least upper bound property of the real line because the set $\{|f(x)-f(y)|\mid x,y\in E,\rho(x,y)<\delta\}$ is bounded above (f is a bounded function). The following are the properties of the function w:

Lemma 6.2 Let (E,ρ) be a metric space, and $F=\mathbb{R}^n$ or $\mathbb{C}$. Let $w:\mathcal{C}_b(E,F)\times(0,+\infty)\to[0,+\infty)$ be defined as above. The following hold true:

i) For any $f\in\mathcal{C}_b(E,F)$, the section $\delta\mapsto w(f,\delta)$ is increasing in δ.

ii) For any $\delta>0$, the section $f\mapsto w(f,\delta)$ is uniformly continuous on $\mathcal{C}_b(E,F)$ with respect to the supremum metric d.

iii) For any $f\in\mathcal{C}_b(E,F)$, f is uniformly continuous on E if and only if

$$\lim_{\delta\to0}w(f,\delta)=0.$$

Proof) The first claim is obvious; for any $f\in\mathcal{C}_b(E,F)$ and $0<h_1<h_2$,

$$\{|f(x)-f(y)|\mid x,y\in E,\rho(x,y)<h_1\}\subset\{|f(x)-f(y)|\mid x,y\in E,\rho(x,y)<h_2\},$$

so $w(f,h_1)\le w(f,h_2)$.

Now choose any $\delta>0$. For any $f,g\in\mathcal{C}_b(E,F)$, assume without loss of generality that $w(f,\delta)\ge w(g,\delta)$. Then, by the definition of the supremum, for any $\epsilon>0$ there exist $x,y\in E$ such that $\rho(x,y)<\delta$ and

$$w(f,\delta)-\epsilon<|f(x)-f(y)|\le w(f,\delta).$$

It follows that

$$\begin{aligned}|w(f,\delta)-w(g,\delta)|&=w(f,\delta)-w(g,\delta)\\&\le|f(x)-f(y)|-|g(x)-g(y)|+\epsilon\\&\le|f(x)-f(y)-g(x)+g(y)|+\epsilon\\&\le|f(x)-g(x)|+|f(y)-g(y)|+\epsilon=2\cdot d(f,g)+\epsilon.\end{aligned}$$

This holds for any $\epsilon > 0$, so

$$|w(f,\delta) - w(g,\delta)| \leq 2 \cdot d(f,g),$$

which shows that $w(\cdot,\delta)$ is uniformly continuous on $\mathcal{C}_b(E,F)$ with respect to the metric d.

To show the final claim, choose any $f \in \mathcal{C}_b(E,F)$, and suppose it is uniformly continuous on E. Then, for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|f(x) - f(y)| < \epsilon$$

for any $x, y \in E$ such that $\rho(x,y) < \delta$. Then, for any $0 < h < \delta$,

$$w(f,h) \leq w(f,\delta) = \sup\{|f(x) - f(y)| \mid x,y \in E, \rho(x,y) < \delta\} \leq \epsilon.$$

This holds for any $\epsilon > 0$, so by definition,

$$\lim_{h \to 0} w(f,h) = 0.$$

Conversely, suppose that $w(f,h) \to 0$ as $h \to 0$. Then, for any $\epsilon > 0$, there exists a $\delta > 0$ such that $w(f,h) < \epsilon$ for any $0 < h \leq \delta$. In particular,

$$w(f,\delta) = \sup\{|f(x) - f(y)| \mid x,y \in E, \rho(x,y) < \delta\} < \epsilon,$$

which implies that, for any $x, y \in E$ such that $\rho(x,y) < \delta$,

$$|f(x) - f(y)| \leq w(f,\delta) < \epsilon.$$

This holds for any $\epsilon > 0$, so by definition, f is uniformly continuous on E.
Q.E.D.

Due to the last property, which shows that the convergence of $w(f,\cdot)$ to 0 is equivalent to the uniform continuity of f, the function $w(f,\cdot)$ is often called the modulus of continuity of f. We will see later that the modulus of continuity helps characterize relative compactness and tightness of continuous function spaces.

6.1.3 Relative Compactness and Equicontinuity

In continuous function spaces, there exists a particularly convenient characterization of relative compactness that involves the concept of equicontinuity. Let (E, ρ) be a metric space, d the supremum metric on $\mathcal{C}_b(E, F)$ and A a subset of the space $\mathcal{C}_b(E, F)$. We say that A is equicontinuous at $x \in E$ if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that, for any $y \in E$ satsifying $\rho(x, y) < \delta$,

$$|f(x) - f(y)| < \epsilon$$

for any $f \in A$. In other words, A is equicontinuous at x if every function in A is continuous at x to roughly the same degree.

A stronger notion than equicontinuity at a point is uniform equicontinuity. As the word suggests, A is uniformly equicontinuous if every function in A is continuous at any point on E to roughly the same degree. Formally, A is uniformly equicontinuous on E if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that, for any $x, y \in E$ satisfying $\rho(x, y) < \delta$, we have

$$|f(x) - f(y)| < \epsilon$$

for any $f \in A$. An equivalent formulation can be given in terms of the modulus of continuity: A is uniformly equicontinuous on E if and only if

$$\lim_{h \to 0} \sup_{f \in A} w(f, h) = 0.$$

It is easy to see the equivalence:

- **Necessity**
  If A is uniformly equicontinuous on E, then for any ϵ, there exists a $\delta > 0$ such that, for any $x, y \in E$ satisfying $\rho(x, y) < \delta$, we have

  $$|f(x) - f(y)| < \epsilon$$

  for any $f \in A$. It follows that, for any $f \in A$,

  $$w(f, \delta) = \sup\{|f(x) - f(y)| \mid x, y \in E, \rho(x, y) < \delta\} \leq \epsilon,$$

  so

  $$\sup_{f \in A} w(f, \delta) \leq \epsilon.$$

Because $w(f,\cdot)$ is increasing on $(0,+\infty)$ for any $f \in \mathcal{C}_b(E,F)$, for any $0 < h < \delta$,

$$\sup_{f\in A} w(f,h) \leq \epsilon.$$

This holds for any $\epsilon > 0$, so

$$\lim_{h\to 0} \sup_{f\in A} w(f,h) = 0.$$

- **Sufficiency**
  Suppose that

  $$\lim_{h\to 0} \sup_{f\in A} w(f,h) = 0.$$

  Then, for any $\epsilon > 0$, there exists a $\delta > 0$ such that

  $$\sup_{f\in A} w(f,\delta) < \epsilon.$$

  By definition, for any $x, y \in E$ such that $\rho(x,y) < \delta$,

  $$|f(x) - f(y)| \leq \sup_{g\in A} w(g,\delta) < \epsilon$$

  for any $f \in A$. Thsi holds for any $\epsilon > 0$, so A is uniformly equicontinuous on E.

The next theorem characterizes relative compactness on $\mathcal{C}_b(E,F)$ in terms of uniform equicontinuity for the case when E is compact.

Theorem 6.3 (The Arzela-Ascoli Theorem)
Let (E,ρ) be a compact metric space, and $F=\mathbb{R}^n$ or $\mathbb{C}$. A subset A of $\mathcal{C}(E,F)$ is relatively compact if and only if

i) For any $x\in E$, $\sup_{f\in A}|f(x)|<+\infty$, and

ii) A is uniformly equicontinuous on E.

Furthermore, in this case $\sup_{f\in A}\|f\|_{\mathcal{C}}<+\infty$ as well.

Proof) **Sufficiency**

Note first that, for any $q\in\mathbb{Q}$, the collection $\{B_\rho(x,q)\}_{x\in E}$ is an open cover of E; by the compactness of E, there exists a finite collection E_q of points in E such that

$$E\subset\bigcup_{x\in E_q}B_\rho(x,q).$$

Define $E^0=\bigcup_{q\in\mathbb{Q}}E_q$; E^0 is clearly a countable subset of E.

Suppose that A is uniformly equicontinuous on E and that

$$\sup_{f\in A}|f(x)|<+\infty$$

for any $x\in E$.
Choose any sequence $\{f_n\}_{n\in N_+}$ of continuous functions in A. Since $\{f_n\}_{n\in N_+}$ is pointwise bounded and E^0 is countable, there exists a subsequence $\{f_{n_k}\}_{k\in N_+}$ of $\{f_n\}_{n\in N_+}$ that converges pointwise on E^0. The proof will be complete if we can show that $\{f_{n_k}\}_{k\in N_+}$ is Cauchy in d; then, the completeness of $\mathcal{C}(E,F)$ ensures that the sequence has a limit in $\mathcal{C}(E,F)$.

By the uniform equicontinuity of A, for any $\epsilon>0$ there exists a $\delta>0$ such that, for any $x,y\in E$ satisfying $\rho(x,y)<\delta$,

$$|f(x)-f(y)|<\frac{\epsilon}{3}$$

for any $f\in A$. Choosing $q\in\mathbb{Q}$ so that $q<\delta$, by construction,

$$E\subset\bigcup_{y\in E_q}B_\rho(y,q),$$

where E_q is finite. Since $\{f_{n_k}(y)\}_{k\in N_+}$ is convergent and thus Cauchy with respect to the euclidean metric on F for any $y\in E_q$, the finiteness of E_q tells us that there exists an $N\in N_+$ such that

$$|f_{n_k}(y)-f_{n_m}(y)|<\frac{\epsilon}{3}$$

for any $k, m \geq N$ and $y \in E_q$.
Suppose $k, m \geq N$, where N depends only on ϵ (it also depends on δ, but δ depends only on ϵ). Choose any $x \in E$. Then, there exists a $y \in E_q$ such that $x \in B_\rho(y, q)$, which implies that

$$\rho(x, y) < q < \delta.$$

Therefore,

$$|f_{n_k}(x) - f_{n_k}(y)| < \frac{\epsilon}{3} \quad \text{and} \quad |f_{n_m}(x) - f_{n_m}(y)| < \frac{\epsilon}{3}$$

by the equicontinuity result above, and putting everything together,

$$|f_{n_k}(x) - f_{n_m}(x)| \leq |f_{n_k}(x) - f_{n_k}(y)| + |f_{n_k}(y) - f_{n_m}(y)| + |f_{n_m}(x) - f_{n_m}(y)| < \epsilon.$$

This holds for any $x \in E$, which tells us that

$$d(f_{n_k}, f_{n_m}) = \sup_{x \in E} |f_{n_k}(x) - f_{n_m}(x)| \leq \epsilon$$

for any $k, m \geq N$. This in turn holds for any $\epsilon > 0$, so $\{f_{n_k}\}_{k \in N_+}$ is Cauchy with respect to d.

Necessity
Turning now to necessity, suppose that A is relatively compact. For any $\epsilon > 0$, $\{B_d(f, \epsilon)\}_{f \in \overline{A}}$ is an open cover of the compact set $\overline{A}$; by compactness, there exist $f_1, \cdots, f_n \in \overline{A}$ such that

$$A \subset \overline{A} \subset \bigcup_{i=1}^{n} B_d(f_i, \epsilon).$$

Each f_i is bounded because E is a compact set and f_i is continuous; letting

$$M = \max_{1 \leq i \leq n} \|f_i\|_{\mathcal{C}} < +\infty,$$

for any $f \in A$ there exists an $1 \leq i \leq n$ such that $d(f, f_i) < \epsilon$, which implies that

$$\|f\|_{\mathcal{C}} \leq d(f, f_i) + \|f_i\|_{\mathcal{C}} \leq \epsilon + M.$$

Therefore,

$$\sup_{f \in A} \|f\|_{\mathcal{C}} \leq M + \epsilon < +\infty,$$

from which it follows that

$$\sup_{f\in A}|f(x)|<+\infty$$

for any $x\in A$.
To show that A is uniformly equicontinuous, note that, $\{f_1,\cdots,f_n\}$ is a finite collection of functions that are uniformly continuous on the compact set E. Therefore, there exists a $\delta>0$ such that

$$w(f_i,\delta)<\epsilon$$

for $1\leq i\leq n$. Choosing any $f\in A$ and letting $f\in B_d(f_i,\epsilon)$ for some $1\leq i\leq n$, it now follows that, for any $x,y\in E$ such that $\rho(x,y)<\delta$,

$$|f(x)-f_i(x)|,|f(y)-f_i(y)|\leq d(f,f_i)<\epsilon$$

and

$$|f_i(x)-f_i(y)|<\epsilon,$$

so that

$$|f(x)-f(y)|\leq|f(x)-f_i(x)|+|f_i(x)-f_i(y)|+|f(y)-f_i(y)|<3\epsilon.$$

Thus,

$$w(f,\delta)\leq 3\epsilon,$$

and because this holds for any $f\in A$,

$$\sup_{f\in A}w(f,\delta)\leq 3\epsilon.$$

Such a $\delta>0$ exists for any $\epsilon>0$, so by definition

$$\lim_{h\to 0}\sup_{f\in A}w(f,h)=0.$$

Q.E.D.

6.2 The Stone-Weierstrass Theorem

We take a brief detour to focus on the Stone-Weierstrass theorem, which allows us to approximate all kinds of functions with simpler and more elementary functions, such as polynomials. This result is then used to prove the separability of spaces of continuous functions defined on compact metric spaces.

6.2.1 The Weierstrass Approximation Theorem

We first state and prove the Weierstrass approxmiation theorem, which the Stone-Weierstrass theorem generalizes. It states that any complex continuous function on a closed interval can be uniformly approximated by a sequence of polynomial functions on the real line.

Theorem 6.4 (Weierstrass Approximation Theorem)
For any continuous complex valued function $f : [a,b] \to \mathbb{C}$ there exists a sequence of polynomial functions $\{P_n\}_{n\in N_+}$ on $\mathbb{R}$ such that $P_n \to f$ uniformly. Moreover, each P_n can be chosen to be real-valued if f is.

Proof) We gradually generalize the types of functions for which the above theorem holds.

Stage 1: $[a,b] = [0,1]$ and $f(0) = f(1) = 0$

Assume initially that $[a,b] = [0,1]$ and that $f(0) = f(1) = 0$. In this case, we can extend the domain of f to $\mathbb{R}$ by assuming $f(x) = 0$ for any $x \notin [0,1]$, and f remains a continuous function. In fact, f is a continuous function on $\mathbb{R}$ with support contained in the compact set $[0,1]$, so it is uniformly continuous on the real line. By the extreme value theorem, f is bounded on the compact set $[0,1]$ because it is continuous; let $|f(x)| < M$ for any $x \in \mathbb{R}$, where $M < +\infty$.

For any $n \in N_+$, since the mapping $x \mapsto (1-x^2)^n$ is non-negative polynomial function on $[-1,1]$, if

$$\int_{-1}^{1} (1-x^2)^n dx = 0,$$

then by the vanishing property for non-negative functions $x^2 = 1$ for any $x \in [-1,1]$, a contradiction. Therefore, we can define

$$c_n = \left(\int_{-1}^{1} (1-x^2)^n dx \right)^{-1} > 0.$$

Define the function $h : \mathbb{R} \to \mathbb{R}$ as

$$h(x) = (1-x^2)^n - 1 + nx^2$$

for any $x \in \mathbb{R}$. Then, for any $x \in (0,1)$,

$$h'(x) = -2nx(1-x^2)^{n-1} + 2nx = 2nx\left(1-(1-x^2)^{n-1}\right) > 0$$

and $h(0) = 0$, which tells us that $h(x) > h(0) = 0$ for any $x > 0$. In other words,

$$(1-x^2)^n \geq 1 - nx^2$$

for any $x \in [0,1]$. This result allows us to find the upper bound of c_n as follows:

$$\begin{aligned}
\int_{-1}^{1}(1-x^2)^n dx &= 2\int_0^1 (1-x^2)^n dx && \text{(The integrand is an even function)} \\
&\geq 2\int_0^{\frac{1}{\sqrt{n}}} (1-x^2)^n dx \\
&\geq 2\int_0^{\frac{1}{\sqrt{n}}} (1-nx^2) dx && \left((1-x^2)^n \geq 1-nx^2 \text{ on } [0,1]\right) \\
&= 2\left[x - \frac{n}{3}x^3\right]_0^{\frac{1}{\sqrt{n}}} = 2n^{-\frac{1}{2}} - \frac{2n}{3}n^{-\frac{3}{2}} \\
&= \frac{4}{3\sqrt{n}} > \frac{1}{\sqrt{n}},
\end{aligned}$$

which implies that

$$\sqrt{n} > \left(\int_{-1}^{1}(1-x^2)^n dx\right)^{-1} = c_n > 0.$$

Using this c_n, we define

$$Q_n(x) = c_n(1-x^2)^n$$

for any $x \in \mathbb{R}$, so that Q_n is non-negative on $[-1,1]$ with integral

$$\int_{-1}^{1} Q_n(x)dx = 1.$$

We now use the Q_n defined above to construct the sequence of polynomials that converges uniformly to f. We proceed in smaller steps.

Step 1: Construction of Polynomials P_n

For any $n \in N_+$, define the function $P_n : [0,1] \to \mathbb{C}$ as

$$P_n(x) = \int_{-1}^{1} f(x+t)Q_n(t)dt$$

for any $x \in [0,1]$. This integral is well-defined because f is bounded on $\mathbb{R}$, Q_n is bounded on $[-1,1]$ (being a polynomial, it is continuous and thus bounded on a compact set) and the Lebesgue measure is finite on $[-1,1]$.

To verify that P_n is indeed a polynomial function, we employ a linear change of variables to see that, for any $x \in [0,1]$,

$$\begin{aligned}
P_n(x) &= \int_{-x}^{1-x} f(x+t)Q_n(t)dt \qquad (f = 0 \text{ outside } [0,1] \text{ and } [-x,1-x] \subset [-1,1]) \\
&= \int_{\mathbb{R}} f(x+t)Q_n(t)I_{[-x,1-x]}(t)dt \\
&= \int_{\mathbb{R}} f(t)Q_n(t-x)I_{[-x,1-x]}(t-x)dt \qquad \text{(Linear Change of Variables)} \\
&= \int_0^1 f(t)Q_n(t-x)dt = c_n \int_0^1 f(t)((1-t^2)+2tx-x^2)^n dt.
\end{aligned}$$

Expanding the last term and evaluating the integrals with respect to t (which are finite because of the boundedness of f and the Lebesgue measure on $[0,1]$) yields a polynomial of degree $2n$ with respect to x, so it follows that P_n is indeed a polynomial function on $[0,1]$. Thus, there exist constants $a_0^{(n)}, \cdots, a_{2n}^{(n)} \in \mathbb{C}$ such that

$$P_n(x) = \sum_{i=0}^{2n} a_i^{(n)} \cdot x^i$$

for any $x \in [0,1]$. We can now define

$$P_n(x) = \sum_{i=0}^{2n} a_i^{(n)} \cdot x^i$$

for any $x \in \mathbb{R}$, which extends P_n to a polynomial function on the real line. Note that the coefficients $a_0^{(n)}, \cdots, a_{2n}^{(n)}$ are real if f is real-valued; thus, P_n can be taken to be a real polynomial if f is real-valued.

Step 2: Uniform Convergence of P_n

Note that, for any $\epsilon > 0$, the uniform continuity of f on $\mathbb{R}$ tells us that there exists a $\delta > 0$ such that

$$|f(x) - f(y)| < \epsilon$$

for any $x, y \in \mathbb{R}$ such that $|x - y| < \delta$. Now choose any $x \in [0,1]$, $n \in N_+$ and note that

$$\begin{aligned} |P_n(x) - f(x)| &= \left| \int_{-1}^{1} f(x+t) Q_n(t) dt - \int_{-1}^{1} f(x) Q_n(t) dt \right| \leq \int_{-1}^{1} |f(x+t) - f(x)| Q_n(t) dt \\ &= \int_{-1}^{-\delta} |f(x+t) - f(x)| Q_n(t) dt + \int_{\delta}^{1} |f(x+t) - f(x)| Q_n(t) dt \\ &\quad + \int_{-\delta}^{\delta} |f(x+t) - f(x)| Q_n(t) dt. \end{aligned}$$

Since

$$|f(x+t) - f(x)| < \epsilon$$

for any $|t| < \delta$,

$$0 \leq Q_n(t) = c^n (1 - t^2)^n \leq \sqrt{n} (1 - \delta^2)^n$$

for any $t \in [-1, 1]$ such that $|t| \geq \delta$, and

$$\int_{-\delta}^{\delta} Q_n(t) dt \leq \int_{-1}^{1} Q_n(t) dt = 1,$$

we can see that

$$|P_n(x) - f(x)| \leq 4M \cdot \sqrt{n} (1 - \delta^2)^n + \epsilon \cdot \int_{-\delta}^{\delta} Q_n(t) dt \leq 4M \cdot \sqrt{n} (1 - \delta^2)^n + \epsilon.$$

This holds for any $x \in [0,1]$, so

$$\sup_{x \in [0,1]} |P_n(x) - f(x)| \leq 4M \cdot \sqrt{n} (1 - \delta^2)^n + \epsilon$$

for any $n \in N_+$. Taking $n \to \infty$ on both sides, we can see that

$$\limsup_{n \to \infty} \sup_{x \in [0,1]} |P_n(x) - f(x)| \leq \epsilon.$$

This in turn holds for any $\epsilon > 0$, so

$$\lim_{n \to \infty} \sup_{x \in [0,1]} |P_n(x) - f(x)| = 0,$$

which tells us that $P_n \to f$ uniformly on $[0,1]$.

Stage 2: $[a,b] = [0,1]$ but arbitrary $f(0), f(1)$

Given a continuous complex valued function $f : [0,1] \to \mathbb{C}$, define $g : [0,1] \to \mathbb{C}$ as

$$g(x) = f(x) - f(0) + x \cdot (f(0) - f(1))$$

for any $x \in [0,1]$. g, being the sum of two continuous functions, is a continuous function on $[0,1]$, and $g(0) = g(1) = 0$ by construction. By the preceding result, there exists a sequence $\{P_n\}_{n \in N_+}$ of polynomials (that are real if g is real) such that

$$\sup_{x \in [0,1]} |g(x) - P_n(x)| \to 0$$

as $n \to \infty$. Since

$$f(x) - g(x) = (f(1) - f(0))x + f(0)$$

is a polynomial with respect to x, so is

$$R_n(x) = P_n(x) + f(x) - g(x).$$

for any $n \in N_+$. Moreover, if f is real, then so is g, which implies that R_n is a real polynomial function. For any $x \in [0,1]$ and $n \in N_+$,

$$|f(x) - R_n(x)| = |g(x) - P_n(x)|,$$

so it follows that

$$\sup_{x \in [0,1]} |f(x) - R_n(x)| \to 0$$

as $n \to \infty$; $\{R_n\}_{n \in N_+}$ is a sequence of polynomial functions that converges uniformly to f on $[0,1]$, and are real valued if f is real.

Stage 3: Arbitrary $[a,b]$ and $f(0), f(1)$

Now we generalize this result to continuous complex valued functions $f : [a,b] \to \mathbb{C}$ for arbitrary closed intervals $[a,b]$. Define $g : [0,1] \to \mathbb{C}$ as

$$g(x) = f\left((b-a)x + a\right)$$

for any $x \in [0,1]$; g, being the composition of two continuous functions, is itself continuous, so by the preceding result there exists a sequence of polynomial functions $\{P_n\}_{n \in N_+}$ of polynomials on $[0,1]$ such that $P_n \to g$ uniformly on $[0,1]$ and are real valued if g is real. For any $n \in N_+$, $R_n : \mathbb{R} \to \mathbb{C}$ defined as

$$R_n(x) = P_n\left(\frac{x-a}{b-a}\right)$$

for any $x \in \mathbb{R}$ is a polynomial (the composition of polynomials is a polynomial), and it is real valued if f is real valued (since g is also real valued in this case). We can now see that

$$\sup_{x \in [a,b]} |f(x) - R_n(x)| = \sup_{x \in [a,b]} \left| g\left(\frac{x-a}{b-a}\right) - P_n\left(\frac{x-a}{b-a}\right) \right| \leq \sup_{x \in [0,1]} |g(x) - P_n(x)|$$

for any $n \in N_+$, so $R_n \to f$ uniformly on $[0,1]$.
Q.E.D.

The following corollary of the approximation theorem is of particular interest:

Corollary to the Approximation Theorem For any $a > 0$ there exists a sequence $\{P_n\}_{n \in N_+}$ of real valued polynomial functions on $\mathbb{R}$ such that $P_n(0) = 0$ and

$$\lim_{n \to \infty} \sup_{x \in [-a,a]} |P_n(x) - |x|| = 0,$$

that is, $P_n(x) \to |x|$ uniformly on $[-a,a]$ as $n \to \infty$.

Proof) Define $f : [-a,a] \to \mathbb{R}$ as $f(x) = |x|$ for any $x \in [-a,a]$. Since f is a real valued continuous function, by the Weierstrass approximation theorem there exists a sequence $\{P_n^*\}_{n \in N_+}$ of real polynomial functions on $\mathbb{R}$ such that $P_n^* \to f$ uniformly on $[-a,a]$.
For any $n \in N_+$, define $P_n : \mathbb{R} \to \mathbb{R}$ as $P_n(x) = P_n^*(x) - P_n^*(0)$ for any $x \in \mathbb{R}$. Then, $\{P_n\}_{n \in N_+}$ is a sequence of real polynomial functions on $\mathbb{R}$ such that $P_n(0) = 0$, and

$$\sup_{x \in [-a,a]} |P_n(x) - f(x)| \leq \sup_{x \in [-a,a]} |P_n^*(x) - f(x)| + |P_n^*(0)|.$$

Since $P_n^*(0) \to f(0) = 0$ as $n \to \infty$, it follows that $P_n \to f$ uniformly on $[-a,a]$.
Q.E.D.

6.2.2 Algebras Over a Field

The central objects of Stone's generalization of the approximation theorem are algebras over a field, not to be confused with algebras on a set, which is a measure-theoretic structure containing the entire set and closed under complements and finite unions. An algebra $\mathcal{A}$ over a field F is a vector space over F that is also closed under the additional operation of elementwise multiplication $\times$, which satisfies the following axioms:

- **The Distributive Law**
  For any $x, y, z \in \mathcal{A}$,

$$(x+y) \times z = x \times z + y \times z \quad \text{and}$$
$$z \times (x+y) = z \times x + z \times y.$$

- **Commutativity with Scalar Multiplication**
  For any $\alpha, \beta \in F$ and $x, y \in \mathcal{A}$,

$$(\alpha x) \times (\beta y) = (\alpha\beta)(x \times y).$$

The main types of algebras over a field that are of interest are spaces of functions. Let E be a set, F a field, and $\mathcal{F}$ the collection of all functions from E to F. Then, we already know that $\mathcal{F}$ is a vector space over F.
We can extend $\mathcal{F}$ into an algebra over F by defining the multiplicative operation $\times$ on $\mathcal{F}$ as follows: for any $f, g \in \mathcal{F}$, $f \times g$ is the function on E defined as

$$(f \times g)(x) = f(x)g(x) \in F$$

for any $x \in E$. Note that this operation is well-defined because the target space F of the functions in $\mathcal{F}$ is a field and element-wise multiplication is defined on F. This operation also satisfies the two axioms above:

- For any $f, g, h \in \mathcal{F}$,

$$((f+g) \times h)(x) = (f(x)+g(x))h(x) = f(x)h(x) + g(x)h(x) = (f \times h)(x) + (g \times h)(x)$$

  for any $x \in E$, so that

$$(f+g) \times h = f \times h + g \times h.$$

  The left distributive law can be similarly shown.

- For any $f, g \in \mathcal{F}$ and $a, b \in F$,

$$((af) \times (bg))(x) = (af(x)) \cdot (bg(x)) = (ab) \cdot f(x)g(x) = ((ab) \cdot (f \times g))(x)$$

for any $x \in E$, so that

$$(af) \times (bg) = (ab) \cdot (f \times g).$$

Therefore, $\mathcal{F}$ is a well-defined algebra over the field F. We write fg instead of $f \times g$ for notational brevity.
Often, we will be interested in subsets of $\mathcal{F}$ that are themselves algebras over F. The two subsets of interest are given below:

- **The Algebra of Bounded Functions**
  Let $\mathcal{F}$ collect all bounded functions from E to F, where $F = \mathbb{R}$ or $\mathbb{C}$. Then, $\mathcal{F}$ is an algebra over F because it contains the zero function and is closed under addition (the sum of bounded functions is bounded), scalar multiplication and element-wise multiplication (the product of bounded functions is bounded).

- **The Algebra of Bounded Continuous Functions**
  Let (E, τ) be a topological space and $F = \mathbb{R}$ or $\mathbb{C}$. Since both $\mathbb{R}$ and $\mathbb{C}$ are fields, $\mathcal{C}_b(E, F)$ is a subspace of the algebra $\mathcal{F}$ over the field F. Furthermore, since the zero function is contained in $\mathcal{C}_b(E, F)$ and the product of any two bounded continuous functions is also bounded and continuous, $\mathcal{C}_b(E, F)$ is a subalgebra of $\mathcal{F}$ over the field F.

Let $\mathcal{F}$ once again denote the collection of all functions from some set E into the field $F = \mathbb{R}$ or $\mathbb{C}$. The uniform closure of a subalgebra $\mathcal{A}$ of $\mathcal{F}$ is defined as the set of all functions $f \in \mathcal{F}$ such that there exists a sequence $\{f_n\}_{n \in N_+} \subset A$ that converges uniformly to f. If $F = \mathbb{R}$ or $\mathbb{C}$ and A were a subset of the collection of all bounded continuous functions on E, then the uniform closure would be the closure of A with respect to the supremum metric, hence the name uniform "closure".
An important result is that the uniform closure of a subalgebra $\mathcal{A}$ is also a subalgebra of $\mathcal{F}$, given that the functions in $\mathcal{F}$ are bounded.

Lemma 6.5 For any set E and field $F = \mathbb{R}$ or $\mathbb{C}$, let $\mathcal{F}$ be the collection of all functions from E to F. Then, the uniform closure of any subalgebra $\mathcal{A}$ of $\mathcal{F}$ is a linear subspace of $\mathcal{F}$.
If $\mathcal{F}$ instead collects all bounded functions from E to F, then the uniform closure of $\mathcal{A}$ is a subalgebra of $\mathcal{F}$.

Proof) Denote the uniform closure of $\mathcal{A}$ by $\overline{\mathcal{A}}$. Because the zero function is contained in $\mathcal{A}$ (it is a subalgebra of $\mathcal{F}$), it is also contained in $\overline{\mathcal{A}}$. It remains to see whether $\overline{\mathcal{A}}$ is closed under addition, scalar multiplication and element-wise multiplication; the first

two ensure that it is a vector space over F, and the last condition is necessary for $\overline{\mathcal{A}}$ to be considered a subalgebra.

Choose any $f, g \in \overline{\mathcal{A}}$ and $z \in F$. Then, by definition there exist sequences $\{f_n\}_{n \in N_+}$ and $\{g_n\}_{n \in N_+}$ such that $f_n \to f$ and $g_n \to g$ uniformly. For any $n \in N_+$,

$$z \cdot f_n + g_n, f_n g_n \in \mathcal{A}$$

because $\mathcal{A}$ is a subalgebra of $\mathcal{F}$.
Since

$$|(z \cdot f_n(x) + g_n(x)) - (z \cdot f(x) + g(x))| \leq |z| \cdot |f_n(x) - f(x)| + |g_n(x) - g(x)|$$

for any $x \in E$, the uniform convergence of $\{f_n\}_{n \in N_+}$ and $\{g_n\}_{n \in N_+}$ implies that the sequence $\{z \cdot f_n + g_n\}_{n \in N_+}$ is a sequence of functions in $\mathcal{A}$ that converges uniformly to $z \cdot f + g$. Therefore, $z \cdot f + g \in \overline{\mathcal{A}}$, and it follows that $\overline{\mathcal{A}}$ is a vector space over F.

Suppose that $\mathcal{F}$ collects all bounded functions from E to F, and that $\mathcal{A}$ is a subalgebra of this algebra over F. Then, $f, g \in \mathcal{F}$ are bounded; suppose $|f|, |g| \leq M$ on E for some $M \in (0, +\infty)$. Then, for any $x \in E$,

$$\begin{aligned} |f_n(x)g_n(x) - f(x)g(x)| &\leq |f_n(x) - f(x)||g_n(x) - g(x)| + |f(x)||g_n(x) - g(x)| + |g(x)||f_n(x) - f(x)| \\ &\leq |f_n(x) - f(x)||g_n(x) - g(x)| + M\left(|g_n(x) - g(x)| + |f_n(x) - f(x)|\right). \end{aligned}$$

Therefore, the uniform convergence of $\{f_n\}_{n \in N_+}$ and $\{g_n\}_{n \in N_+}$ implies that the sequence $\{f_n g_n\}_{n \in N_+}$ is a sequence of functions in $\mathcal{A}$ that converges uniformly to fg. It follows that $\overline{\mathcal{A}}$ is a subalgebra, in addition to being a subspace, of $\mathcal{F}$.
Q.E.D.

Let E be a set, F a field with additive identity 0_F, and $\mathcal{F}$ the algebra of all functions from E to F. Letting $\mathcal{A}$ be a subalgebra of $\mathcal{F}$, we say that $\mathcal{A}$:

- **Separates points on E**
  If, for any $x_1, x_2 \in E$ such that $x_1 \neq x_2$, there exists an $f \in \mathcal{A}$ such that $f(x_1) \neq f(x_2)$.

- **Vanishes at no point in E**
  If, for any $x \in E$, there exists an $f \in \mathcal{A}$ such that $f(x) \neq 0_F$.

The following result is an important consequence of the above properties:

Lemma 6.6 For any set E and field F with additive identity 0_F, let $\mathcal{F}$ be the collection of all functions from E to F and $\mathcal{A}$ a subalgebra of $\mathcal{F}$. If $\mathcal{A}$ separates points on E and vanishes at no point in E, then for any $c_1, c_2 \in F$ and distinct points $x_1, x_2 \in E$ there exists an $f \in \mathcal{A}$ such that

$$f(x_1) = c_1 \quad \text{and} \quad f(x_2) = c_2.$$

Proof) Choose any $c_1, c_2 \in F$ and $x_1, x_2 \in E$ such that $x_1 \neq x_2$. Because $\mathcal{A}$ separates points on E and vanishes at no point in E, there exist $g, h_1, h_2 \in \mathcal{A}$ such that

$$g(x_1) \neq g(x_2), \quad h_1(x_1) \neq 0_F, \quad h_2(x_2) \neq 0_F.$$

Denote $a = g(x_1) - g(x_2) \neq 0_F$. Then, defining $f : E \to F$ as

$$f = \frac{c_1}{a \cdot h_1(x_1)}(gh_1) - \frac{c_1 \cdot g(x_2)}{a \cdot h_1(x_1)}h_1 - \frac{c_2}{a \cdot h_2(x_2)}(gh_2) + \frac{c_2 \cdot g(x_1)}{a \cdot h_2(x_2)}h_2,$$

since $gh_1, gh_2, h_1, h_2 \in \mathcal{A}$ and f is a linear combination of these functions, $f \in \mathcal{A}$. Moreover,

$$f(x_1) = \frac{c_1}{a}g(x_1) - \frac{c_1 \cdot g(x_2)}{a} = c_1\frac{g(x_1) - g(x_2)}{a} = c_1,$$

and likewise, $f(x_2) = c_2$.
Q.E.D.

6.2.3 Stone's Generalization of the Weierstrass Approximation Theorem

Here we present the main result of this section, which extends the Weierstrass approximation theorem to functions in arbitrary algebras that satisfy the given properties. We first formulate the result for real-valued functions:

Theorem 6.7 (The Stone-Weierestrass Theorem: Real Version)
Let (E,τ) be a topological space, and $\mathcal{A}$ a subalgebra of the algebra $\mathcal{C}(E,\mathbb{R})$ over the real field. If $\mathcal{A}$ separates points on E, then (E,τ) is a Hausdorff space.
If, in addition, $\mathcal{A}$ vanishes at no point in E and (E,τ) is a compact space, then $\mathcal{C}(E,\mathbb{R})$ is the uniform closure of $\mathcal{A}$, or in other words, the closure of $\mathcal{A}$ with respect to the supremum metric d on $\mathcal{C}(E,\mathbb{R})$.

Proof) We first show that (E,τ) is a Hausdorff space if $\mathcal{A}$ separates points on E.
Choose any distinct points $x_1,x_2 \in E$. Then, because $\mathcal{A}$ separates points on E, there exists an $f \in \mathcal{A}$ such that $f(x_1) \neq f(x_2)$. Because f is continuous ($\mathcal{A}$ is a subalgebra of the space of all real continuous functions on E), defining $\epsilon = \frac{|f(x_1)-f(x_2)|}{2} > 0$, the inverse image

$$V_i = f^{-1}\left(B_{|\cdot|}(f(x_i),\epsilon)\right)$$

is an open subset of E that contains x_i for $i = 1,2$. Furthermore, if $t = V_1 \cap V_2$, then

$$|f(t)-f(x_1)| < \epsilon \quad \text{and} \quad |f(t)-f(x_2)| < \epsilon,$$

which implies that

$$|f(t)-f(x_2)| \geq |f(x_1)-f(x_2)| - |f(t)-f(x_1)| > |f(x_1)-f(x_2)| - \epsilon = \epsilon,$$

a contradiction. Therefore, V_1, V_2 are disjoint open subsets of E such that $x_1 \in V_1$ and $x_2 \in V_2$, so that (E,τ) is Hausdorff by definition.

Now we can show the second claim. Suppose $\mathcal{A}$ separates points on E, vanishes at no point in E and that (E,τ) is a compact topological space.
As above, let $\overline{\mathcal{A}}$ denote the uniform closure of $\mathcal{A}$. Note initially that, because $\mathcal{C}(E,\mathbb{R})$ is a collection of bounded real valued functions on E due to the compactness of E and the continuity of the functions comprising $\mathcal{C}(E,\mathbb{R})$, by lemma 6.5 $\overline{\mathcal{A}}$ is a subalgebra of $\mathcal{C}(E,\mathbb{R})$. Thus, it remains to verify that $\mathcal{C}(E,\mathbb{R})$ is a subalgebra of $\overline{\mathcal{A}}$.
We proceed in steps:

Step 1: If $f \in \overline{\mathcal{A}}$, then $|f| \in \overline{\mathcal{A}}$

Choose any $f \in \overline{\mathcal{A}}$. Since f is a bounded continuous function, there exists an $M \in (0,+\infty)$ such that $|f| < M$ on E. By the corollary to the Weierstrass approximation

theorem, there exists a sequence $\{P_n\}_{n\in N_+}$ of real polynomial functions on $\mathbb{R}$ such that $P_n(0)=0$ for any $n\in N_+$ and

$$\sup_{x\in[-M,M]}|P_n(x)-|x||\to 0$$

as $n\to\infty$. For any $n\in N_+$, because P_n is a polynomial function on $\mathbb{R}$, there exist $a_0,\cdots,a_m\in\mathbb{R}$ such that

$$P_n(x)=\sum_{i=0}^{m}a_i\cdot x^i$$

for any $x\in\mathbb{R}$; $P_n(0)=0$ implies that $a_0=0$, so that

$$P_n\circ f=\sum_{i=1}^{m}a_i\cdot f^i.$$

Because $\overline{\mathcal{A}}$ is an algebra over the real field, $f,f^1,\cdots,f^m\in\overline{\mathcal{A}}$, and since $\overline{\mathcal{A}}$ is closed under linear combinations of its elements, $P_n\circ f\in\overline{\mathcal{A}}$.
This holds for any $n\in N_+$, so for any $x\in E$,

$$|(P_n\circ f)(x)-|f(x)||\le\sup_{y\in[-M,M]}|P_n(y)-|y||$$

since $f(x)\in[-M,M]$, which implies that

$$d(P_n\circ f,|f|)\le\sup_{y\in[-M,M]}|P_n(y)-|y||.$$

Taking $n\to\infty$ on both sides, we can see that $|f|$ is the uniform limit of the sequence $\{P_n\circ f\}_{n\in N_+}$ of bounded continuous functions in $\overline{\mathcal{A}}$. Since $\overline{\mathcal{A}}$ is closed with respect to the supremum metric d, this implies that $|f|\in\overline{\mathcal{A}}$.

Step 2: If $f,g\in\overline{\mathcal{A}}$, then $\max(f,g),\min(f,g)\in\overline{\mathcal{A}}$

Choose any $f,g\in\overline{\mathcal{A}}$. Then,

$$\max(f,g)=\frac{f+g}{2}+\frac{|f-g|}{2}\quad\text{and}\quad\min(f,g)=\frac{f+g}{2}-\frac{|f-g|}{2},$$

where $f+g,f-g,|f-g|\in\overline{\mathcal{A}}$ by step 1 and the fact that $\overline{\mathcal{A}}$ is an algebra over $\mathbb{R}$, so $\max(f,g),\min(f,g)\in\overline{\mathcal{A}}$.
This can easily be generalized to finite collections of functions in $\overline{\mathcal{A}}$. Specifically, if $\{g_1,\cdots,g_n\}\subset\overline{\mathcal{A}}$, then

$$\max(g_1,\cdots,g_n),\min(g_1,\cdots,g_n)\in\overline{\mathcal{A}}.$$

Step 3: Approximating $f \in \mathcal{C}(E, \mathbb{R})$ with functions in $\overline{\mathcal{A}}$

Let $f \in \mathcal{C}(E, \mathbb{R})$ and $\epsilon > 0$.
Choose any distinct $x, y \in E$. Then, because $\mathcal{A}$ separates points on E and vanishes at no point in E, there exists a function $h_y \in \mathcal{A}$ such that

$$h_y(x) = f(x) \quad \text{and} \quad h_y(y) = f(y).$$

By the continuity of h_y and thus $k_y = h_y - f$, the inverse image

$$J_y = k_y^{-1}\left(B_{|\cdot|}(0, \epsilon)\right)$$

is an open subset of E that contains y and x. In particular, for any $t \in J_y$,

$$|k_y(t)| = |h_y(t) - f(t)| < \epsilon,$$

and in particular $h_y(t) > f(t) - \epsilon$.

$\{J_y\}_{y \in E, y \neq x}$ is an open cover of E, and by the compactness of E, there exist $y_1, \cdots, y_n \in E$ such that $y_i \neq x$ for $1 \leq i \leq n$ and

$$E \subset J_{y_1} \cup \cdots \cup J_{y_n}.$$

Then, defining $g_x = \max(h_{y_1}, \cdots, h_{y_n}) \in \overline{\mathcal{A}}$,

$$g_x(x) = \max(h_{y_1}(x), \cdots, h_{y_n}(x)) = f(x)$$

and, for any $t \in E$ such that $t \neq x$, letting $t \in J_{y_i}$ for some $1 \leq i \leq n$,

$$g_x(t) \geq h_{y_i}(t) > f(t) - \epsilon.$$

By the continuity of g_x and thus $r_x = g_x - f$, the inverse image

$$P_x = r_x^{-1}\left(B_{|\cdot|}(0, \epsilon)\right)$$

is an open subset of E that contains x and such that, for any $t \in P_x$,

$$|r_x(t)| = |g_x(t) - f(t)| < \epsilon,$$

and in particular $f(t) + \epsilon > g_x(t)$.

The above holds for any $x \in E$, so the collection $\{P_x\}_{x \in E}$ is an open cover of E. By the

compactness of E again, there exist $x_1, \cdots, x_m \in E$ such that

$$E \subset P_{x_1} \cup \cdots \cup P_{x_m}.$$

Define $g = \min(g_{x_1}, \cdots, g_{x_m}) \in \overline{\mathcal{A}}$. Then, for any $t \in E$, because

$$g_{x_i}(t) > f(t) - \epsilon$$

for any $1 \leq i \leq m$ by construction, we have

$$g(x) = \min(g_{x_1}(t), \cdots, g_{x_m}(t)) > f(t) - \epsilon.$$

Furthermore, letting $t \in P_{x_i}$ for some $1 \leq i \leq m$,

$$g(t) \leq g_{x_i}(t) < f(t) + \epsilon,$$

so that

$$|f(t) - g(t)| < \epsilon.$$

This holds for any $t \in E$, so

$$d(f,g) = \sup_{t \in E} |f(t) - g(t)| \leq \epsilon.$$

We have thus shown that, for any $\epsilon > 0$, there exists a $g \in \overline{\mathcal{A}}$ such that $g \in B_d(f, \epsilon)$. In other words, f is contained in the closure of $\overline{\mathcal{A}}$ with respec to d; because $\overline{\mathcal{A}}$ is closed, it is equal to its closure and thus $f \in \overline{\mathcal{A}}$. Finally, this holds for any $f \in \mathcal{C}(E, \mathbb{R})$, so $\mathcal{C}(E, \mathbb{R}) = \overline{\mathcal{A}}$.
Q.E.D.

The above theorem only holds for algebras of continuous real-valued functions: for the result to hold for algebras of complex continuous functions, we require the additional condition of self-adjointness. An algebra $\mathcal{A}$ of functions from a set E to $\mathbb{C}$ is said to be self-adjoint if the conjugate of every function in $\mathcal{A}$ is also contained in $\mathcal{A}$, that is, if for any $f \in \mathcal{A}$, the function $\overline{f}$ defined as

$$\overline{f}(x) = \overline{f(x)}$$

for any $x \in E$ is also contained in $\mathcal{A}$.

Theorem 6.8 (The Stone-Weierestrass Theorem: Complex Version)
Let (E,τ) be a topological space, and $\mathcal{A}$ a subalgebra of the algebra $\mathcal{C}(E,\mathbb{C})$ over the complex field. If $\mathcal{A}$ separates points on E, then (E,τ) is a Hausdorff space.
If, in addition, $\mathcal{A}$ vanishes at no point in E, is self-adjiont, and (E,τ) is a compact space, then $\mathcal{C}(E,\mathbb{C})$ is the uniform closure of $\mathcal{A}$, or in other words, the closure of $\mathcal{A}$ with respect to the supremum metric d on $\mathcal{C}(E,\mathbb{C})$.

Proof) The fact that (E,τ) must be a Hausdorff space if $\mathcal{A}$ separates points follows from the continuity of the functions in $\mathcal{A}$ in the same manner as in the previous theorem.

Suppose now that $\mathcal{A}$ separates points on E, vanishes at no point in E, is self-adjoint, and that (E,τ) is a compact topological space. By lemma 4.5, we know once again that $\overline{\mathcal{A}}$ is a subalgebra of $\mathcal{C}(E,\mathbb{C})$.
Define $\mathcal{A}_{\mathbb{R}}$ as the collection of all real-valued functions in $\mathcal{A}$; $\mathcal{A}_{\mathbb{R}}$ is non-empty since it contains the zero function on E. In addition, for any $f \in \mathcal{A}$, since $\overline{f} \in \mathcal{A}$ as well,

$$Re(f) = \frac{f+\overline{f}}{2}, Im(f) = \frac{1}{2i}\left(f - \overline{f}\right) \in \mathcal{A}$$

as well, and because $Re(f), Im(f)$ are real-valued functions, they are contained in $\mathcal{A}_{\mathbb{R}}$.
We can show that $\mathcal{A}_{\mathbb{R}}$ separates points on E and vanishes at no point in E. For the first property, choose any $x_1, x_2 \in E$; because $\mathcal{A}$ separates points on E, there exists an $f \in \mathcal{A}$ such that $f(x_1) \neq f(x_2)$. This implies that either $Re(f)$ or $Im(f)$ must give different values for x_1 and x_2, so that $\mathcal{A}_{\mathbb{R}}$ also separates points on E.
Now choose any $x \in E$; because $\mathcal{A}$ vanishes at no point in E, there exists an $f \in \mathcal{A}$ such that $f(x) \neq 0$. This implies that the value of either $Re(f)$ or $Im(f)$ must be non-zero at x, so that $\mathcal{A}_{\mathbb{R}}$ also vanishes at no point in E.
By the real version of the Stone-Weierstrass theorem, it now follows that the uniform closure of $\mathcal{A}_{\mathbb{R}}$ is precisely the collection $\mathcal{C}(E,\mathbb{R})$ of all real-valued continuous functions on E.

Choose any $f \in \mathcal{C}(E,\mathbb{C})$. Then, $Re(f), Im(f)$ are real-valued continuous functions on E, so by the preceding result, there exist sequences $\{g_n\}_{n\in N_+}$ and $\{h_n\}_{n\in N_+}$ in $\mathcal{A}_{\mathbb{R}}$

such that

$$d(g_n, Re(f)), d(h_n, Im(f)) \to 0$$

as $n \to \infty$. Defining $f_n = g_n + i \cdot h_n$ for any $n \in N_+$, $f_n \in \mathcal{A}$ because $\mathcal{A}$ is closed under addition and scalar multiplication, and for any $x \in E$,

$$|f_n(x) - f(x)| \leq |g_n(x) - Re(f)(x)| + |h_n(x) - Im(f)(x)| \leq d(g_n, Re(f)) + d(h_n, Im(f)),$$

so that

$$d(f_n, f) \leq d(g_n, Re(f)) + d(h_n, Im(f)).$$

It follows that $\{f_n\}_{n \in N_+}$ is a sequence of functions in $\mathcal{A}$ that converges uniformly to f, so that, by definition, $f \in \overline{\mathcal{A}}$. This holds for any $f \in \mathcal{C}(E, \mathbb{C})$, so $\mathcal{C}(E, \mathbb{C}) = \overline{\mathcal{A}}$.
Q.E.D.

6.2.4 The Separability of Continuous Function Spaces

The Stone-Weierstrass Theorem can now be used to show that the space of continuous functions defined on compact metric spaces are separable. This result is proved by first constructing a countable collection of continuous functions using the separability of compact sets and the associated metric, and then showing that this collection is an algebra of continuous functions that separates points and does not vanish (if the algebra is over $\mathbb{C}$, we also show that it is self-adjoint). Then, the Stone-Weierstrass theorem can be used to establish the separability of the function space.

Theorem 6.9 (Separability of Continuous Function Spaces)
Let (E,ρ) is a compact metric space, and $F=\mathbb{R}^n$ or $\mathbb{C}$. Then, letting d be the supremum metric on $\mathcal{C}(E,F)$, the pair $(\mathcal{C}(E,F),d)$ defines a complete and separable metric space.

Proof) We showed earlier that $(\mathcal{C}(E,F),d)$ is a complete metric space. We must now show that it is separable. To this end, note that the compactness of E implies its separability; let E_0 be the countable subset of E that is dense in E.

First let $F=\mathbb{R}$ or $\mathbb{C}$, and define

$$\mathbb{B}=\{\rho(\cdot,x)\mid x\in E_0\}\cup\{\mathbf{0}\},$$

where $\mathbf{0}$ is the zero function. For any $x\in E$, the mapping $y\mapsto\rho(y,x)$ is uniformly continuous on E by the triangle inequality and takes values in $[0,+\infty)$, so $\mathbb{B}\subset\mathcal{C}(E,F)$. Every function in $\mathcal{C}(E,F)$ is bounded, so the functions in $\mathbb{B}$ are also bounded. Now define

$$\mathbb{B}'=\Big\{\prod_{i=1}^{n}f_i\mid f_1,\cdots,f_n\in\mathbb{B}\Big\},$$

or the collection of the products of all the functions in $\mathbb{B}$. Because of the finiteness of the products, $\mathbb{B}'$ continues to be a collection of bounded and continuous functions. Finally, define

$$\mathcal{A}=\Big\{\sum_{i=1}^{n}a_i\cdot f_i\mid n\in N_+,a_1,\cdots,a_n\in F\text{ and }f_1,\cdots,f_n\in\mathbb{B}'\Big\}.$$

Each function in $\mathcal{A}$ is a finite linear combination of continuous bounded functions, so $\mathcal{A}\subset\mathcal{C}(E,F)$. It is also the case that $\mathcal{A}$ is an algebra over F: for any functions

$$f=\sum_{i=1}^{n}a_i\cdot f_i,\quad g=\sum_{i=1}^{m}b_i\cdot g_i$$

in $\mathcal{A}$ and $z\in F$, $zf+g$ is a linear combination of the functions $f_1,\cdots,f_n,g_1,\cdots,g_m\in\mathbb{B}'$,

so $zf+g \in \mathcal{A}$. Likewise,

$$fg = \sum_{i=1}^{n}\sum_{j=1}^{m}(a_i b_j)\cdot f_i g_j,$$

where each $f_i g_j \in \mathbb{B}'$ and $a_i b_j \in F$, so $fg \in \mathcal{A}$ as well. Finally, the zero function is included in $\mathbb{B}$ and thus $\mathcal{A}$. By definition, $\mathcal{A}$ is a subalgebra of $\mathcal{C}(E,F)$.

We will now show that $\mathcal{A}$ separates points on E and vanishes at no point in E.
Choose any $x_1, x_2 \in E$ such that $x_1 \neq x_2$; then, because E_0 is dense in E, choosing $q \in \mathbb{Q}$ so that $q < \frac{\rho(x_1,x_2)}{2}$, there exists a $y \in E_0$ such that

$$\rho(x_1, y) < q.$$

It follows that

$$\rho(x_2,y) \geq \rho(x_1,x_2) - \rho(x_1,y) > \rho(x_1,x_2) - q > \frac{\rho(x_1,x_2)}{2} > q.$$

Defining $f : E \to \mathbb{R}$ as $f(x) = \rho(x,y)$ for any $x \in E$, $f \in \mathcal{A}$ because $y \in E_0$, and

$$f(x_1) = \rho(x_1,y) < q < \rho(x_2,y) = f(x_2),$$

so that $f(x_1) \neq f(x_2)$. This shows us that $\mathcal{A}$ separates points on E.
Next, choose some $x \in E$. Then, for any $y \in E_0$ such that $x \neq y$ (such a y exists because E contains more than one element), $d(\cdot, y) \in \mathcal{A}$ and $d(x,y) > 0$. This shows us that $\mathcal{A}$ vanishes at no point in E.
If $F = \mathbb{C}$, then $\mathcal{A}$ is also self-adjoint because each function in $\mathbb{B}'$ is real-valued.

By the Stone-Weierstrass theorem, the uniform closure $\overline{\mathcal{A}}$ equals the collection $\mathcal{C}(E,F)$ of all continuous functions on E taking values in F. Because $\mathcal{A}$ is a countable subset of $\mathcal{C}(E,F)$, by definition $(\mathcal{C}(E,F), d)$ is a separable metric space.

Suppose now that $F = \mathbb{R}^n$, and let

$$\mathcal{A} = \Big\{ \sum_{i=1}^{n} a_i \cdot f_i \mid n \in N_+, a_1, \cdots, a_n \in \mathbb{R} \text{ and } f_1, \cdots, f_n \in \mathbb{B}' \Big\}$$

as before. Defining

$$\mathcal{A}^n = \{ f : E \to \mathbb{R}^n \mid f_i \in \mathcal{A} \text{ for any } 1 \leq i \leq n \},$$

because each $\mathcal{A}$ is countable and consists of real continuous bounded functions on E, $\mathcal{A}^n$ is also a countable collection of n-dimensional real continuous bounded functions on E. It remains to show that $\mathcal{A}^n$ is dense in $\mathcal{C}(E,F)$.

Choose any $f \in \mathcal{C}(E,F)$. Then, for any $\epsilon > 0$ and $1 \leq i \leq n$, there exists a $g_i \in \mathcal{A}$ such that

$$\sup_{x \in E} |f_i(x) - g_i(x)| < \frac{\epsilon}{n}.$$

Defining $g = (g_1, \cdots, g_n) \in \mathcal{A}^n$, we now have

$$|f(x) - g(x)| \leq \sum_{i=1}^{n} |f_i(x) - g_i(x)| < \epsilon$$

for any $x \in E$, so that

$$d(f,g) = \sup_{x \in E} |f(x) - g(x)| \leq \epsilon.$$

This holds for any $\epsilon > 0$, so $\mathcal{A}^n$ is dense in $\mathcal{C}(E,F)$ and $(\mathcal{C}(E,F), d)$ is a separable metric space.
Q.E.D.

6.3 Borel σ-algebras on Continuous Function Spaces

Let (E,ρ) be a compact metric space, τ the topology induced by ρ and $\mathcal{E}$ the Borel σ-algebra generated by τ. For $F=\mathbb{R}^n$ or $\mathbb{C}$, we showed above that $(\mathcal{C}(E,F),d)$ is a complete and separable metric space, where d is the supremum metric on $\mathcal{C}(E,F)$. We denote by $\mathcal{B}_{\mathcal{C}}(E,F)$ the Borel σ-algebra generated by the topology on $\mathcal{C}(E,F)$ induced by d.
Below we study two topics closely related to $\mathcal{B}_{\mathcal{C}}(E,F)$. The first concerns a π-system that generates this σ-algebra. Afterwards, we investigate how to characterize tight sequences of probability measures on $(\mathcal{C}(E,F),\mathcal{B}_{\mathcal{C}}(E,F))$.

6.3.1 Finite Dimensional Sets

It is of interest further down the line to have a π-system that generates $\mathcal{B}_{\mathcal{C}}(E,F)$. This will be furnished in the form of the collection of finite-dimensional sets $\mathcal{C}_f$. First, we define what is meant by the projection of a function $f\in\mathcal{C}(E,F)$. Let $F=\mathbb{R}^n$ in what follows.
For any $t_1,\cdots,t_k\in E$ that may or may not be distinct, the projection function $\pi_{t_1,\cdots,t_k}:\mathcal{C}(E,F)\to F^k$ is defined as

$$\pi_{t_1,\cdots,t_k}(f)=(f(t_1),\cdots,f(t_k))$$

for any $f\in\mathcal{C}(E,F)$. $\pi_{t_1,\cdots,t_k}$ has the following basic propreties:

- **Uniform Continuity**
  For any $f,g\in\mathcal{C}(E,F)$,

$$\begin{aligned}|\pi_{t_1,\cdots,t_k}(f)-\pi_{t_1,\cdots,t_k}(g)| &= |(f(t_1)-g(t_1),\cdots,f(t_k)-g(t_k))|\\ &\leq \sum_{i=1}^{k}|f(t_i)-g(t_i)|\leq k\cdot d(f,g),\end{aligned}$$

  so that $\pi_{t_1,\cdots,t_k}$ is uniformly continuous on $\mathcal{C}(E,F)$.

- **Inverse Images of Projections**
  For any $A\in\mathcal{B}(F^k)$, where $\mathcal{B}(F^k)$ is the standard Borel σ-algebra on the euclidean space F^k,

$$\pi^{-1}_{t_1,\cdots,t_k}(A)\in\mathcal{B}_{\mathcal{C}}(E,F)$$

  because the continuity of $\pi_{t_1,\cdots,t_k}$ ensures that it is measurable relative to $\mathcal{B}_{\mathcal{C}}(E,F)$ and $\mathcal{B}(F^k)$.

The inverse images above are referred to as finite-dimensional sets, and the collection of all finite

dimensional sets is thus defined as

$$\mathcal{C}_f = \left\{ \pi^{-1}_{t_1,\cdots,t_k}(A) \mid t_1,\cdots,t_k \in E, A \in \mathcal{B}(F^k) \right\}.$$

We can show that $\mathcal{C}_f$ is a π-sysetm on $\mathcal{C}(E,F)$ that generates $\mathcal{B}_{\mathcal{C}}(E,F)$.

Theorem 6.10 Let (E,ρ) be a compact metric space, $F=\mathbb{R}^n$, d the supremum metric on $\mathcal{C}(E,F)$, and $\mathcal{B}_{\mathcal{C}}(E,F)$ the associated Borel σ-algebra on $\mathcal{C}(E,F)$. Then, the collection $\mathcal{C}_f$ of all finite-dimensional sets is a π-system that generates $\mathcal{B}_{\mathcal{C}}(E,F)$.

Proof) We first show that $\mathcal{C}_f$ is a π-system on $\mathcal{C}(E,F)$.
Choose any $H_1, H_2 \in \mathcal{C}_f$. There exist $t_1,\cdots,t_k \in E$ and $s_1,\cdots,s_m \in E$ and $A \in \mathcal{B}(F^k)$, $B \in \mathcal{B}(F^m)$ such that

$$H_1 = \pi^{-1}_{t_1,\cdots,t_k}(A) \quad \text{and} \quad H_2 = \pi^{-1}_{s_1,\cdots,s_m}(B).$$

Then, since $A \times B \in \mathcal{B}(F^k) \bigotimes \mathcal{B}(F^m) = \mathcal{B}(F^{k+m})$,

$$H_1 \cap H_2 = \pi^{-1}_{t_1,\cdots,t_k,s_1,\cdots,s_m}(A \times B) \in \mathcal{C}_f,$$

whcih tells us that $\mathcal{C}_f$ is a π-system.

Every set in $\mathcal{C}_f$ is contained in $\mathcal{B}_{\mathcal{C}}(E,F)$, so the σ-algebra $\sigma\mathcal{C}_f$ generated by $\mathcal{C}_f$ is contained in $\mathcal{B}_{\mathcal{C}}(E,F)$. We now show the reverse inclusion.
Choose any $f \in \mathcal{C}(E,F)$ and $\epsilon > 0$. By the compactness of E, E is separable and there exists a countable subset E_0 that is dense in E. It can then be seen that

$$\begin{aligned} \overline{B}_d(f,\epsilon) &= \{ g \in \mathcal{C}(E,F) \mid d(f,g) \le \epsilon \} \\ &= \bigcap_{x \in E_0} \{ g \in \mathcal{C}(E,F) \mid |f(x) - g(x)| \le \epsilon \}. \end{aligned}$$

To show this, first choose any $g \in \overline{B}_d(f,\epsilon)$. Then,

$$|f(x) - g(x)| \le d(f,g) \le \epsilon$$

for any $x \in E$, so $|f(x)-g(x)| \le \epsilon$ for any $x \in E_0$ and thus g is contained in the intersection on the right hand side.
Conversely, choose any $g \in \mathcal{C}(E,F)$ such that $|f(y)-g(y)| \le \epsilon$ for any $y \in E_0$. Then, for any $x \in E$, by the continuity of f and g at x, for any $\eta > 0$ there exists a $\delta > 0$ such that

$$|f(z) - f(x)|, |g(z) - g(x)| < \frac{\eta}{2}$$

for any $z \in E$ such that $\rho(x,z) < \delta$. Because E_0 is dense in E, there exists a $y \in E_0$ such

that $\rho(x,y) < \delta$, which implies that

$$|f(x) - g(x)| \leq |f(x) - f(y)| + |f(y) - g(y)| + |g(x) - g(y)| \leq \epsilon + \eta.$$

This holds for any $\eta > 0$, so

$$|f(x) - g(x)| \leq \epsilon,$$

and because this in turn holds for any $x \in E$,

$$d(f,g) = \sup_{x \in E} |f(x) - g(x)| \leq \epsilon,$$

or equivalently, $g \in \overline{B}_d(f,\epsilon)$.
The above equality can also be written as

$$\overline{B}_d(f,\epsilon) = \bigcap_{x \in E_0} \pi_x^{-1}\left(\overline{B}_{|\cdot|}(f(x),\epsilon)\right),$$

where

$$\overline{B}_{|\cdot|}(f(x),\epsilon) = \{y \in F \mid |f(x) - y| \leq \epsilon\} \in \mathcal{B}(F).$$

Since E_0 is countable and each $\pi_x^{-1}\left(\overline{B}_{|\cdot|}(f(x),\epsilon)\right) \in \mathcal{C}_f$, it follows that

$$\overline{B}_d(f,\epsilon) \in \sigma\mathcal{C}_f.$$

Note that any open ball $B_d(f,\epsilon)$ is the countable union of closed balls:

$$B_d(f,\epsilon) = \bigcup_n \overline{B}_d\left(f, \epsilon - \frac{1}{n}\right),$$

and because each closed ball on the right hand side is contained in the σ-algebra $\sigma\mathcal{C}_f$, so is the open ball $B_d(f,\epsilon)$.
Finally, since $(\mathcal{C}(E,F),d)$ is a separable metric space, there exists a countable base on $\mathcal{C}(E,F)$ consisting of open balls that generates the metric topology induced by d. This implies that any open set in $\mathcal{C}(E,F)$ is the countable union of open balls, so that any open set in $\mathcal{C}(E,F)$ is also contained in $\sigma\mathcal{C}_f$. Since the collection of all open sets in $\mathcal{C}(E,F)$ generates $\mathcal{B}_{\mathcal{C}}(E,F)$, we can see that $\mathcal{B}_{\mathcal{C}}(E,F) \subset \sigma\mathcal{C}_f$.
Q.E.D.

In the above proof, the separability of both (E,ρ) and $(\mathcal{C}(E,F),d)$ are needed, which is why the claim only holds when (E,ρ) is a compact metric space.

6.3.2 Characterizing Tightness on Continuous Function Spaces

Let (E,ρ) be a compact metric space, and $F=\mathbb{R}^n$. So far, we have seen that:

- $(\mathcal{C}(E,F),d)$ is a complete and separable metric space, where d is the supremum metric on $\mathcal{C}(E,F)$
- **(The Arzela-Ascoli Theorem)** A subset A of $\mathcal{C}(E,F)$ is relatively compact if and only if A is pointwise/uniformly bounded and uniformly equicontinuous on E
- The set $\mathcal{C}_f$ of all finite-dimensional sets in $\mathcal{C}(E,F)$ is a π-system generating $\mathcal{B}_{\mathcal{C}}(E,F)$, the Borel σ-algebra associated with d.

Since $(\mathcal{C}(E,F),\mathcal{B}_{\mathcal{C}}(E,F))$ is a measurable space, we can define probability measures on this space, and by extension talk of the tightness of a collection of probability measures on $\mathcal{C}(E,F)$. Since this involves compact sets in $\mathcal{C}(E,F)$, it is possible for us to make use of the Arzela-Ascoli theorem, which provides us a characterization of (relatively) compact sets in $\mathcal{C}(E,F)$. This is precisely what we do below:

Theorem 6.11 (Tightness on Continuous Function Spaces)
Let (E,ρ) be a compact metric space, $F=\mathbb{R}^n$, d the supremum metric on $\mathcal{C}(E,F)$, and $\mathcal{B}_{\mathcal{C}}(E,F)$ the associated Borel σ-algebra on $\mathcal{C}(E,F)$.
Suppose $\Pi=\{\mu_n\}_{n\in N_+}$ is a sequence of probability measures on $(\mathcal{C}(E,F),\mathcal{B}_{\mathcal{C}}(E,F))$. Then, Π is tight if and only if

i) For any $\eta>0$ and $x\in E$, there exists an $a>0$ such that

$$\mu_n\left(\{f \mid |f(x)|>a\}\right)\leq\eta$$

for any $n\in N_+$.

ii) For any $\epsilon>0$,

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mu_n\left(\{f \mid w(f,\delta)>\epsilon\}\right)=0.$$

Proof) Note initially that, for any $x\in E$ and $a>0$,

$$\{f \mid |f(x)|>a\}=\pi_x^{-1}\left(\overline{B}_{|\cdot|}(\mathbf{0},a)^c\right).$$

Since π_x is continuous on $\mathcal{C}(E,F)$ and the complement of the closed ball $\overline{B}_{|\cdot|}(\mathbf{0},a)$ is open, the inverse image on the right hand side is an open set in $\mathcal{C}(E,F)$. Therefore,

$$\mu_n\left(\{f \mid |f(x)|>a\}\right)$$

is well-defined for any $n \in N_+$.
Likewise, for any $\delta > 0$ and $\epsilon > 0$,

$$\{f \mid w(f,\delta) \geq \epsilon\} = w(\cdot,\delta)^{-1}((\epsilon,+\infty))$$

is an open set in $\mathcal{C}(E,F)$ because the mapping $f \mapsto w(f,\delta)$ is uniformly continous on $\mathcal{C}(E,F)$, so that

$$\mu_n(\{f \mid w(f,\delta) > \epsilon\})$$

is well-defined for any $n \in N_+$.

Necessity

Suppose that Π is a tight sequence of probability measures on $(\mathcal{C}(E,F), \mathcal{B}_{\mathcal{C}}(E,F))$. Then, for any $\eta > 0$, there exists a compact set K in $\mathcal{C}(E,F)$ such that

$$\mu_n(K^c) < \eta$$

for any $n \in N_+$. Because K is also a relatively compact set, by the Arzela-Ascoli theorem K is pointwise bounded and uniformly equicontinuous; formally,

$$\sup_{f \in K} |f(x)| < +\infty$$

for any $x \in E$, and

$$\lim_{h \to 0} \sup_{f \in K} w(f,h) = 0.$$

By the pointwise boundedness of functions in K, for any $x \in E$ there exists an $M \in (0,+\infty)$ such that $|f(x)| \leq M$ for any $f \in K$, so that

$$K \subset \{f \mid |f(x)| \leq M\},$$

and by implication,

$$\mu_n(\{f \mid |f(x)| > M\}) \leq \mu_n(K^c) < \eta$$

for any $n \in N_+$. This proves that i) holds.
Next, note that the uniform equicontinuity result implies that, for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\sup_{f \in K} w(f,h) \leq \epsilon$$

for any $0 < h < \delta$. Therefore, for any $0 < h < \delta$,

$$K \subset \{f \mid w(f,h) \leq \epsilon\},$$

so that

$$\mu_n(\{f \mid w(f,h) > \epsilon\}) \leq \mu_n(K^c) < \eta$$

for any $n \in N_+$. It follows that

$$\limsup_{n\to\infty} \mu_n(\{f \mid w(f,h) > \epsilon\}) \leq \eta$$

for any $0 < h < \delta$. Such a $\delta > 0$ exists for any $\eta > 0$, so

$$\lim_{h\to 0} \limsup_{n\to\infty} \mu_n(\{f \mid w(f,h) > \epsilon\}) = 0,$$

which shows that ii) holds.

Sufficiency

Now suppose conditions i) and ii) hold. By the compactness of E, there exists a countable subset E^0 of E that is dense in E. Let $\{x_k\}_{k\in N_+}$ be the arrangement of the elements of E^0 into a sequence.

For any $\eta > 0$ and $k \in N_+$, i) tells us that we can choose an $a_k > 0$ such that, for $A_k = \{f \mid |f(x_k)| > a_k\}$,

$$\mu_n(A_k) \leq \frac{\eta}{2^{k+1}}$$

for any $n \in N_+$. Define $B = \bigcup_k A_k$.

Furthermore, by ii), for any $k \in N_+$ there exists a $\delta_{k,0} > 0$ such that

$$\limsup_{n\to\infty} \mu_n(\{f \mid w(f,h) > 1/k\}) < \frac{\eta}{2^{k+1}}$$

for any $0 < h \leq \delta_{k,0}$, and therefore there exists an $N_k \in N_+$ such that

$$\mu_n(\{f \mid w(f,\delta_{k,0}) > 1/k\}) \leq \sup_{m\geq N_k} \mu_m(\{f \mid w(f,\delta_{k,0}) > 1/k\}) < \frac{\eta}{2^{k+1}}$$

for any $n \geq N_k$.

Because $\mu_1, \cdots, \mu_{N_k}$ are probability measures on $(\mathcal{C}(E,F), \mathcal{B}_{\mathcal{C}}(E,F))$ and $(\mathcal{C}(E,F), d)$ is a complete and separable metric space, the singletons $\{\mu_i\}$ for $1 \leq i \leq N_k$ are all tight.

It follows that, for any $1 \leq i \leq N_k$, there exists a compact set K_i in $\mathcal{C}(E,F)$ such that

$$\mu_i(K_i^c) < \frac{\eta}{2^{k+1}},$$

and because the finite union of compact sets is compact, defining $K_0 = \bigcup_{i=1}^{N_k} K_i$, K_0 is a compact set such that

$$\mu_i(K^c) < \frac{\eta}{2^{k+1}},$$

for any $1 \leq i \leq N_k$. Because K_0 is uniformly equicontinuous,

$$\lim_{h \to 0} \sup_{f \in K_0} w(f,h) = 0$$

and there exists a δ_0 such that $\sup_{f \in K_0} w(f,h) \leq \frac{1}{k}$ for any $0 < h \leq \delta_0$. As such,

$$K_0 \subset \{f \mid w(f,\delta_0) \leq 1/k\}$$

and

$$\mu_i\left(\{f \mid w(f,\delta_0) \leq 1/k\}\right) \leq \mu_i(K_0^c) < \frac{\eta}{2^{k+1}}$$

for any $1 \leq i \leq N_k$. Therefore, defining $\delta_k = \min(\delta_0, \delta_{k,0}) > 0$, we can see that

$$\mu_n\left(\{f \mid w(f,\delta_k) \leq 1/k\}\right) < \frac{\eta}{2^{k+1}}$$

for any $n \in N_+$. Define $B_k = \{f \mid w(f,\delta_k) \leq 1/k\}$.

Now define

$$A = B^c \cap \left(\bigcap_k B_k^c\right)$$

and $K = \overline{A}$. We will show that K is compact by showing that A is relatively compact. This will be done via the Arzela-Ascoli theorem.

– **Property 1: Uniform Equicontinuity**
For any $\epsilon > 0$, letting $k \in N_+$ be chosen so that $\frac{1}{k} < \epsilon$, there exists a $\delta_k > 0$ such that

$$w(f,h) < w(f,\delta_k) \leq \frac{1}{k} < \epsilon$$

for any $0 < h < \delta_k$ and $f \in A$, since $A \subset B_k^c$. By implication,

$$\sup_{f \in A} w(f,h) < \epsilon$$

for any $0 < h < \delta_k$, and because this holds for any $\epsilon > 0$,

$$\lim_{h \to 0} \sup_{f \in A} w(f, h) = 0,$$

or in other words, A is uniformly equicontinuous on $\mathcal{C}(E, F)$.

– **Property 2: Pointwise Boundedness**
For any $k \in N_+$ and $f \in A$, because $f \in A_k^c$ as well (A_k^c contains the countable intersection B^c), we have $|f(x_k)| \leq a_k$ by design.
The uniform equicontinuity of A on E tells us that there exists a $\zeta > 0$ such that

$$|f(z) - f(y)| < 1$$

for any $f \in A$ and $z, y \in E$ such that $\rho(z, y) < \zeta$.
Now choose any $x \in E$. By the denseness of E^0 in E there exists a $k \in N_+$ such that $\rho(x, x_k) < \zeta$; it follows that

$$|f(x)| \leq |f(x) - f(x_k)| + |f(x_k)| \leq a_k + 1.$$

for any $f \in A$. Therefore,

$$\sup_{f \in A} |f(x)| \leq a_k + 1 < +\infty,$$

and because this holds for any $x \in E$, A is a pointwise bounded collection of functions in $\mathcal{C}(E, F)$.

By the Arzela-Ascoli theorem, it follows that A is relatively compact, so that K is a compact set.
Finally, since

$$K^c \subset A^c = B \cup \left(\bigcup_k B_k \right) = \bigcup_k (A_k \cup B_k),$$

by countably subadditivity

$$\mu_n(K^c) \leq \sum_{k=1}^{\infty} (\mu_n(B_k) + \mu_n(A_k)) \leq \eta$$

for any $n \in N_+$. Our choice of $\eta > 0$ was arbitrary, so this shows that $\Pi = \{\mu_n\}_{n \in N_+}$ is a tight sequence of probability measures on $(\mathcal{C}(E, F), \mathcal{B}_{\mathcal{C}}(E, F))$.
Q.E.D.

6.4 Random Functions

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space, (E, ρ) a compact metric space and $F = \mathbb{R}^n$. A random (continuous) function $X : \Omega \to \mathcal{C}(E, F)$ is a random variable taking values in the measurable space $(\mathcal{C}(E, F), \mathcal{B}_{\mathcal{C}}(E, F))$, that is, a function that is measurable relative to $\mathcal{H}$ and $\mathcal{B}_{\mathcal{C}}(E, F)$.

Often times it is convenient to work with stochastic processes that correspond to some random function instead of the random function itself. Consider a random function X. Then, for any $t \in E$, define $X_t = \pi_t \circ X$; since π_t is a continuous function on $\mathcal{C}(E, F)$ taking values in F, it is measurable relative to $\mathcal{B}_{\mathcal{C}}(E, F)$ and $\mathcal{B}(F)$, and because measurability is preserved across compositions, X_t is a random variable taking values in $(F, \mathcal{B}(F))$.
As such, the stochastic process $\{X_t\}_{t \in E}$ is well-defined. Furthermore, for any $\omega \in \Omega$, the mapping $t \mapsto X_t(\omega)$ on E is precisely the continuous function $X(\omega)$ from E to F. Therefore, $\{X_t\}_{t \in E}$ is a stochastic process taking values in $(F, \mathcal{B}(F))$ with continuous paths.

Conversely, consider some stochastic process $\{X_t\}_{t \in E}$ taking values in $(F, \mathcal{B}(F))$ with continuous paths, and define $X : \Omega \to \mathcal{C}(E, F)$ so that, for any $\omega \in \Omega$, $X(\omega)$ is the function $t \mapsto X_t(\omega)$. To show that X is a random function, we use the fact that the collection $\mathcal{C}_f$ of all finite-dimensional sets in $\mathcal{C}(E, F)$ is a π-system generating $\mathcal{B}_{\mathcal{C}}(E, F)$.
For any $A \in \mathcal{C}_f$, there exist $t_1, \cdots, t_k \in E$ and $H \in \mathcal{B}(F^k)$ such that

$$A = \pi^{-1}_{t_1, \cdots, t_k}(H).$$

We can now see that

$$X^{-1}(A) = (\pi_{t_1, \cdots, t_k} \circ X)^{-1}(H) = (X_{t_1}, \cdots, X_{t_k})^{-1}(H) \in \mathcal{H},$$

where the last inclusion holds because $X_{t_1}, \cdots, X_{t_k}$ are random variables taking values in $(F, \mathcal{B}(F))$. Thus, the fact that $\mathcal{C}_f$ generates $\mathcal{B}_{\mathcal{C}}(E, F)$ implies that

$$X^{-1}(A) \in \mathcal{H}$$

for any $A \in \mathcal{B}_{\mathcal{C}}(E, F)$, and by definition X is a random function.

We call the stochastic process $\{X_t\}_{t \in E}$ taking values in $(F, \mathcal{B}(F))$ with continuous paths and the random function X taking values in $\mathcal{C}(E, F)$ constructed as above the process and function corresponding to one another.

6.4.1 Sufficienct Conditions for Weak Convergence

The equivalence of random functions and stochastic processes, as well as the characterization of tightness on continuous function spaces derived in the previous section, allow us to formulate the following sufficient conditions for the weak convergence of random functions:

Theorem 6.12 (Conditions for Weak Convergence)
Let (E,ρ) be a compact metric space, $F=\mathbb{R}^m$, d the supremum metric on $\mathcal{C}(E,F)$, and $\mathcal{B}_{\mathcal{C}}(E,F)$ the associated Borel σ-algebra on $\mathcal{C}(E,F)$.
Let $\{X^n\}_{n\in N_+}$ be a sequence of random functions and X a random function on $(\mathcal{C}(E,F),\mathcal{B}_{\mathcal{C}}(E,F))$. Letting $\{X_t^n\}_{t\in E}$ and $\{X_t\}_{t\in E}$ be the stochastic processes corresponding to X^n and X, if:

i) $(X_{t_1}^n,\cdots,X_{t_k}^n)\xrightarrow{d}(X_{t_1},\cdots,X_{t_k})$ for any $t_1,\cdots,t_k\in E$

ii) For any $\epsilon>0$,

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\left(w(X^n,\delta)>\epsilon\right)=0,$$

then $X^n\xrightarrow{d}X$.

Proof) Let $\Pi=\{\mu_n\}_{n\in N_+}$ be the distributions of $\{X^n\}_{n\in N_+}$, and μ that of X. The first condition tells us that, for any $t_1,\cdots,t_k\in E$,

$$\pi_{t_1,\cdots,t_k}\circ X^n\xrightarrow{d}\pi_{t_1,\cdots,t_k}\circ X,$$

or equivalently,

$$\mu_n\circ\pi_{t_1,\cdots,t_k}^{-1}\to\mu\circ\pi_{t_1,\cdots,t_k}^{-1}$$

weakly. Suppose that $\{\mu_{n_k}\}_{k\in N_+}$ is a weakly convergent subsequence of $\{\mu_n\}_{n\in N_+}$. Letting the weak limit be the probability measure v on $\mathcal{C}(E,F)$, by the continuous mapping theorem

$$\mu_{n_k}\circ\pi_{t_1,\cdots,t_k}^{-1}\to v\circ\pi_{t_1,\cdots,t_k}^{-1}$$

weeakly as $k\to\infty$, so that, by the uniqueness of weak limits,

$$\mu\circ\pi_{t_1,\cdots,t_k}^{-1}=v\circ\pi_{t_1,\cdots,t_k}^{-1}.$$

This holds for any $t_1,\cdots,t_k\in E$, so $\mu(H)=v(H)$ for any finite-dimensional set $H\in\mathcal{C}_f$. Since $\mathcal{C}_f$ is a π-system generating $\mathcal{B}_{\mathcal{C}}(E,F)$ and μ,v are probability measures, it follows that $\mu=v$ on $\mathcal{B}_{\mathcal{C}}(E,F)$. Therefore, every weakly convergent subsequence of $\{\mu_n\}_{n\in N_+}$ has the weak limit μ.

In light of lemma 4.10, we need only prove that Π is relatively compact to show that $\mu_n \to \mu$ weakly. Furthermore, the sufficiency part of Prohorov's theorem implies that this can be shown by simply proving that Π is tight. This is in turn shown using the characterization of tightness on continuous function spaces.
By i), for any $t \in E$

$$\mu_n \circ \pi_t^{-1} \to \mu \circ \pi_t^{-1}$$

weakly. As such, the sequence $\{\mu_n \circ \pi_t^{-1}\}_{n \in N_+}$ of probability measures on $(F, \mathcal{B}(F))$ is trivially relatively compact, so that the necessity part of Prohorov's theorem implies that it is tight. By definition, for any $\eta > 0$, there exists a compact set K in F such that

$$\mu_n \left(\pi_t^{-1}(K^c) \right) < \eta$$

for any $n \in N_+$. K, being bounded, is contained in some closed ball around $\mathbf{0}$ with radius $a > 0$, and as such

$$\mu_n \left(\{ f \mid |f(t)| > a \} \right) \leq \mu_n \left(\pi_t^{-1}(K^c) \right) < \eta$$

for any $n \in N_+$.
Furthermore, ii) tells us that

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mu_n \left(\{ f \mid w(f, \delta) > \epsilon \} \right) = 0$$

for any $\epsilon > 0$, so Π is a tight collection of probability measures on $\mathcal{C}(E, F)$ (by theorem 6.11).
Q.E.D.

6.4.2 Construction of the Wiener Measure and Donsker's Theorem

Here we construct the m-dimensional Wiener measure, and by extension the Wiener process, on $[0,1]$. The m-dimensional Wiener process on more general closed intervals can be constructed by stitching together the processes defined on $[0,1]$ (technically, we stitch together Brownian bridges instead of the Wiener processes themselves). The Wiener measure in question is derived as the weak limit of a sequence of random functions defined on the compact space $[0,1]$ equipped with the euclidean metric.

We start by constructing a sequence of stochastic processes with index set $[0,1]$ and a sequence of corresponding random functions in $\mathcal{C}([0,1],\mathbb{R}^m)$. The Wiener measure will be found as the limit of this sequence.

Let $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ be a sequence of independent and identically distributed m-dimensional random vectors with mean 0, variance I_m and finite fourth moments. For any $T \in N_+$, define the stochastic process $\{X_T(r)\}_{r\in[0,1]}$ as

$$\begin{aligned} X_T(r) &= \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \varepsilon_t + \frac{1}{\sqrt{T}} \left(Tr - \lfloor Tr \rfloor\right) \varepsilon_{\lfloor Tr \rfloor + 1} \\ &= \frac{1}{\sqrt{T}} S_{\lfloor Tr \rfloor} + \frac{1}{\sqrt{T}} \left(Tr - \lfloor Tr \rfloor\right) \varepsilon_{\lfloor Tr \rfloor + 1} \end{aligned}$$

for any $r \in [0,1]$, where $\{S_t\}_{t\in\mathbb{N}}$ is the partial sum process of $\{\varepsilon_t\}_{t\in N_+}$ with $S_0 = \mathbf{0}$. Being the linear combination of random vectors, each $X_T(r)$ is a random vector, and $\{X_T(r)\}_{r\in[0,1]}$ clearly has continuous paths. Therefore, there exists a random function X^T taking values in $\mathcal{C}([0,1],\mathbb{R}^m)$ that corresponds to $\{X_T(r)\}_{r\in[0,1]}$. Let μ_T be the distribution of X^T; it is a probability measure on $(\mathcal{C}([0,1],\mathbb{R}^m), \mathcal{B}_{\mathcal{C}}([0,1],\mathbb{R}^m))$.

We start with a characterization of tightness of the sequence $\{\mu_T\}_{T\in N_+}$. We once again make use of theorem 4.11, which furnishes necessary and sufficient conditions for a sequence of probability measures on a continuous function space to be tight.

Theorem 6.13 (Conditions for Tightness of $\{\mu_T\}$)
Let $\{\mu_T\}_{T\in N_+}$ be the sequence of probability measures on $(\mathcal{C}([0,1],\mathbb{R}^m), \mathcal{B}_{\mathcal{C}}([0,1],\mathbb{R}^m))$ defined above. Then, $\Pi = \{\mu_T\}_{T\in N_+}$ is tight if

$$\lim_{\lambda\to\infty} \limsup_{n\to\infty} \lambda^2 \mathbb{P}\left(\max_{k\le n} |S_k| > \lambda\sqrt{n}\right) = 0.$$

Proof) Since $[0,1]$ is compact when equipped with the euclidean metric on $\mathbb{R}$, by theorem 6.11 $\Pi = \{\mu_T\}_{T\in N_+}$ is a tight sequence of probability measures if and only if:

- For any $\eta > 0$ and $r \in [0,1]$, there exists an $a > 0$ such that

$$\mu_T(\{f \mid |f(r)| > a\}) \le \eta$$

for any $T \in N_+$.

– For any $\epsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{T \to \infty} \mu_T \left(\{ f \mid w(f, \delta) > \epsilon \} \right) = 0.$$

Condition 1

We will first show that the pointwise boundedness condition holds. For any $T \in N_+$ and $r \in [0, 1]$,

$$\mu_T \circ \pi_r^{-1}$$

is the distribution of $\pi_r \circ X_T$, which is given as

$$\pi_r \circ X_T = \frac{1}{\sqrt{T}} S_{\lfloor Tr \rfloor} + \frac{1}{\sqrt{T}} (Tr - \lfloor Tr \rfloor) \varepsilon_{\lfloor Tr \rfloor + 1}.$$

Since

$$\left| \frac{1}{\sqrt{T}} (Tr - \lfloor Tr \rfloor) \varepsilon_{\lfloor Tr \rfloor + 1} \right| \leq \frac{1}{\sqrt{T}} \left| \varepsilon_{\lfloor Tr \rfloor + 1} \right|,$$

we can see that, for any $\delta > 0$,

$$\mathbb{P} \left(\left| \frac{1}{\sqrt{T}} (Tr - \lfloor Tr \rfloor) \varepsilon_{\lfloor Tr \rfloor + 1} \right| > \delta \right) \leq \frac{1}{\delta^2 \cdot T} \operatorname{tr} \left(\mathbb{E} \left[\varepsilon_{\lfloor Tr \rfloor + 1} \varepsilon'_{\lfloor Tr \rfloor + 1} \right] \right) = \frac{m}{\delta^2 \cdot T}$$

by Chebyshev's inequality; taking $T \to \infty$ on both sides yields

$$\lim_{T \to \infty} \mathbb{P} \left(\left| \frac{1}{\sqrt{T}} (Tr - \lfloor Tr \rfloor) \varepsilon_{\lfloor Tr \rfloor + 1} \right| > \delta \right) = 0.$$

Therefore,

$$\frac{1}{\sqrt{T}} (Tr - \lfloor Tr \rfloor) \varepsilon_{\lfloor Tr \rfloor + 1} \xrightarrow{p} \mathbf{0}$$

as $T \to \infty$.

On the other hand, note that

$$\frac{1}{\sqrt{T}} S_{\lfloor Tr \rfloor} = \frac{\sqrt{\lfloor Tr \rfloor}}{\sqrt{T}} \cdot \left(\frac{1}{\sqrt{\lfloor Tr \rfloor}} \sum_{t=1}^{\lfloor Tr \rfloor} \varepsilon_t \right).$$

By the Lindeberg-Levy CLT and Slutsky's theorem, we now have

$$\pi_r \circ X_T \xrightarrow{d} r \cdot Z,$$

where $Z \sim N(\mathbf{0}, I_m)$. This implies that the sequence

$$\{\mu_T \circ \pi_r^{-1}\}_{T \in N_+}$$

converges weakly, and the proof of theorem 6.12 shows us that this implies, for any $\eta > 0$, the existence of an $a > 0$ such that

$$\mu_T\left(\{f \mid |f(r)| > a\}\right) \leq \eta$$

for any $T \in N_+$.

Condition 2

Now we show that our assumption implies that, for any $\epsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{T \to \infty} \mu_T\left(\{f \mid w(f, \delta) > \epsilon\}\right) = 0.$$

Choose any $\epsilon > 0$ and $\eta > 0$. By assumption,

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} \lambda^2 \mathbb{P}\left(\max_{k \leq n} |S_k| > \lambda \sqrt{n}\right) = 0,$$

so there exists an $M > 0$ such that

$$\inf_{d \in N_+} \left(\sup_{n \geq d} \lambda^2 \mathbb{P}\left(\max_{k \leq n} |S_k| > \lambda \sqrt{n}\right)\right) < \eta,$$

for any $\lambda > M$, and in turn, for any $\lambda > M$, by the definition of the infimum there exists a $d_\lambda \in N_+$ such that

$$\sup_{n \geq d_\lambda} \lambda^2 \mathbb{P}\left(\max_{k \leq n} |S_k| > \lambda \sqrt{n}\right) < \frac{\eta \epsilon^2}{288}.$$

Letting $\lambda(\delta) = \frac{\epsilon}{6 \cdot \sqrt{2\delta}}$, we choose $\delta \in (0, 1)$ so that

$$\lambda(\delta) > M.$$

Now we proceed in steps:

Step 1: Bounding the Measure

For any $T \in N_+$, define

$$m(T, \delta) = \lceil \delta T \rceil, \quad v(T, \delta) = \lceil T / m(T, \delta) \rceil,$$

and let

$$t_{iT} = \begin{cases} \frac{i \cdot m(T,\delta)}{T} & \text{if } 0 \le i \le v(T,\delta) - 1 \\ 1 & \text{if } i = v(T,\delta). \end{cases}$$

$m(T,\delta)$ is an integer smaller than or equal to T, $v(T,\delta)$ is a natural number, and $\{t_{0T}, \cdots, t_{v(T,\delta)T}\}$ is a partition of $[0,1]$, where

$$t_{v(T,\delta)-1,T} = \frac{(v(T,\delta)-1) \cdot m(T,\delta)}{T} < 1 = t_{v(T,\delta)T}.$$

For any $T \in N_+$, choose any $f \in \mathcal{C}([0,1], \mathbb{R}^m)$ such that $w(f,\delta) > \epsilon$. Then, there exists some $x, y \in [0,1]$ such that $|x-y| < \delta$ and

$$w(f,\delta) \ge |f(x) - f(y)| > \epsilon.$$

Because the length of the intervals in the partition $\{t_{0T}, \cdots, t_{v(T,\delta)T}\}$ is greater than or equal to δ, x and y are either contained in the same interval or contained in adjacent intervals. Consider two cases:

- **x and y are in the same interval**
  Assuming that $t_{i-1,T} \le x, y \le t_{iT}$ for some $1 \le i \le v(T,\delta)$,

  $$|f(x) - f(y)| \le |f(t_{i-1,T}) - f(x)| + |f(y) - f(t_{i-1,T})| \le 2 \cdot \sup_{t_{i-1,T} \le z \le t_{iT}} |f(t_{i-1,T}) - f(z)|.$$

- **x and y are contained in different intervals**
  Assuming that $t_{i-1,T} \le x \le t_{iT}$ and $t_{iT} \le y \le t_{i+1,T}$ for some $1 \le i \le v(T,\delta) - 1$, we have

  $$\begin{aligned} |f(x) - f(y)| &\le |f(t_{i-1,T}) - f(x)| + |f(t_{i-1,T}) - f(y)| \\ &\le |f(t_{i-1T}) - f(x)| + |f(t_{i,T}) - f(y)| + |f(t_{i,T}) - f(t_{i-1,T})| \\ &\le 2 \cdot \sup_{t_{i-1,T} \le z \le t_{iT}} |f(t_{i-1,T}) - f(z)| + \sup_{t_{i,T} \le z \le t_{i+1,T}} |f(t_{i,T}) - f(z)|. \end{aligned}$$

Therefore, it must be the case that

$$\max_{1 \le i \le v(T,\delta)} \sup_{t_{i-1,T} \le z \le t_{iT}} |f(t_{i-1,T}) - f(z)| > \frac{\epsilon}{3}$$

for $|f(x) - f(y)| > \epsilon$ to hold. In other words,

$$\{f \mid w(f,\delta) > \epsilon\} \subset \Big\{f \mid \max_{1 \le i \le v(T,\delta)} \sup_{t_{i-1,T} \le z \le t_{iT}} |f(t_{i-1,T}) - f(z)| > \frac{\epsilon}{3}\Big\},$$

and it follows that

$$\mu_T\left(\{f \mid w(f,\delta) > \epsilon\}\right) \leq \mu_T\left(\left\{f \mid \max_{1\leq i\leq v(T,\delta)} \sup_{t_{i-1,T}\leq z\leq t_{iT}} |f(t_{i-1,T}) - f(z)| > \frac{\epsilon}{3}\right\}\right)$$
$$\leq \sum_{i=1}^{v(T,\delta)} \mathbb{P}\left(\sup_{t_{i-1,T}\leq z\leq t_{iT}} |X_T(t_{i-1,T}) - X_T(z)| > \frac{\epsilon}{3}\right).$$

Step 2: Using the iid Condition

Choose any $1 \leq i \leq v(T,\delta)$. Then, since $t_{i-1,T} = \frac{i\cdot m(T,\delta)}{T}$, for any $t_{i-1,T} \leq z \leq t_{iT}$ we have

$$i \cdot m(T,\delta) \leq Tz \leq \min\left((i+1)\cdot m(T,\delta), T\right).$$

and therefore

$$\mathbb{P}\left(\sup_{t_{i-1,T}\leq z\leq t_{iT}} |X_T(t_{i-1,T}) - X_T(z)| > \frac{\epsilon}{3}\right) \leq \mathbb{P}\left(\sup_{i\cdot m(T,\delta)\leq k\leq (i+1)\cdot m(T,\delta)} \frac{\left|S_k - S_{i\cdot m(T,\delta)}\right|}{\sqrt{T}} > \frac{\epsilon}{6}\right)$$
$$+ \mathbb{P}\left(\max_{i\cdot m(T,\delta)\leq k\leq (i+1)\cdot m(T,\delta)} |\varepsilon_{k+1}| > \frac{\epsilon\sqrt{T}}{6}\right)$$
$$= \mathbb{P}\left(\sup_{k\leq m(T,\delta)} |S_k| > \frac{\epsilon\sqrt{T}}{6}\right) + \mathbb{P}\left(\max_{1\leq i\leq m(T,\delta)} |\varepsilon_i| > \frac{\epsilon\sqrt{T}}{6}\right),$$

where the last equality follows because $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is i.i.d. Thus,

$$\mu_T\left(\{f \mid w(f,\delta) > \epsilon\}\right) \leq v(T,\delta)\cdot\mathbb{P}\left(\sup_{k\leq m(T,\delta)} |S_k| > \frac{\epsilon\sqrt{T}}{6}\right) + v(T,\delta)\cdot\mathbb{P}\left(\max_{1\leq i\leq m(T,\delta)} |\varepsilon_i| > \frac{\epsilon\sqrt{T}}{6}\right).$$

Step 3: Taking $T \to \infty$

By construction,

$$\delta T - 1 < m(T,\delta) \leq \delta T;$$

thus,

$$\frac{1}{\delta} \leq \frac{T}{m(T,\delta)} < \frac{T}{m(T,\delta)+1} < \frac{1}{\delta},$$

and sending $T \to \infty$

$$\frac{T}{m(T,\delta)} \to \frac{1}{\delta}.$$

Therefore, there exists an $N \in N_+$, dependent on δ, such that

$$\left| \frac{T}{m(T,\delta)} - \frac{1}{\delta} \right| < \frac{1}{2\delta} < \frac{1}{\delta}$$

for any $T \geq N$, so that, for any $T \geq N$,

$$\frac{1}{2\delta} < \frac{T}{m(T,\delta)} < \frac{2}{\delta}.$$

Likewise, because

$$\frac{T}{m(T,\delta)} - 1 < v(T,\delta) \leq \frac{T}{m(T,\delta)},$$

for any $T \geq N$ we have

$$v(T,\delta) \leq \frac{T}{m(T,\delta)} < \frac{2}{\delta}.$$

Therefore, for any $T \geq N$, because

$$T > \frac{m(T,\delta)}{2\delta},$$

we have

$$\mu_T\left(\{f \mid w(f,\delta) > \epsilon\}\right) \leq \frac{2}{\delta} \cdot \underbrace{\mathbb{P}\left(\sup_{k \leq m(T,\delta)} |S_k| > \frac{\epsilon}{6} \cdot \sqrt{\frac{m(T,\delta)}{2\delta}} \right)}_{I} + \frac{2}{\delta} \cdot \underbrace{\mathbb{P}\left(\max_{1 \leq i \leq m(T,\delta)} |\varepsilon_i| > \frac{\epsilon\sqrt{T}}{6} \right)}_{II}.$$

We study each term in turn:

i) **Term I**

Under the definition of $\lambda(\delta)$, term I can be written as

$$I = \frac{144}{\epsilon^2} \cdot \lambda(\delta)^2 \mathbb{P}\left(\sup_{k \leq m(T,\delta)} |S_k| > \lambda(\delta) \cdot \sqrt{m(T,\delta)} \right).$$

$m(T,\delta) \to \infty$ as $T \to \infty$, so there exists an $N_1 \in N_+$ such that $N_1 \geq N$ and, for any $T \geq N_1$,

$$m(T,\delta) > d_{\lambda(\delta)}.$$

Then, for any $T \geq N_1$,

$$\frac{144}{\epsilon^2}\cdot\lambda(\delta)^2\mathbb{P}\left(\sup_{k\le m(T,\delta)}|S_k|>\lambda(\delta)\cdot\sqrt{m(T,\delta)}\right)$$
$$\le\frac{24}{\epsilon}\cdot\left(\sup_{n\ge d_{\lambda(\delta)}}\lambda(\delta)^2\mathbb{P}\left(\sup_{k\le n}|S_k|>\lambda(\delta)\sqrt{n}\right)\right)<\frac{\eta}{2}.$$

ii) **Term II**

Since

$$\left\{max_{1\le i\le m(T,\delta)}|\varepsilon_i|>\frac{\epsilon\sqrt{T}}{6}\right\}\subset\bigcup_{i=1}^{m(T,\delta)}\left\{|\varepsilon_i|>\frac{\epsilon\sqrt{T}}{6}\right\},$$

we have

$$\mathbb{P}\left(\max_{1\le i\le m(T,\delta)}|\varepsilon_i|>\frac{\epsilon\sqrt{T}}{6}\right)\le\sum_{i=1}^{m(T,\delta)}\mathbb{P}\left(|\varepsilon_i|>\frac{\epsilon\sqrt{T}}{6}\right)$$
$$\le\frac{6^4\cdot m(T,\delta)}{\epsilon^4\cdot T^2}\cdot\mathbb{E}|\varepsilon_1|^4.$$

where we used the fact that $\varepsilon_1,\varepsilon_2,\cdots$ are iid. The preceding result tells us that $\frac{m(T,\delta)}{T}\to\delta$ as $T\to\infty$, and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ has finite fourth moments, so

$$\lim_{T\to\infty}\mathbb{P}\left(\max_{1\le i\le m(T,\delta)}|\varepsilon_i|>\frac{\epsilon\sqrt{T}}{6}\right)=0$$

Thus, there exists an $N_2\in N_+$ such that $N_2\ge N_1$ and, for any $T\ge N_2$, we have

$$\mathbb{P}\left(\max_{1\le i\le m(T,\delta)}|\varepsilon_i|>\frac{\epsilon\sqrt{T}}{6}\right)<\frac{\eta\delta}{4}.$$

Putting the results together, we can see that, for any $T\ge N_2$, where this N_2 depends only on δ,

$$\mu_T\left(\{f\mid w(f,\delta)>\epsilon\}\right)<\eta.$$

Therefore,

$$\limsup_{T\to\infty}\mu_T\left(\{f\mid w(f,\delta)>\epsilon\}\right)\le\eta.$$

Because the modulus of continuity w is decreasing in its second argument, for any $h\in(0,\delta)$, we also have

$$\limsup_{T\to\infty}\mu_T\left(\{f\mid w(f,h)>\epsilon\}\right)\le\eta.$$

Our choice of $\eta > 0$ was arbitrary, so this implies that

$$\lim_{h\to 0} \limsup_{T\to\infty} \mu_T\left(\{f \mid w(f,h) > \epsilon\}\right) = 0,$$

which is exactly what we wanted to show.

Q.E.D.

A series of lemmas show us that the condition introduced in the theorem above is satisfied.

Lemma 6.14 (Etemadi's Inequality)
Let $\{e_t\}_{t\in N_+}$ be an independent sequence of m-dimensional random vectors and $\{V_T\}_{T\in N_+}$ their partial sum process. Then, for any $n \in N_+$ and $\alpha > 0$,

$$\mathbb{P}\left(\max_{k\le n} |V_k| > \alpha\right) \le 3 \cdot \max_{k\le n} \mathbb{P}\left(|V_k| > \frac{\alpha}{3}\right).$$

Proof) For any $n \in N_+$ and $1 \le k \le n$, define

$$B_k = \{|V_k| > \alpha\} \cap \left(\bigcap_{j=1}^{k-1} \{|V_j| \le \alpha\}\right).$$

Then, the collection $\{B_1, \cdots, B_n\}$ of $\mathcal{H}$-measurable sets are disjoint and have the union

$$\bigcup_{k=1}^{n} B_k = \left\{\max_{k\le n} |V_k| > \alpha\right\}.$$

Therefore, by countable additivity,

$$\begin{aligned}
\mathbb{P}\left(\max_{k\le n} |V_k| > \alpha\right) &= \sum_{k=1}^{n} \mathbb{P}(B_k) \\
&= \sum_{k=1}^{n} \mathbb{P}\left(\left\{|V_n| > \frac{\alpha}{3}\right\} \cap B_k\right) + \sum_{k=1}^{n} \mathbb{P}\left(\left\{|V_n| \le \frac{\alpha}{3}\right\} \cap B_k\right) \\
&\le \mathbb{P}\left(|V_n| > \frac{\alpha}{3}\right) + \sum_{k=1}^{n} \mathbb{P}\left(\left\{|V_n| \le \frac{\alpha}{3}\right\} \cap B_k\right).
\end{aligned}$$

For each $1 \le k \le n$, since $|V_k| > \alpha$ and $|V_n| \le \frac{\alpha}{3}$, we have

$$|V_n - V_k| \ge |V_k| - |V_n| > \frac{2\alpha}{3},$$

and $V_n - V_k$ is independent of $V_1, \cdots, V_k$ and by implication B_k since $V_n - V_k$ is the sum

of e_t for $k+1 \leq t \leq n$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(\max_{k\leq n}|V_k| > \alpha\right) &\leq \mathbb{P}\left(|V_n| > \frac{\alpha}{3}\right) + \sum_{k=1}^{n}\mathbb{P}\left(\left\{|V_n - V_k| > \frac{2\alpha}{3}\right\} \cap B_k\right) \\
&= \mathbb{P}\left(|V_n| > \frac{\alpha}{3}\right) + \sum_{k=1}^{n}\mathbb{P}\left(|V_n - V_k| \leq \frac{2\alpha}{3}\right) \cdot \mathbb{P}(B_k) \\
&\leq \mathbb{P}\left(|V_n| > \frac{\alpha}{3}\right) + \left(\max_{k\leq n}\mathbb{P}\left(|V_n - V_k| > \frac{2\alpha}{3}\right)\right)\left(\sum_{k=1}^{n}\mathbb{P}(B_k)\right).
\end{aligned}$$

By countably additivity again,

$$\sum_{k=1}^{n}\mathbb{P}(B_k) = \mathbb{P}\left(\bigcup_{k=1}^{n} B_k\right) \leq 1$$

and $|V_n - V_k| > \frac{2\alpha}{3}$ implies either $|V_n| > \frac{\alpha}{3}$ or $|V_k| > \frac{\alpha}{3}$, we can see that

$$\begin{aligned}
\mathbb{P}\left(\max_{k\leq n}|V_k| > \alpha\right) &\leq 2 \cdot \mathbb{P}\left(|V_n| > \frac{\alpha}{3}\right) + \left(\max_{k\leq n}\mathbb{P}\left(|V_k| > \frac{\alpha}{3}\right)\right) \\
&\leq 3 \cdot \max_{k\leq n}\mathbb{P}\left(|V_k| > \frac{\alpha}{3}\right).
\end{aligned}$$

Q.E.D.

Lemma 6.15 Let $\{\mu_T\}_{T\in N_+}$ be the sequence of probability measures defined above. $\{\mu_T\}_{T\in N_+}$ is a tight sequence of probability measures.

Proof) By Etemadi's inequality and the independence of $\{\varepsilon_t\}_{t\in\mathbb{Z}}$, we can see that

$$\lambda^2\mathbb{P}\left(\max_{k\leq n}|S_k| > \lambda\sqrt{n}\right) \leq 3\lambda^2 \cdot \left(\max_{k\leq n}\mathbb{P}\left(|S_k| > \frac{\lambda}{3}\sqrt{n}\right)\right)$$

for any $\lambda > 0$ and $n \in N_+$. For any $1 \leq k \leq n$,

$$\begin{aligned}
\mathbb{P}\left(|S_k| > \frac{\lambda}{3}\sqrt{n}\right) &\leq \mathbb{P}\left(|S_k| > \frac{\lambda}{3}\sqrt{k}\right) \\
&\leq \frac{81}{\lambda^4 k^2}\mathbb{E}|S_k|^4 \\
&= \frac{81}{\lambda^4 k^2}\left(\sum_{t=1}^{k}\mathbb{E}|\varepsilon_t|^4 + \sum_{t=1}^{k}\sum_{s\neq t}^{k}\mathbb{E}\left[\varepsilon_t'\varepsilon_t\varepsilon_s'\varepsilon_s\right] + \sum_{t=1}^{k}\sum_{s\neq t}^{k}\mathbb{E}\left[\varepsilon_t'\varepsilon_s\varepsilon_t'\varepsilon_s\right] + \sum_{t=1}^{k}\sum_{s\neq t}^{k}\mathbb{E}\left[\varepsilon_t'\varepsilon_s\varepsilon_s'\varepsilon_t\right]\right) \\
&= \frac{81}{\lambda^4 k^2}\left(k \cdot \mathbb{E}|\varepsilon_1|^4 + k(k-1)(m^2+2m)\right) \\
&= \frac{81}{\lambda^4 k} \cdot \mathbb{E}|\varepsilon_1|^4 + \frac{81}{\lambda^4} \cdot \frac{k(k-1)(m^2+2m)}{k^2} \\
&\leq \frac{81}{\lambda^4} \cdot \mathbb{E}|\varepsilon_1|^4 + \frac{81}{\lambda^4} \cdot (m^2+2m).
\end{aligned}$$

Therefore,

$$\max_{k\leq n}\mathbb{P}\left(|S_k|>\frac{\lambda}{3}\sqrt{n}\right)\leq\frac{81}{\lambda^4}\cdot\mathbb{E}|\varepsilon_1|^4+\frac{81}{\lambda^4}\cdot(m^2+2m),$$

and by the finiteness of the fourth moments of $\{\varepsilon_t\}_{t\in\mathbb{Z}}$, we have

$$\lim_{\lambda\to\infty}\limsup_{k\to\infty}3\lambda^2\cdot\left(\max_{k\leq n}\mathbb{P}\left(|S_k|>\frac{\lambda}{3}\sqrt{n}\right)\right)=0.$$

By implication,

$$\lim_{\lambda\to\infty}\limsup_{n\to\infty}\lambda^2\mathbb{P}\left(\max_{k\leq n}|S_k|>\lambda\sqrt{n}\right)=0$$

as well, and by theorem 6.13, $\{\mu_T\}_{T\in N_+}$ is tight.
Q.E.D.

Because the metric space $(\mathcal{C}([0,1],\mathbb{R}^m),d)$, where d is the supremum metric on $\mathcal{C}([0,1],\mathbb{R}^m)$, is complete, and the sequence $\{\mu_T\}_{T\in N_+}$ is tight, by the sufficiency part of Prohorov's theorem $\{\mu_T\}_{T\in N_+}$ is relatively compact. By implication, there exists a subsequence of $\{\mu_T\}_{T\in N_+}$ that converges weakly to some probability measure μ^W on $(\mathcal{C}([0,1],\mathbb{R}^m),\mathcal{B}_{\mathcal{C}}([0,1],\mathbb{R}^m))$.

This measure μ^W is called the standard m-dimensional Wiener measure on $[0,1]$, and the random function W with distribution μ^W is referred to as the standard m-dimensional Wiener function on $[0,1]$. Similarly, the m-dimensional stochastic process $\{W(r)\}_{r\in[0,1]}$ with continuous paths corresponding to W is the standard m-dimensional Wiener process on $[0,1]$.

6.4.3 Donsker's Theorem

Before studying some of the properties of the Wiener measure, we first derive the analogue of the Lindeberg-Levy CLT for sequences of random continuous functions. Specifically, we prove that $\{\mu_T\}_{T\in N_+}$ converges weakly to μ^W, or equivalently, that $\{X^T\}_{T\in N_+}$ converges in distribution to W. This result is called Donsker's Theorem, or the Functional Central Limit Theorem (FCLT).

Theorem 6.16 (Donsker's Theorem)
Let $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ be a sequence of independent and identically distributed m-dimensional random vectors with mean 0, variance I_m and finite fourth moments. For any $T \in N_+$, define the stochastic process $\{X_T(r)\}_{r\in[0,1]}$ as

$$X_T(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \varepsilon_t + \frac{1}{\sqrt{T}} (Tr - \lfloor Tr \rfloor)\, \varepsilon_{\lfloor Tr \rfloor + 1}$$

for any $r \in [0,1]$. Let X^T be the random function taking values in $\mathcal{C}([0,1],\mathbb{R}^m)$ that corresponds to $\{X_T(r)\}_{r\in[0,1]}$, and μ_T the distribution of X^T.
Then, the sequence $\{\mu_T\}_{T\in N_+}$ of probability measures on $\mathcal{C}([0,1],\mathbb{R}^m)$ converges weakly to the Wiener measure μ^W.

Proof) We have already shown above that $\{\mu_T\}_{T\in N_+}$ is a tight sequence of probability measures. By the sufficiency part of Prohorov's theorem and the completeness of the metric space $(\mathcal{C}([0,1],\mathbb{R}^m), d)$, where d is the supremum metric, $\{\mu_T\}_{T\in N_+}$ is relatively compact. In light of lemma 4.10, we need only show that every weakly convergent subsequence of $\{\mu_T\}_{T\in N_+}$ converges to μ^W.

We will show that any weak limit of a convergent subsequence of $\{\mu_T\}_{T\in N_+}$ has the same set of finite-dimensional distributions. Let v be a probability measure on $\mathcal{C}([0,1],\mathbb{R}^n)$ such that there exists a weakly convergent subsequence $\{\mu_{T_k}\}_{k\in N_+}$ of $\{\mu_T\}_{T\in N_+}$ with weak limit equal to v.

Choose any $0 = t_0 < \cdots < t_k \leq 1$. Then, for large enough T, $Tt_0, \cdots, Tt_k$ are all integers, so that

$$X_T(t_i) = \frac{1}{\sqrt{T}} S_{Tt_i}$$

for $0 \leq i \leq k$. By implication, for large T,

$$\begin{aligned} X_T(t_i) - X_T(t_{i-1}) &= \frac{1}{\sqrt{T}} \sum_{t=Tt_{i-1}+1}^{Tt_i} \varepsilon_t \\ &= \sqrt{t_i - t_{i-1}} \left(\frac{1}{\sqrt{T(t_i - t_{i-1})}} \sum_{t=Tt_{i-1}+1}^{Tt_i} \varepsilon_t \right), \end{aligned}$$

and because $T(t_i - t_{i-1}) \to +\infty$ as $T \to \infty$, by the Lindeberg-Levy CLT

$$X_T(t_i) - X_T(t_{i-1}) \xrightarrow{d} N\left[\mathbf{0}, (t_i - t_{i-1}) \cdot I_m\right].$$

Furthermore, since $X_T(t_i) - X_T(t_{i-1})$ involve non-overlapping terms in $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ across $1 \le i \le k$ for any T that is large enough, we can see that these increments are independent. Defining

$$I_T = \begin{pmatrix} X_T(t_1) - X_T(t_0) \\ \vdots \\ X_T(t_k) - X_T(t_{k-1}) \end{pmatrix},$$

the characteristic function for I_T at any $r = (r_1, \cdots, r_k) \in \mathbb{R}^{mk}$ is therefore

$$\mathbb{E}\left[\exp(ir' I_T)\right] = \prod_{j=1}^{k} \mathbb{E}\left[\exp\Big(ir_j'(X_T(t_i) - X_T(t_{i-1}))\Big)\right].$$

By the continuity theorem,

$$\lim_{T\to\infty} \mathbb{E}\left[\exp\Big(ir_j'(X_T(t_i) - X_T(t_{i-1}))\Big)\right] = \exp\Big(-\frac{1}{2}(t_i - t_{i-1}) r_j' r_j\Big)$$

for $1 \le i \le k$, so we have

$$\begin{aligned} \lim_{T\to\infty} \mathbb{E}\left[\exp(ir' I_T)\right] &= \prod_{j=1}^{k} \exp\Big(-\frac{1}{2}(t_i - t_{i-1}) r_j' r_j\Big) \\ &= \exp\Big(-\frac{1}{2} r' V r\Big), \end{aligned}$$

where

$$V = \begin{pmatrix} (t_1 - t_0) \cdot I_m & \cdots & O \\ \vdots & \ddots & \vdots \\ O & \cdots & (t_k - t_{k-1}) \cdot I_m \end{pmatrix}.$$

By the continuity theorem again,

$$I_T \xrightarrow{d} N\left[\mathbf{0}, V\right].$$

We can easily find that

$$\pi_{t_1,\cdots,t_k} \circ X^T = \begin{pmatrix} X_T(t_1) \\ \vdots \\ X_T(t_k) \end{pmatrix} = \underbrace{\begin{pmatrix} I_m & \cdots & O \\ \vdots & \ddots & \vdots \\ I_m & \cdots & I_m \end{pmatrix}}_{R} I_T,$$

so by the continuous mapping theorem,

$$\pi_{t_1,\cdots,t_k} \circ X^T \xrightarrow{d} N\left[\mathbf{0}, RVR'\right].$$

Furthermore, since $\{X^{T_k}\}_{k\in N_+}$ is a subsequence of $\{X^T\}_{T\in N_+}$ that converges in distribution, and

$$\mu_{T_k} \circ \pi^{-1}_{t_1,\cdots,t_k} \to v \circ \pi^{-1}_{t_1,\cdots,t_k},$$

the uniqueness of weak limits tells us that

$$v \circ \pi^{-1}_{t_1,\cdots,t_k} \sim N\left[\mathbf{0}, RVR'\right].$$

Note that the right hand side depends only on $t_1,\cdots,t_k$, and that $v(0) = \mathbf{0}$ trivially. Therefore, any weak limit of a convergent subsequence of $\{\mu_T\}_{T\in N_+}$ has the same finite-dimensional distributions.

The proof is now essentially complete. Choose any weakly convergent subsequence $\{\mu_{T_k}\}_{k\in N_+}$ of $\{\mu_T\}_{T\in N_+}$, and denote its weak limit by μ. Because μ^W was also found as the weak limit of some convergent subsequence of $\{\mu_T\}_{T\in N_+}$, the preceding result tells us that μ^W and μ have the same finite-dimensional distributions, that is, they agree on the set $\mathcal{C}_f$. Since $\mathcal{C}_f$ is a π-system that generates $\mathcal{B}_{\mathcal{C}}([0,1],\mathbb{R}^m)$, by implication $\mu = \mu^W$, and we can conclude that $\{\mu_{T_k}\}_{k\in N_+}$ converges weakly to μ^W. Finally, by lemma 4.10, because every weakly convergent subsequence of $\{\mu_T\}_{T\in N_+}$ converges weakly to μ^W, the relative compacness of $\{\mu_T\}_{T\in N_+}$ tells us that

$$\mu_T \to \mu^W$$

weakly.

Q.E.D.

6.4.4 Properties of the Wiener Measure

We can show that $\{W(r)\}_{r\in[0,1]}$ possesses the following properties:

- **Initial Value**
  Because $X_T(0) = \mathbf{0}$ for any $T \in N_+$, and $X^T \xrightarrow{d} W$ by construction, the continuous mapping theorem tells us that $\pi_0 \circ X^T \xrightarrow{d} \pi_0 \circ W$. Therefore, $\pi_0 \circ W = W(0) = \mathbf{0}$ as well.

- **Continuous Paths**
  $\{W(r)\}_{r\in[0,1]}$ has continuous paths by construction.

- **Stationary and Independent Increments**
  From the proof of Donsker's theorem, we can infer that, for any $0 = t_0 < t_1, \cdots, t_k \leq 1$, defining

$$\Delta W = \begin{pmatrix} \Delta W(t_0) \\ \vdots \\ \Delta W(t_k) \end{pmatrix},$$

  where

$$\Delta W(t_i) = W(t_i) - W(t_{i-1})$$

  for $0 \leq i \leq k$, we have

$$\Delta W \sim N[\mathbf{0}, V]$$

  for

$$V = \begin{pmatrix} (t_1 - t_0) \cdot I_m & \cdots & O \\ \vdots & \ddots & \vdots \\ O & \cdots & (t_k - t_{k-1}) \cdot I_m \end{pmatrix}.$$

  In other words, $\{\Delta W(t_0), \cdots, \Delta W(t_k)\}$ are independent random vectors such that

$$\Delta W(t_i) \sim N[\mathbf{0}, (t_i - t_{i-1}) \cdot I_m]$$

  for $1 \leq i \leq k$.

  This also implies that

$$W(t_1) \sim N[\mathbf{0}, t_1 \cdot I_m]$$

for any $t_1 \in (0, 1]$, since $W(t_0) = \mathbf{0}$.