Notes on Advanced Probability

August 2018

1 Distribution and Measure—C1, C2

An increasing function f always has left and right limits. Indeed, $f(x-) = \sup_{-\infty<t<x} f(t)$ and $f(x+) = \inf_{x<t<+\infty} f(t)$. Certainly, $f(x-) \leq f(x) \leq f(x+)$. Let

$$\delta_t(x) = \begin{cases} 0 & \text{for } x < t, \\ 1 & \text{for } x \geq t. \end{cases}$$

Call the function δ_t the point mass at t. Let $\{a_n,\ n \geq 1\}$ be any given enumeration of the set of all rational numbers and let $\{b_n,\ n \geq 1\}$ be a set of positive numbers such that $\sum_{n=1}^{+\infty} b_n < +\infty$. Consider

$$f(x) = \sum_{n=1}^{+\infty} b_n \cdot \delta_{a_n}(x).$$

Then f is an increasing function with everywhere dense jumps. On the other hand, the set of discontinuity of an increasing f is countable: each interval $I_x \equiv [f(x-), f(x+)]$ at a jump x contains a rational number and hence the set of jumps corresponds to a subset of the set of rational numbers. When two increasing functions f_1 and f_2 agree on a set D that is dense on the real line $\Re$, the two functions will have the same points of jump of the same size, and they coincide except possibly at some of these points of jump.

For any $x \in \Re$, we can have sequences t_n and t'_n in D with $t_n \uparrow x$ and $t'_n \downarrow x$. Then,

$$f_1(x-) = \lim_{n\longrightarrow+\infty} f_1(t_n) = \lim_{n\longrightarrow+\infty} f_2(t_n) = f_2(x-),$$

$$f_1(x+) = \lim_{n\longrightarrow+\infty} f_1(t'_n) = \lim_{n\longrightarrow+\infty} f_2(t'_n) = f_2(x+).$$

These two functions may assume different values at the countable points x where $f_1(x+) - f_1(x-) = f_2(x+) - f_2(x-) > 0$. If we put $\tilde{f}(x) = f(x+)$ at every $x \in \Re$, then $\tilde{f}$ is increasing and right continuous everywhere. When the increasing function f is only defined on a dense subset D of $\Re$, we can define $\tilde{f}$ on $\Re$ by letting

$$\tilde{f}(x) = \inf_{x<t\in D} f(t).$$

Then $\tilde{f}$ is both increasing and right continuous everywhere.

A real-valued function F with domain $\Re \equiv (-\infty, +\infty)$ that is increasing and right continuous with $F(-\infty) = 0$ and $F(+\infty) = 1$ is called a distribution function. Let $\{a_j\}$ be the countable set of points of jump of F and b_j the size at jump at a_j. Consider the function

$$F_d(x) = \sum_j b_j \delta_{a_j}(x),$$

which represents the sum of all the jumps of F in the half-line $(-\infty, x]$. It is increasing and right continuous with

$$F_d(-\infty) = 0, \qquad\qquad F_d(+\infty) = \sum_j b_j \leq 1.$$

Let $F_c(x) = F(x) - F_d(x)$; then F_c is positive, increasing, and continuous. A distribution function F's decomposition into a jump part F_d and continuous part F_c is also unique. Suppose neither F_c nor F_d is 0, then we may set $\alpha = F_d(+\infty)$ so that $0 < \alpha < 1$,

$$F_1 = \frac{1}{\alpha} F_d, \qquad\qquad F_2 = \frac{1}{1-\alpha} F_c,$$

and write $F = \alpha F_1 + (1-\alpha) F_2$. Thus, every distribution function is a unique convex combination of a discrete and a continuous one.

The Lebesgue measure is denoted by m; "almost everywhere" on the real line without qualification will refer to it and be abbreviated to "a.e.'; an integral written in the form $\int \cdots dt$ is a Lebesgue integral; a function f is said to be "integrable" in (a, b) if and only if (iff) $\int_a^b f(t)dt$ is defined and finite. The class of such functions will be denoted by $L^1(a, b)$, and $L^1(-\infty, +\infty)$ is abbreviated to L^1.

A function F is called absolutely continuous [in $(-\infty, +\infty)$ and with respect to the Lebesgue measure] if and only if there exists a function $f \in L^1$ such that for every $x < x'$,

$$F(x') - F(x) = \int_x^{x'} f(t)dt.$$

It is well known that such a function F has a derivative equal to f a.e. In particular, if F is a distribution function, then

$$f \geq 0 \text{ a.e.} \qquad \text{and} \qquad \int_{-\infty}^{+\infty} f(t) = 1.$$

Conversely, given any $f \in L^1$ satisfying the above, the function F defined by

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

is easily seen to be a distribution function that is absolutely continuous.

A function F is called singular if and only if it is not identically zero and yet F' is in existence and equal to zero almost everywhere. Let F be bounded increasing with $F(-\infty) = 0$, and let F' denote its derivative wherever existing. Then the following are true.

(a) If S denotes the set of all x for which $F'(x)$ exists with $0 \leq F'(x) < +\infty$, then $m(S^c) = 0$.

(b) This F' belongs to L^1, and we have for every $x < x'$,

$$\int_{x}^{x'} F'(t)dt \leq F(x') - F(x).$$

(c) If we put

$$F_{ac}(x) = \int_{-\infty}^{x} F'(t)dt, \qquad F_s(x) = F(x) - F_{ac}(x),$$

then $F'_{ac} = F'$ a.e. so that $F'_s = F' - F'_{ac} = 0$ a.e. and consequently F_s is singular if it is not identically zero. Any positive function f that is equal to F' a.e. is called a density of F. In the above, F_{ac} is called the absolutely continuous part, F_s the singular part of F. The previous F_d is part of F_s as defined here.

It is clear that F_{ac} is increasing and $F_{ac} \leq F$. Also for $x < x'$,

$$F_s(x') - F_s(x) = F(x') - F(x) - \int_{x}^{x'} f(t)dt \geq 0.$$

Hence, F_s is also increasing and $F_s \leq F$. Now, it may be realized that every distribution function can be written as the convex combination of a discrete, a singular continuous, and an absolutely continuous distribution function. Such a decomposition is unique. A singular continuous distribution F can be constructed. Let $J_{n,k}$ be the kth set that has been taken away in the nth round of sculpting the Cantor set. Let $F(x) = k/2^n$ for any $x \in J_{n,k}$.

Let Ω be an "abstract space", namely a nonempty set of elements to be called "points" and denoted generically by ω. A nonempty collection $\mathscr{A}$ of subsets of Ω may have certain "closure properties". Let us list some of them below.

(i) $E \in \mathscr{A} \Longrightarrow E^c \in \mathscr{A}$.

(ii) $E_1 \in \mathscr{A},\ E_2 \in \mathscr{A} \Longrightarrow E_1 \cup E_2 \in \mathscr{A}$.

(iii) $E_1 \in \mathscr{A},\ E_2 \in \mathscr{A} \Longrightarrow E_1 \cap E_2 \in \mathscr{A}$.

(iv) $\forall n \geq 2$: $E_j \in \mathscr{A}, 1 \leq j \leq n \Longrightarrow \bigcup_{j=1}^{n} E_j \in \mathscr{A}$.

(v) $\forall n \geq 2$: $E_j \in \mathscr{A}, 1 \leq j \leq n \Longrightarrow \bigcap_{j=1}^{n} E_j \in \mathscr{A}$.

(vi) $E_j \in \mathscr{A}$; $E_j \subseteq E_{j+1}, 1 \leq j < +\infty \Longrightarrow \bigcup_{j=1}^{+\infty} E_j \in \mathscr{A}$.

(vii) $E_j \in \mathscr{A}$; $E_j \supseteq E_{j+1}, 1 \leq j < +\infty \Longrightarrow \bigcap_{j=1}^{+\infty} E_j \in \mathscr{A}$.

(viii) $E_j \in \mathscr{A}$, $1 \leq j < +\infty \Longrightarrow \bigcup_{j=1}^{+\infty} E_j \in \mathscr{A}$.

(ix) $E_j \in \mathscr{A}$, $1 \leq j < +\infty \Longrightarrow \bigcap_{j=1}^{+\infty} E_j \in \mathscr{A}$.

(x) $E_1 \in \mathscr{A}$, $E_2 \in \mathscr{A}$, $E_1 \subseteq E_2 \Longrightarrow E_2 \backslash E_1 \in \mathscr{A}$.

Under (i), (ii) and (iii) are equivalent; (vi) and (vii) are equivalent; (viii) and (ix) are equivalent. Also, (ii) implies (iv) and (iii) implies (v) by induction. It is trivial that (viii) implies (ii) and (iv); (ix) implies (iii) and (vii).

A nonempty collection $\mathscr{F}$ of subsets of Ω is called a field iff (i) and (ii) hold. It is called a monotone class (M.C.) iff (vi) and (vii) hold. It is called a Borel field (B.F.) iff (i) and (viii) hold. A field is a B.F. if and only if it is also an M.C.

The collection $\mathscr{I}$ of all subsets of Ω is a B.F. called the total B.F.; the collection of the two sets $\{\emptyset, \Omega\}$ is a B.F. called the trivial B.F. If A is any index set and if for every $\alpha \in A$, the collection $\mathscr{F}_\alpha$ is a B.F. (or M.C.) then the intersection $\bigcap_{\alpha \in A} \mathscr{F}_\alpha$ of all these B.F.'s (or M.C.'s) is also a B.F. (or M.C.). Given any nonempty collection $\mathscr{C}$ of sets, there is a minimal B.F. (or M.C.) containing it, that is just the intersection of all B.F.'s (or M.C.'s) containing $\mathscr{C}$. This minimal B.F. (or M.C.) is also said to be generated by $\mathscr{C}$. Let $\mathscr{F}_0$ be a field, $\mathscr{G}$ the minimal M.C. containing $\mathscr{F}_0$, $\mathscr{F}$ the minimal B.F. containing $\mathscr{F}_0$, then $\mathscr{F} = \mathscr{G}$.

Let Ω be a space and $\mathscr{F}$ a B.F. of subsets of Ω. A probability measure $\mathbb{P}(\cdot)$ on $\mathscr{F}$ is a positively valued set function on $\mathscr{F}$ with $\mathbb{P}(\Omega) = 1$ that also satisfies countable additivity:

$$\mathbb{P}\left(\bigcup_j E_j\right) = \sum_j \mathbb{P}(E_j),$$

when $\{E_j\}$ is a countable collection of disjoint sets in $\mathscr{F}$. Derivable properties include:

(a) $\mathbb{P}(E) \leq 1$.

(b) $\mathbb{P}(\emptyset) = 0$.

(c) $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$.

(d) $\mathbb{P}(E \cup F) + \mathbb{P}(E \cap F) = \mathbb{P}(E) + \mathbb{P}(F)$.

(e) $E \subseteq F \Longrightarrow \mathbb{P}(E) = \mathbb{P}(F) - \mathbb{P}(F \backslash E) \leq \mathbb{P}(F)$.

(f) Monotone property, $E_n \uparrow E$ or $E_n \downarrow E \Longrightarrow \mathbb{P}(E_n) \longrightarrow \mathbb{P}(E)$.

(g) Boole's inequality, $\mathbb{P}\left(\bigcup_j E_j\right) \leq \sum_j \mathbb{P}(E_j)$.

The axiom of continuity refers to

$$E_n \downarrow \emptyset \Longrightarrow \mathbb{P}(E_n) \longrightarrow 0.$$

The axioms of finite additivity and of continuity together are equivalent to the axiom of countable additivity. The triple $(\Omega, \mathscr{F}, \mathbb{P})$ is called a probability space; Ω alone is called the sample space, and ω is then a sample point.

Let $\Delta \subseteq \Omega$, then the trace of the B.F. $\mathscr{F}$ on Δ is the collection of all sets of the form $\Delta \cap F$, where $F \in \mathscr{F}$. This is a B.F. of subsets of Δ, and we denote it by $\Delta \cap \mathscr{F}$. Suppose $\Delta \in \mathscr{F}$ and $\mathbb{P}(\Delta) > 0$; then we may define the set function $\mathbb{P}_\Delta$ on $\Delta \cap \mathscr{F}$ as follows:

$$\mathbb{P}_\Delta(E) = \frac{\mathbb{P}(E)}{\mathbb{P}(\Delta)}, \qquad \forall E \in \Delta \cap \mathscr{F}.$$

It is easy to see that $\mathbb{P}_\Delta$ is a probability measure on $\Delta \cap \mathscr{F}$.

Typical probability spaces include the discrete case and the case where $\Omega = (0, 1]$, $\mathscr{F}$ is the minimal B.F. generated by intervals $(-\infty, a]$, $(a, b]$, and $(b, +\infty)$, and $\mathbb{P}$ is the Borel-Lebesgue measure m. Another case is where $\Omega = (-\infty, +\infty)$ and $\mathscr{F}$ is $\mathscr{B}^1$ that is generated from intervals $(a, b]$. The question of probability measures on $\mathscr{B}^1$ is closely related to the theory of distribution functions. There is in fact a one-to-one correspondence between the set functions on the one hand, and the point functions on the other.

Each probability measure μ on $\mathscr{B}^1$ determines a distribution function F through

$$F(x) = \mu((-\infty, x]), \qquad \forall x \in \Re.$$

Consequently, $F(b) - F(a) = \mu((a, b])$, $F(b-) - F(a) = \mu((a, b))$, $F(b-) - F(a-) = \mu([a, b))$, and $F(b) - F(a-) = \mu([a, b])$. Especially for $x_n \uparrow x$, note $(-\infty, x_n] \uparrow (-\infty, x)$ and hence

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

On the flip side, each distribution function F determines a probability measure μ on $\mathscr{B}^1$. This is the classical theory of Lebesgue-Stieltjes measure. Starting from $\mu((a, b]) = F(b) - F(a)$, we can obtain $\mu((a, b)) = F(b-) - F(a)$ and then μ for all open sets and closed sets. For subset S of $\Re$, define outer and inner measures

$$\mu^*(S) = \inf_{U \text{ open, } U \supseteq S} \mu(U), \qquad \mu_*(S) = \sup_{C \text{ closed, } C \subseteq S} \mu(C) = 1 - \mu^*(S^c).$$

We call S measurable when $\mu_*(S) = \mu^*(S)$. It can be shown that (a) all measurable sets form a B.F., say $\mathscr{L}$; (b) on it the function $\mu = \mu^* = \mu_*$ is a probability measure.

Note $\mathscr{L}$ may be larger than $\mathscr{B}^1$; indeed it is. The outer and inner measures are useful for approximations. For each measurable set S and $\epsilon > 0$, there exist an open set U and a closed set C such that $U \supseteq S \supseteq C$ and

$$\mu(U) - \epsilon \leq \mu(S) \leq \mu(C) + \epsilon.$$

An alternative definition of measurability is through the outer measure alone and a criterion by Caratheodory. The question of whether there exists another ν that corresponds to the given distribution function F is answered in the following manner.

A measure μ on $(\Omega, \mathscr{F})$ is σ-finite when there exists a sequence of sets $E_n \in \mathscr{F}$, $E_n \uparrow \Omega$ with $\mu(E_n) < +\infty$. Let μ and ν be two measures defined on the same B.F. $\mathscr{F}$, which is generated by the field $\mathscr{F}_0$. If either μ or ν is σ-finite on $\mathscr{F}_0$, and $\mu(E) = \nu(E)$ for every $E \in \mathscr{F}_0$, then the same is true for every $E \in \mathscr{F}$, and thus $\mu = \nu$. Thus, when μ and ν are σ-finite measures on $\mathscr{B}^1$ that agree on all intervals of the eight kinds: $(a, b]$, (a, b), $[a, b)$, $[a, b]$, $(-\infty, b]$, $(-\infty, b)$, $[a, +\infty)$, and $(a, +\infty)$ or merely on those with the endpoints in a given dense set D, then they agree on $\mathscr{B}^1$.

Using the technique of trace, the correspondence between distribution functions and probability measures can be restricted to $[0, 1]$. The most interesting is the uniform distribution. The corresponding measure m on $\mathscr{B}$ is the usual Borel measure on $[0, 1]$, while its extension on the strictly larger $\mathscr{L}$ is the Lebesgue measure. Indeed, $(\mathscr{L}, m)$ is the completion of $(\mathscr{B}, m)$. Note a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is said to be complete iff any subset of a set in $\mathscr{F}$ with $\mathbb{P}(F) = 0$ also belongs to $\mathscr{F}$. Any probability space $(\Omega, \mathscr{F}, \mathbb{P})$ can be completed: there exists a complete space $(\Omega, \overline{\mathscr{F}}, \overline{\mathbb{P}})$ such that $\mathscr{F} \subseteq \overline{\mathscr{F}}$ and yet $\mathbb{P} = \overline{\mathbb{P}}$ on $\mathscr{F}$. Indeed, let

$$\overline{\mathscr{F}} = \{E \subseteq \Omega : \ E \triangle F \in \mathscr{N} \quad \text{for some } F \in \mathscr{F}\},$$

where $\mathscr{N}$ is the collection of sets that are subsets of null sets and for each $E \in \overline{\mathscr{F}}$, put $\overline{\mathbb{P}}(E) = \mathbb{P}(F)$ where F is any set that satisfies the condition above for the given E.

What is the advantage of completion? Suppose a certain property is known to hold outside a certain set N with $\mathbb{P}(N) = 0$. Then the exact set N' on which it fails to hold is a subset of N, which is not necessarily in $\mathscr{F}$. But N' will be in $\overline{\mathscr{F}}$ with $\overline{\mathbb{P}}(N') = 0$. We need the measurability of the exact exceptional set to facilitate certain dispositions.

2 Random Variable and Expectation—C3

Let the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ be given. Let $\Re^1 \equiv (-\infty, +\infty)$ be the real line, $\Re^* \equiv [-\infty, +\infty]$ the extended real line, $\mathscr{B}^1$ te Euclidean Borel field on $\Re^1$, $\mathscr{B}^*$ the extended Borel field. A set in $\mathscr{B}^*$ is just a set in $\mathscr{B}^1$ possibly enlarged by one or both points $\pm\infty$. A real, extended-valued random variable is a function X whose domain is a set Δ in $\mathscr{F}$ and whose range is contained in $\Re^*$ such that for each $B \in \mathscr{B}^*$,

$$\{\omega :\ X(\omega) \in B\} \in \Delta \cap \mathscr{F},$$

where $\Delta \cap \mathscr{F}$ is the trace of $\mathscr{F}$ on Δ. Mostly, we suppose $\Delta = \Omega$ and X is real-valued.

The inverse mapping X^{-1} is defined so that

$$X^{-1}(A) = \{\omega :\ X(\omega) \in A\}, \qquad\qquad \forall A \subseteq \Re^1.$$

It carries members of $\mathscr{B}^1$ onto members of $\mathscr{F}$:

$$X^{-1}(B) \in \mathscr{F}, \qquad\qquad \forall B \in \mathscr{B}^1.$$

Thus, a random variable is a measurable function from Ω to $\Re^1$. Note the inverse mapping has the following properties:

$$X^{-1}(A^c) = (X^{-1}(A))^c,$$
$$X^{-1}\left(\bigcup_\alpha A_\alpha\right) = \bigcup_\alpha X^{-1}(A_\alpha),$$
$$X^{-1}\left(\bigcap_\alpha A_\alpha\right) = \bigcap_\alpha X^{-1}(A_\alpha).$$

X is a random variable iff for each real number x, or each real number in a dense subset,

$$\{\omega :\ X(\omega) \leq x\} \in \mathscr{F}.$$

To prove the “if” part, we construct

$$\mathscr{A} \equiv \{S \subseteq \Re^1 :\ X^{-1}(S) \in \mathscr{F}\},$$

and show that $\mathscr{A}$ is a B.F. containing all the $(-\infty, x]$'s; thus, $\mathscr{B}^1 \subseteq \mathscr{A}$ and $X^{-1}(\mathscr{B}^1) \subseteq \mathscr{F}$.

Each random variable X on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ induces a probability space $(\Re^1, \mathscr{B}^1, \mu)$ by means of

$$\mu(B) = \mathbb{P}(X^{-1}(B)) \equiv \mathbb{P}[X \in B], \qquad\qquad \forall B \in \mathscr{B}^1.$$

The collection $\{X^{-1}(B), B \in \mathscr{B}^1\}$ is called the B.F. generated by X. It is the smallest Borel subfield of $\mathscr{F}$ which contains all sets of the form $\{\omega : X(\omega) \leq x\}$, where $x \in \Re^1$. The above

$\mu \equiv \mathbb{P} \circ X^{-1}$ is the "probability distribution measure" of X, and its associated distribution function F will be called the distribution function of X. Specifically, F is given by

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

While the random variable X determines μ and therefore F, the converse is false. A family of random variables having the same distribution is said to be "identically distributed".

If X is a random variable and f a Borel measurable function on $(\Re^1, \mathscr{B}^1)$, then $f(X)$ is a random variable. In the two-dimensional Euclidean space $\Re^2$, the Euclidean Borel field $\mathscr{B}^2$ is generated by rectangles of the form

$$\{(x, y) : \ a \leq x \leq b, c \leq y \leq d\}.$$

A *fortiori*, it is also generated by product sets of the form

$$B_1 \times B_2 \equiv \{(x, y) : x \in B_1, y \in B_2\},$$

where B_1 and B_2 belong to $\mathscr{B}^1$. A function from $\Re^2$ into $\Re^1$ is called a Borel measurable function iff $f^{-1}(\mathscr{B}^1) \subseteq \mathscr{B}^2$. Now let X and Y be two random variables on $(\Omega, \mathscr{F}, \mathbb{P})$. The random vector (X, Y) induces a probability ν on $\mathscr{B}^2$ as follows:

$$\nu(A) = \mathbb{P}[(X, Y) \in A].$$

The inverse mapping $(X, Y)^{-1}$ can be defined by

$$(X, Y)^{-1}(A) = \{\omega : (X, Y) \in A\}, \qquad \forall A \in \mathscr{B}^2.$$

If X and Y are two random variables and f is a Borel measurable function of two variables, then $f(X, Y)$ is a random variable. In particular, $X \vee Y$, $X \wedge Y$, $X + Y$, $X - Y$, $X \cdot Y$, and X/Y provided Y does not vanish are all random variables.

If $\{X_j, j \geq 1\}$ is a sequence of random variables, then $\inf_j X_j$, $\sup_j X_j$, $\liminf_j X_j$, and $\limsup_j X_j$ are random variables, not necessarily finite-valued with probability one though everywhere defined, and $\lim_j X_j$ is a random variable on the set Δ on which there is either convergence or divergence to $\pm\infty$. The indicator function 1_Δ is defined to indicate whether or not $\omega \in \Delta$ for any given $\Delta \subseteq \Omega$. It is a random variable iff $\Delta \in \mathscr{F}$.

The concept of "expectation" is the same as that of integration in the probability space with respect to the measure $\mathbb{P}$. For each positive discrete random variable belonging to the weighted partition $\{\Lambda_j; b_j\}$, we define the expectation to be

$$\mathbb{E}[X] \equiv \sum_j b_j \mathbb{P}(\Lambda_j).$$

This is either a positive number or $+\infty$. Let X be a positive random variable. The set

$$\Lambda_{mn} \equiv \left\{\omega : \frac{n}{2^m} \leq X(\omega) < \frac{n+1}{2^m}\right\}$$

belongs to $\mathscr{F}$. For each m, let X_m denote the random variable belonging to the weighted partition $\{\Lambda_{mn}; n/2^m\}$. It is easy to see that

$$X_m(\omega) \leq X_{m+1}(\omega); \qquad\qquad 0 \leq X(\omega) - X_m(\omega) < \frac{1}{2^m}.$$

Consequently, there is monotone convergence:

$$\lim_m X_m(\omega) \uparrow X(\omega), \qquad\qquad \forall \omega \in \Omega.$$

The expectation of X_m has just been defined; it is

$$\mathbb{E}[X_m] = \sum_{n=0}^{+\infty} \frac{n}{2^m} \cdot \mathbb{P}\left[\frac{n}{2^m} \leq X < \frac{n+1}{2^m}\right].$$

If for one m we have $\mathbb{E}[X_m] = +\infty$, then we define $\mathbb{E}[X] = +\infty$; otherwise, define

$$\mathbb{E}[X] = \lim_m \uparrow \mathbb{E}[X_m].$$

For an arbitrary X, we again decompose it into the positive and negative portions X^+ and X^-, and define its expectation as

$$\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-],$$

unless both terms on the right are $+\infty$. The expectation, when in existence, is also denoted

$$\int_\Omega X(\omega)\mathbb{P}(d\omega).$$

For each $\Lambda \in \mathscr{F}$, we define

$$\int_\Lambda X(\omega) \cdot \mathbb{P}(d\omega) \equiv \mathbb{E}[X \cdot 1_\Lambda].$$

We shall say X is integrable with respect to $\mathbb{P}$ over Λ iff the integral exists and is finite.

The general integral has the familiar properties of the Lebesgue integral on $[0, 1]$.

(i) Absolute integrability. $\int_\Lambda X d\mathbb{P}$ is finite iff

$$\int_\Lambda |X| d\mathbb{P} < +\infty.$$

(ii) Linearity.

$$\int_\Lambda (aX + bY) d\mathbb{P} = a\int_\Lambda X d\mathbb{P} + b\int_\Lambda Y d\mathbb{P},$$

provided that the right-hand side is meaningful.

(iii) Additivity over sets. If the Λ_n's are disjoint, then

$$\int_{\bigcup_n \Lambda_n} X d\mathbb{P} = \sum_n \int_{\Lambda_n} X d\mathbb{P}.$$

(iv) Positivity. If $X \geq 0$ a.e. on Λ, then

$$\int_\Lambda X d\mathbb{P} \geq 0.$$

(v) Monotonicity. If $X_1 \leq X \leq X_2$ a.e. on Λ, then

$$\int_\Lambda X_1 d\mathbb{P} \leq \int_\Lambda X d\mathbb{P} \leq \int_\Lambda X_2 d\mathbb{P}.$$

(vi) Mean value theorem. If $a \leq X \leq b$ a.e. on Λ, then

$$a\mathbb{P}(\Lambda) \leq \int_\Lambda X d\mathbb{P} \leq b\mathbb{P}(\Lambda).$$

(vii) Modulus inequality.

$$\left| \int_\Lambda X d\mathbb{P} \right| \leq \int_\Lambda |X| d\mathbb{P}.$$

(viii) Dominated convergence theorem. If $\lim_n X_n = X$ a.e. on Λ, for any n we have $|X_n| \leq Y$ a.e. on Λ with $\int_\Lambda Y d\mathbb{P} < +\infty$, then

$$\int_\Lambda X d\mathbb{P} = \lim_n \int_\Lambda X_n d\mathbb{P}.$$

(ix) Bounded convergence theorem. If $\lim_n X_n = X$ a.e. on Λ and there is a constant M such that for any n we have $|X_n| \leq M$ a.e. on Λ, then the above is still true.

(x) Monotone convergence theorem. If $X_n \geq 0$ and $X_n \uparrow X$ a.e. on Λ, then

$$\lim_n \uparrow \int_\Lambda X_n d\mathbb{P} = \int_\Lambda X d\mathbb{P}.$$

(xi) Integration term by term. If

$$\sum_n \int_\Lambda |X_n| d\mathbb{P} < +\infty,$$

then $\sum_n |X_n| < +\infty$ a.e. on Λ so that $\sum_n X_n$ converges a.e. on Λ and

$$\int_\Lambda \sum_n X_n d\mathbb{P} = \sum_n \int_\Lambda X_n d\mathbb{P}.$$

(xii) Fatou's lemma. If $X_n \geq 0$ a.e. on Λ, then

$$\int_\Lambda \liminf_n X_n d\mathbb{P} \leq \liminf_n \int_\Lambda X_n d\mathbb{P}.$$

In addition, we have

$$\sum_{n=1}^{+\infty} \mathbb{P}[|X| \geq n] \leq \mathbb{E}[|X|] \leq 1 + \sum_{n=1}^{+\infty} \mathbb{P}[|X| \geq n].$$

If X takes only positive integer values, then

$$\mathbb{E}[X] = \sum_{n=1}^{+\infty} \mathbb{P}[X \geq n].$$

Let X on $(\Omega, \mathscr{F}, \mathbb{P})$ induce the probability space $(\Re^1, \mathscr{B}^1, \mu)$ and let f be measurable. Then

$$\int_\Omega f(X(\omega))\mathbb{P}(d\omega) = \int_{\Re^1} f(x)\mu(dx).$$

The proof starts from $f = 1_B$ for some $B \in \mathscr{B}^1$ and proceed with monotone convergence theorem. Let μ_X and F_X denote, respectively, the probability measure and distribution function induced by X. Then,

$$\mathbb{E}[f(X)] = \int_{\Re^1} f(x)\mu_X(dx) = \int_{\infty}^{+\infty} f(x)dF_X(x).$$

With two dimensions, we have the following. Let (X, Y) on $(\Omega, \mathscr{F}, \mathbb{P})$ induce the probability space $(\Re^2, \mathscr{B}^2, \mu^2)$ and let f be a measurable function of two variables. Then

$$\int_\Omega f(X(\omega), Y(\omega))\mathbb{P}(d\omega) = \int\int_{\Re^2} f(x, y)\mu^2(dx, dy).$$

All the above can be used to show that

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

For moments, we have

$$\mathbb{E}[(X - a)^r] = \int_{\Re^1} (x - a)^r \mu(dx) = \int_{-\infty}^{+\infty} (x - a)^r dF(x).$$

The expectation is a moment called the mean. At the second order, we have

$$\mathrm{var}[X] = \sigma^2[X] = [(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

The inequality that $\sigma^2[X] \leq \mathbb{E}[X^2]$ will be used a great deal. The well known inequalities of Holder and Minkowski may be written as follows. Let X and Y be random variables, $1 < p < +\infty$, and $1/p + 1/q = 1$. Then

$$|\mathbb{E}[XY]| \leq \mathbb{E}[|XY|] \leq (\mathbb{E}[|X|^p])^{1/p}(\mathbb{E}[|Y|^q])^{1/q},$$

$$(\mathbb{E}[|X+Y|^p])^{1/p} \leq (\mathbb{E}[|X|^p])^{1/p} + (\mathbb{E}[|Y|^p])^{1/p}.$$

If $Y = 1$, the first would become $\mathbb{E}[|X|] \leq (\mathbb{E}[|X|^p])^{1/p}$. At $p = 2$, it is called the Cauchy-Schwarz inequality. It will also lead to

$$(\mathbb{E}[|X|^r])^{1/r} \leq (\mathbb{E}[|X|^{r'}])^{1/r'}, \qquad \forall 0 < r < r' < +\infty.$$

The last will be referred to as the Liapounov inequality, which is a special case of the Jensen's inequality: If φ is a convex function on $\Re^1$, and X and $\varphi(X)$ are integrable, then

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

Chebyshev inequality: If φ is a strictly positive and increasing function on $(0, +\infty)$, $\varphi(u) = \varphi(-u)$, and X is a random variable such that $\mathbb{E}[\varphi(X)] < +\infty$, then for $\mu > 0$,

$$\mathbb{P}[|X| \geq u] \leq \frac{\mathbb{E}[\varphi(X)]}{\varphi(u)}.$$

The most familiar application is when $\varphi(u) = |u|^p$ for $0 < p < +\infty$, so that the inequality yields an upper bound for the tail probability in terms of an absolute moment.

3 Measure and Integral—C10

Given a measure μ on a field $\mathscr{F}_0$ of Ω, we can define μ^* on the total B.F. $\mathscr{I}$:

$$\mu^*(A) = \inf\left\{\sum_j \mu(B_j) : \ B_j \in \mathscr{F}_0 \text{ for all } j \quad \text{and} \ \bigcup_j B_j \supseteq A\right\}, \qquad \forall A \in \mathscr{I}.$$

We have $\mu^* = \mu$ on $\mathscr{F}_0$ and μ^* on $\mathscr{I}$ is an outer measure, i.e., one that satisfies $\mu^*(\emptyset) = 0$, monotonicity, and sub-additivity. Define $\mathscr{F}^* \subseteq \mathscr{I}$ so that

$$\mathscr{F}^* = \{A \subseteq \Omega : \ \mu^*(Z) = \mu^*(AZ) + \mu^*(A^cZ) \quad \text{for every } Z \subseteq \Omega\}.$$

It can be shown that $\mathscr{F}^*$ is a B.F. and on it μ^* is a measure. To prove, we first show that $\mathscr{F}_0 \subseteq \mathscr{F}^*$. Note $\mathscr{F}^*$ is also closed under complementation. After dealing with its closure under union, we can demonstrate that $\mathscr{F}^*$ is a field. The rest will follow.

Let $\mathscr{F}_0$ be a field and $\mathscr{F}$ the Borel field generated by $\mathscr{F}_0$. If $A_n \in \mathscr{F}$ for each n, the countable union $\bigcup_n A_n$ and countable intersection $\bigcap_n A_n$ both belong to $\mathscr{F}$. If μ is a measure on $\mathscr{F}$, two fundamental properties will follow:

(a) (increasing limit) if $A_n \subseteq A_{n+1}$ for all n and $A_n \uparrow A \equiv \bigcup_n A_n$, then

$$\lim_n \uparrow \mu(A_n) = \mu(A).$$

(b) (decreasing limit) if $A_n \supseteq A_{n+1}$ for all n, $A_n \downarrow A \equiv \bigcap_n A_n$, and for some n we have $\mu(A_n) < +\infty$, then

$$\lim_n \downarrow \mu(A_n) = \mu(A).$$

Let μ_1 and μ_2 be two measures on $\mathscr{F}$ that agree on $\mathscr{F}_0$. If one of them, hence both are σ-finite on $\mathscr{F}_0$, then they agree on $\mathscr{F}$. Let $\{\Omega_n\}$ be such that $\bigcup_n \Omega_n = \Omega$ and $\mu_1(\Omega_n) = \mu_2(\Omega_n) < +\infty$. Define

$$\mathscr{C} = \{A \subseteq \Omega : \ \mu_1(\Omega_n A) = \mu_2(\Omega_n A) \text{ for all } n\}.$$

We can show that $\mathscr{C} \supseteq \mathscr{F}_0$, it is a M.C. and hence a B.F. that contains $\mathscr{F}$. From this, we see that under the σ-finite assumption, the outer measure μ^* restricted to the minimal Borel field $\mathscr{F}$ containing $\mathscr{F}_0$ is the unique extension of μ from $\mathscr{F}_0$ to $\mathscr{F}$.

The extension to the bigger $\mathscr{F}^*$ is also unique. Let $\mathscr{F}_{0\sigma\delta}$ be the collection of all sets of the form $\bigcap_{m=1}^{+\infty} \bigcup_{n=1}^{+\infty} B_{mn}$ where each $B_{mn} \in \mathscr{F}_0$ and $\mathscr{F}_{0\delta\sigma}$ be the collection of all sets of the form $\bigcup_{m=1}^{+\infty} \bigcap_{n=1}^{+\infty} B_{mn}$. Both collections belong to $\mathscr{F}$. Suppose $A \in \mathscr{F}^*$. Then there exists $B \in \mathscr{F}_{0\sigma\delta}$ such that $A \subseteq B$ and $\mu^*(A) = \mu^*(B)$; if μ is σ-finite on $\mathscr{F}_0$, then there exists $C \in \mathscr{F}_{0\delta\sigma}$ such that $C \subseteq A$ and $\mu^*(C) = \mu^*(A)$.

Using the above, we can prove that the following collections of subsets of Ω are identical:

(i) $A \subseteq \Omega$ and the outer measure $\mu^*(A) = 0$;

(ii) $A \in \mathscr{F}^*$ and $\mu^*(A) = 0$;

(iii) $A \subseteq B$ where $B \in \mathscr{F}$ and $\mu(B) = 0$.

They all converge on the collection of null sets $\mathscr{N}(\mathscr{F}^*, \mu^*)$. As a consequence, any subset of an $(\mathscr{F}^*, \mu^*)$-null set is an $(\mathscr{F}^*, \mu^*)$-null set. Thus, the measure space $(\Omega, \mathscr{F}^*, \mu^*)$ is complete. It is also the smallest of such. Let $(\Omega, \mathscr{G}, \nu)$ be a complete measure space; $\mathscr{G} \supseteq \mathscr{F}_0$ and $\nu = \mu$ on $\mathscr{F}_0$. If μ is σ-finite on $\mathscr{F}_0$, then

$$\mathscr{G} \supseteq \mathscr{F}^* \quad \text{and} \quad \nu = \mu^* \text{ on } \mathscr{F}^*.$$

The minimal Borel field containing all $(a, b]$ will be denoted by $\mathscr{B}$ and called the Borle field of $\Re$. Since a bounded open interval is the countable union of intervals like $(a, b]$ and any open set in $\Re$ is the countable union of bounded open intervals, the Borel field $\mathscr{B}$ contains all open sets; hence by complementation it contains all closed sets, in particular all compact sets. Starting from one of these collections, forming countable union and countable intersection successively, a countable number of times, one can build up $\mathscr{B}$ through a transfinite induction.

Suppose a measure m satisfying $0 \leq m((a,b]) < +\infty$ has been defined on $\mathscr{B}$. Then we can define a function F on $\Re$ so that $F(0) = 0$, $F(x) = m((0,x])$ for $x > 0$, and $F(x) = -m((x,0])$ for $x < 0$. This function may be called the "generalized distribution" for m. We see that F is finite, $m((a,b]) = F(b) - F(a)$, $F(-\infty) \geq -\infty$, and $F(+\infty) \leq +\infty$. Next, F has unilateral limits everywhere and is right continuous: $F(x-) \leq F(x) = F(x+)$. The right continuity follows from the monotone limit properties of m. The measure of a single point x is $m(x) = F(x) - F(x-)$. The simplest example of F is given by $F(x) = x$. In this case, $m((a,b]) = b - a$. The measure is the length of the line segment from a to b.

Given F as specified above, we can construct a measure m on $\mathscr{B}$ and a larger Borel field $\mathscr{B}^*$ that fulfills $m((a,b]) = F(b) - F(a)$. First, the minimal field $\mathscr{B}_0$ containing all $(a,b]$ is the one that contains all the B's, where for some n,

$$B = \bigcup_{j=1}^{n}(a_j, b_j].$$

For such a B, it must follow that

$$m(B) = \sum_{j=1}^{n}(F(b_j) - F(a_j)).$$

When l is finite, it is easy to show additivity, that

$$m\left(\bigcup_{k=1}^{l} B_k\right) = \sum_{k=1}^{l} m(B_k).$$

To prove the case where $l = +\infty$, we need Borel's Lemma. If $-\infty \leq a < b \leq +\infty$ and $(a,b] = \bigcup_{j=1}^{+\infty}(a_j, b_j]$ where $a_j < b_j$ for each j, and the intervals $(a_j, b_j]$ are disjoint, then

$$F(b) - F(a) = \sum_{j=1}^{+\infty}(F(b_j) - F(a_j)).$$

Once m has been established as a measure for $\mathscr{B}_0$, the general method can be applied to $(\Re, \mathscr{B}_0, m)$. We can define an outer measure m^* that agrees with m on $\mathscr{B}_0$. The Borel field $\mathscr{B}^*$ can be defined using Caratheordory's criterion. Consequently, $(\Re, \mathscr{B}^*, m^*)$ is a complete measure space. Note m is σ-finite on $\mathscr{B}_0$ because $(-n, n] \uparrow (-\infty, +\infty)$ and $m((-n,n])$ is finite. Hence, the restriction of m^* to $\mathscr{B}$ is the unique extension of m from $\mathscr{B}_0$ to $\mathscr{B}$.

For the special case with $F(x) = x$, the measure on $\mathscr{B}$ is called Borel and that on $\mathscr{B}^*$ is called Lebesgue. It was first constructed by Lebesgue from an outer and inner measure. The latter was later bypassed by Caratheodory, whose method has been adopted here. A member of $\mathscr{B}^*$ is usually called Lebesgue-mesurable. The generalization to a generalized distribution function F is sometimes referred to as Borel-Lebesgue-Stiejtjes.

The measure space $(\Omega, \mathscr{F}, \mu)$ is fixed. A function f with domain Ω and range $\Re^* \equiv [-\infty, +\infty]$ is called $\mathscr{F}$-measurable iff for each real number c we have

$$\{f \leq c\} \equiv \{\omega \in \Omega : \ f(\omega) \leq c\} \in \mathscr{F}.$$

We write $f \in \mathscr{F}$ in this case. It follows that for each set $A \in \mathscr{B}$, we have

$$\{f \in A\} \in \mathscr{F};$$

and both $\{f = +\infty\}$ and $\{f = -\infty\}$ also belong to $\mathscr{F}$. A function $f \in \mathscr{F}$ with range a countable set in $[0, +\infty]$ will be called a basic function. Let $\{a_j\}$ be its range and $A_j \equiv \{f = a_j\}$. Then the A_j's are disjoint sets with union Ω and

$$f = \sum_j a_j 1_{A_j},$$

where the sum is over a countable set of j. We proceed to define an integral for functions in $\mathscr{F}$, in three stages, beginning with basic functions.

For the basic function f, its integral is defined to be

$$E(f) = \sum_j a_j \mu(A_j),$$

with $\infty \cdot 0$ and $0 \cdot \infty$ understood as 0. The integral is also denoted by

$$\int f d\mu = \int_\Omega f(\omega) \mu(d\omega).$$

We can rely on the following double limit lemma to derive properties for the integral. Let $\{C_{jk} : j, k \in N\}$ be a doubly indexed array of real numbers with the following properties:

for each fixed j, the sequence $\{C_{jk} : k \in N\}$ is increasing in k;

for each fixed k, the sequence $\{C_{jk} : j \in N\}$ is increasing in j.

Then, we have

$$\lim_j \uparrow \lim_k \uparrow C_{jk} = \lim_k \uparrow \lim_j \uparrow C_{jk} \leq +\infty.$$

If f and g are basic and $f \leq g$, then

$$E(f) \leq E(g).$$

For positive numbers a and b, we also have $af + bg$ being basic and

$$E(af + bg) = aE(f) + bE(g).$$

Let $A \in \mathscr{F}$ and f be a basic function. Then the product $1_A f$ is a basic function and its integral will be denoted by

$$E(A; f) = \int_A f(\omega)\mu(d\omega) = \int_A f d\mu.$$

Let $A_n \in \mathscr{F}$, $A_n \subseteq A_{n+1}$ for all n and $A \equiv \bigcup_n A_n$. Then

$$E(A; f) = \lim_n \uparrow E(A_n; f).$$

Let $\{f_n\}$ and $\{g_n\}$ be two increasing sequences of basic functions such that $\lim_n \uparrow f_n = \lim_n \uparrow g_n$. Then

$$\lim_n \uparrow E(f_n) = \lim_n \uparrow E(g_n).$$

Let f_n and f be basic functions such that $f_n \uparrow f$, then $E(f_n) \uparrow E(f)$.

The collection of positive $\mathscr{F}$-measurable functions will be denoted by $\mathscr{F}_+$. We now approximate a function from the collection by basic functions. Define a function on $[0, +\infty]$ by the symbol $)\cdot]$ such that $)0] = 0$, $) + \infty] = +\infty$, and $)x] = n - 1$ for $x \in (n-1, n]$, $n \in N$. For any $f \in \mathscr{F}_+$, the approximating sequence $\{f^{(m)}\}$, $m \in N$, satisfies

$$f^{(m)}(\omega) = \frac{)2^m f(\omega)]}{2^m}.$$

Each $f^{(m)}$ is a basic function with range in the set of dyadic numbers: $\{k/2^m\}$ where k is a positive integer or $+\infty$. We have $f^{(m)} \leq f^{(m+1)}$ by the magic property of bisection. Finally, $f^{(m)} \uparrow f$ because $f^{(m)}(\omega) \geq f(\omega) - 1/2^m$.

For $f \in \mathscr{F}_+$, its integral is defined to be

$$E(f) = \lim_m \uparrow E(f^{(m)}).$$

We still have $E(f) \leq E(g)$ when $f \leq g$ and $E(af + bg) = aE(f) + bE(g)$. Also, for $f \in \mathscr{F}_+$, the function of sets defined on $\mathscr{F}$ by

$$A \Longrightarrow E(A; f)$$

is a measure. There are three fundamental theorems relating the convergence of functions with the convergence of their integrals. The monotone convergence theorem is as follows. Let $\{f_n\}$ be an increasing sequence of functions in $\mathscr{F}_+$, with limit $f : f_n \uparrow f$. Then

$$\lim_n \uparrow E(f_n) = E(f) \leq +\infty.$$

We can then derive Lebesgue's theorem in its pristine positive guise. Let $f_n \in \mathscr{F}_+$, $n \in N$. Suppose $\lim_n f_n = 0$ and $E(\sup_n f_n) < +\infty$. Then

$$\lim_n E(f_n) = 0.$$

Fatou's lemma is as follows. Let $\{f_n\}$ be an arbitrary sequence of functions in $\mathscr{F}_+$. Then

$$E\left(\liminf_n f_n\right) \leq \liminf_n E(f_n).$$

In the final stage, we can decompose $f \in \mathscr{F}$ into positive and negative portions $f^+, f^- \in \mathscr{F}_+$ so that $f = f^+ - f^-$ and $|f| = f^+ + f^-$. We define

$$E(f) = E(f^+) - E(f^-),$$

as long as $\infty - \infty$ does not occur. We say f is integrable, or $f \in L^1$, iff both $E(f^+)$ and $E(f^-)$ are finite; in this case $E(f)$ is a finite number. A few properties follow:

(i) The function $f \in \mathscr{F}$ is integrable if and only if $|f|$ is integrable; we have

$$|E(f)| \leq E(|f|).$$

(ii) For any $f \in \mathscr{F}$ and any null set A, we have

$$E(A; f) = \int_A f d\mu = 0; \qquad E(f) = E(A^c; f) = \int_{A^c} f d\mu.$$

(iii) If $f \in L^1$, then the set $\{\omega \in \Omega : \ |f(\omega)| = +\infty\}$ is a null set.

(iv) If $f \in L^1$, $g \in \mathscr{F}$, and $|g| \leq |f|$ a.e., then $g \in L^1$.

(v) If $f \in \mathscr{F}$, $g \in \mathscr{F}$, and $g = f$ a.s., then $E(g)$ exists if and only if $E(f)$ exists, and then $E(g) = E(f)$.

(vi) If $\mu(\Omega) < +\infty$, then any a.e. bounded $\mathscr{F}$-measurable function is integrable.

It is convenient to define for any $f \in \mathscr{F}$ a class of functions denoted by $\mathbb{C}(f)$, so that $g \in \mathbb{C}(f)$ iff $g = f$ a.e. When $(\Omega, \mathscr{F}, \mu)$ is a complete measure space, such a g is automatically in $\mathscr{F}$. A member of $\mathbb{C}(f)$ may be called a version of f, and may be substituted for f whenever a null set "does not count". In functional analysis, it is the class $\mathbb{C}(f)$ rather than an individual f that is a member of L^1. For integral on $\mathscr{F}$, we still have $E(f) \leq E(g)$ when $f \leq g$ a.e. and $E(f+g) = E(f) + E(g)$.

Lebesgue's dominated convergence theorem: Let $f_n \in \mathscr{F}$; suppose

(a) $\lim_n f_n = f$ a.e.;

(b) there exists $\phi \in L^1$ such that for all n: $|f_n| \leq \phi$ a.e.

Then, $\lim_n E(|f_n - f|) = 0$. A corollary of this is that

$$\lim_n \int_B f_n d\mu = \int_B f d\mu$$

uniformly in $B \in \mathscr{F}$.

4 Independence and Product Measure, C3, B3, B18

The random variables $\{X_j, 1 \le j \le n\}$ are said to be independent iff for any linear Borel sets $\{B_j, 1 \le j \le n\}$ we have

$$\mathbb{P}\left[\bigcap_{j=1}^{n}(X_j \in B_j)\right] = \prod_{j=1}^{n} \mathbb{P}[X_j \in B_j].$$

The random variables of an infinite family are said to be independent iff those in every finite subfamily are. The above is equivalent to

$$\mathbb{P}\left[\bigcup_{j=1}^{n}(X_j \le x_j)\right] = \prod_{j=1}^{n} \mathbb{P}[X_j \le x_j].$$

In terms of the measure $\otimes_{j=1}^{n}\mu_j$ induced by the random vector $(X_1, ..., X_n)$ on $(\Re^n, \mathscr{B}^n)$– whose existence and uniqueness we shall show later, and $\{\mu_j, 1 \le j \le n\}$ induced by all the X_j's on $(\Re^1, \mathscr{B}^1)$, the above may be written as

$$\otimes_{j=1}^{n}\mu_j\left(\times_{k=1}^{n} B_k\right) = \prod_{j=1}^{n} \mu_j(B_j).$$

We may introduce the n-dimensional distribution function F corresponding to $\otimes_{j=1}^{n}\mu_j$:

$$F(x_1, ..., x_2) = \mathbb{P}[X_j \le x_j, 1 \le j \le n] = \otimes_{j=1}^{n}\mu_j\left(\times_{k=1}^{n}(-\infty, x_k]\right).$$

The independence requirement may be written as

$$F(x_1, ..., x_n) = \prod_{j=1}^{n} F_j(x_j).$$

The events $\{E_j, 1 \le j \le n\}$ are said to be independent iff their indicators are independent; this is equivalent to: for any subset $\{j_1, ..., j_l\}$ of $\{1, ..., n\}$,

$$\mathbb{P}\left[\bigcap_{k=1}^{l} E_{j_k}\right] = \prod_{k=1}^{l} \mathbb{P}[E_{j_k}].$$

If $\{X_j, 1 \le j \le n\}$ are independent random variables and $\{f_j, 1 \le j \le n\}$ are Borel measurable functions, then $\{f_j(X_j), 1 \le j \le n\}$ are independent random variables. If X and Y are independent and both have finite expectations, then

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

The first proof goes through the usual discrete approximation and monotone convergence theorem. In the second proof, we note

$$\mathbb{E}[XY] = \int\int_{\Re^2} xy\mu^2(dx, dy) = \int_{\Re^1} x\mu_1(dx) \cdot \int_{\Re^1} y\mu_2(dy) = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Through induction, we can show that

$$\mathbb{E}\left[\prod_{j=1}^{n} X_j\right] = \prod_{j=1}^{n} \mathbb{E}[X_j],$$

when $\{X_j, 1 \leq j \leq n\}$ are independent random variables with finite expectations.

The fundamental existence theorem of product measures is as follows. Let a finite or infinite sequence of probability measures $\{\mu_j\}$ on $(\Re^1, \mathscr{B}^1)$ or equivalently their distribution functions be given. There exists a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a sequence of independent random variables $\{X_j\}$ defined on it such that μ_j is the probability measure of X_j for any j. For each n, let $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$ be a probability space in which there exists a random variable X_n with μ_n as its probability measure. The proof utilizes the construction that $\Omega = \times_{n=1}^{+\infty}\Omega_n$, the field made up of the sets $E = \bigcup_{k=1}^{n} E^{(k)}$ where all the $E^{(k)}$'s are disjoint sets of the form $\times_{n=1}^{+\infty} F_n$ with all but a finite of the F_n's equal to the Ω_n's, and the probability measure $\mathbb{P}$ satisfying $\mathbb{P}(\times_{n=1}^{+\infty} F_n) = \prod_{n=1}^{+\infty} \mathbb{P}_n(F_n)$. Another useful result is this: Let $\mathscr{F}_0$ be a field of subsets of an abstract space Ω, and $\mathbb{P}$ a probability measure on $\mathscr{F}_0$. There exists a unique probability measure on the Borel field generated by $\mathscr{F}_0$ that agrees with $\mathbb{P}$ on $\mathscr{F}_0$.

In the product space $X \times Y$, a measurable rectangle is a product $A \times B$ for which $A \in \mathscr{X}$ and $B \in \mathscr{Y}$. The natural class of sets in $X \times Y$ to consider is the σ-field $\mathscr{X} \times \mathscr{Y}$ generated by the measurable rectangles. If $E \in \mathscr{X} \times \mathscr{Y}$, then for each x the set $\{y : (x, y) \in E\}$ lies in $\mathscr{Y}$. If f is measurable $\mathscr{X} \times \mathscr{Y}$, then for each x the function $f(x, \cdot)$ is measurable $\mathscr{Y}$. The key lies in the mapping $T_x : Y \longrightarrow X \times Y$ such that $T_x y = (x, y)$ being measurable $\mathscr{Y}/\mathscr{X} \times \mathscr{Y}$. This in turn relies on the following facts:

(i) If $T^{-1}A' \in \mathscr{F}$ for each $A' \in \mathscr{F}'$ and $\mathscr{A}'$ generates $\mathscr{F}'$, then T is measurable $\mathscr{F}/\mathscr{F}'$.

(ii) If T measurable $\mathscr{F}/\mathscr{F}'$ and T' measurable $\mathscr{F}'/\mathscr{F}''$, $T'T$ will be measurable $\mathscr{F}/\mathscr{F}''$.

For (i), consider the class $\mathscr{C}'$ of sets A' such that $T^{-1}A' \in \mathscr{F}$. The class can be shown to be a σ-field that contains $\mathscr{A}'$; hence, $\mathscr{C}' \supseteq \mathscr{F}'$.

A class $\mathscr{C}$ of subsets of Ω is a π-system if it is closed under the formation of finite intersections. A class $\mathscr{C}$ is a λ-system if it contains Ω and is closed under the formation of complements and of finite and countable disjoint unions. When $\mathscr{C}$ is both a π-system and a λ-system, it will be a σ-field. Many uniqueness arguments depend on Dynkin's π-λ theorem: If $\mathscr{P}$ is a π-system and $\mathscr{L}$ a λ-system, then $\mathscr{P} \subseteq \mathscr{L}$ implies $\sigma(\mathscr{P}) \subseteq \mathscr{L}$. Moreover, suppose

μ_1 and μ_2 are probability measures on $\sigma(\mathscr{P})$, where $\mathscr{P}$ is a π-system. If μ_1 and μ_2 agree on $\mathscr{P}$, then they agree on $\sigma(\mathscr{P})$.

Suppose $(X, \mathscr{X}, \mu)$ and $(Y, \mathscr{Y}, \nu)$ are measure spaces with finite μ and ν. Note $\nu[y : (x, y) \in E]$ is a well defined function of x. If $\mathscr{L}$ is the class of E in $\mathscr{X} \times \mathscr{Y}$ for which the function is measurable $\mathscr{X}$, it is a λ-system. Since the function is $I_A(x)\nu(B)$ for $E = A \times B$, $\mathscr{L}$ contains the π-system consisting of the measurable rectangles. Hence $\mathscr{L}$ coincides with $\mathscr{X} \times \mathscr{Y}$ by the π-λ theorem. Now both $\pi'(E) = \int_X \nu[y : (x, y) \in E]\mu(dx)$ and $\pi''(E) = \int_X \mu[x : (x, y) \in E]\nu(dy)$ constitute finite measures on $\mathscr{X} \times \mathscr{Y}$. For measurable rectangles,

$$\pi'(A \times B) = \pi''(A \times B) = \mu(A) \cdot \nu(B).$$

The class of E in $\mathscr{X} \times \mathscr{Y}$ for which $\pi'(E) = \pi''(E)$ thus contains the measurable rectangles; since this class is a λ-system, it contains $\mathscr{X} \times \mathscr{Y}$. The common value $\pi'(E) = \pi''(E)$ is the product measure sought. Furthermore, if $(X, \mathscr{X}, \mu)$ and $(Y, \mathscr{Y}, \nu)$ are σ-finite measure spaces, there is one and only one σ-finite measure π such that $\pi(A \times B) = \mu(A) \cdot \nu(B)$ for measurable rectangles. We often write π as $\mu \times \nu$.

Fubini's theorem states that, for a positive function that is measurable $\mathscr{X} \times \mathscr{Y}$,

$$\int_{X \times Y} f(x, y) \cdot (\mu \times \nu)(d(x, y)) = \int_X \left[\int_Y f(x, y)\nu(dy)\right] \mu(dx).$$

Each $f(x, \cdot)$ is already known to be measurable $\mathscr{Y}$. The question is whether $\int_Y f(\cdot, y)\nu(dy)$ is measurable $\mathscr{X}$, and whether it integrates to the left-hand side. For a positive f, the proof proceeds with $f = 1_E$, simple function, and then monotone convergence theorem. When f is not necessarily positive, we can again use the decomposition $f = f^+ - f^-$. Here, the measurability statement will be slightly weaker.

5 Convergence Concepts—C4

A sequence of random variables $\{X_n\}$ is said to converge almost everywhere to a random variable X iff there exists a null set Ω^0 such that

$$\lim_{n \longrightarrow +\infty} X_n(\omega) = X(\omega) \text{ finite}, \qquad \forall \omega \in \Omega \setminus \Omega^0.$$

The sequence $\{X_n\}$ converges a.e. to X iff for every $\epsilon > 0$,

$$\lim_{m \longrightarrow +\infty} \mathbb{P}[|X_n - X| \leq \epsilon \text{ for all } n \geq m\} = 1.$$

Let $A_n(\epsilon)$ be $\{\omega \in \Omega : \ |X_n - X| \leq \epsilon\}$. Then almost everywhere convergence means

$$\mathbb{P}\left[\bigcap_{k=1}^{+\infty} \bigcup_{m=1}^{+\infty} \bigcap_{n=m}^{+\infty} A_n\left(\frac{1}{k}\right)\right] = 1.$$

Meanwhile, the condition means that for every $\epsilon > 0$,

$$\lim_{m \longrightarrow +\infty} \mathbb{P}\left[\bigcap_{n=m}^{+\infty} A_n(\epsilon)\right] = 1.$$

The sequence $\{X_n\}$ is said to converge in probability to X iff for every $\epsilon > 0$,

$$\lim_{n \longrightarrow +\infty} \mathbb{P}[A_n(\epsilon)] = \lim_{n \longrightarrow +\infty} \mathbb{P}[|X_n - X| \leq \epsilon] = 1.$$

Convergence a.e. implies convergence in pr. Due to the completeness of the real line $\Re$, we have Cauchy-related results. The sequence $\{X_n\}$ converges a.e. iff for every $\epsilon > 0$,

$$\lim_{m \longrightarrow +\infty} \mathbb{P}[|X_n - X_{n'}| > \epsilon \text{ for some } n' > n \geq m] = 0.$$

The sequence converges in pr. iff for every $\epsilon > 0$,

$$\lim_{n,n' \longrightarrow +\infty} \mathbb{P}[|X_n - X_{n'}| > \epsilon] = 0.$$

Suppose $0 < p < +\infty$. The sequence $\{X_n\}$ in L^p is said to converge to X in L^p iff

$$\lim_{n \longrightarrow +\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

In all these definitions above, X_n converges to X iff $X_n - X$ converges to 0. Hence there is no loss of generality if we put $X \equiv 0$ in the discussion, provided that any hypothesis involved can be similarly reduced to this case. We say that X is dominated by Y if $|X| \leq Y$ a.e., and that the sequence $\{X_n\}$ is dominated by Y iff this is true for each X_n with the same Y. We say that X or $\{X_n\}$ is uniformly bounded iff the Y above may be taken to be a constant. If X_n converges to 0 in L^p, then it converges to 0 in pr. The converse is true provided that $\{X_n\}$ is dominated by some Y that belongs to L^p. As a corollary, for a uniformly bounded sequence $\{X_n\}$ convergence in pr. and in L^p are equivalent.

A general result is as follows. $X_n \longrightarrow 0$ in pr. iff

$$\mathbb{E}\left[\frac{|X_n|}{1 + |X_n|}\right] \longrightarrow 0.$$

Furthermore, the function $\rho(\cdot, \cdot)$ given by

$$\rho(X, Y) = \mathbb{E}\left[\frac{|X - Y|}{1 + |X - Y|}\right]$$

is a metric in the space of random variables provided that we identify random variables that are equal a.e. Convergence in pr. does not imply convergence in L^p: let φ_{kj} be the indicator function of $((j-1)/k, j/k)$ and consider $k^{1/p}\varphi_{kj}$; also, the latter does not imply convergence a.e.: just consider φ_{kj}. Convergence a.e. does not imply convergence in L^p either: let

$$X_n(\omega) = \begin{cases} 2^n, & \text{if } \omega \in (0, 1/n); \\ 0, & \text{otherwise.} \end{cases}$$

The sequence of random variables $\{X_n\}$ in L^1 is said to converge weakly in L^1 to X iff for each bounded random variable Y,

$$\lim_{n \longrightarrow +\infty} \mathbb{E}[X_n Y] = \mathbb{E}[XY], \text{ finite.}$$

If X' is another candidate, by taking $Y = 1_{X \neq X'}$ we can show that X and X' must be equivalent. Clearly, convergence in L^1 implies weak convergence; hence the former is sometimes referred to as strong convergence. However, convergence a.e. still does not imply weak convergence; whereas, weak convergence does not imply convergence in pr.

Let E_n be any sequence of subsets of Ω; we define

$$\limsup_n E_n = \bigcap_{m=1}^{+\infty} \bigcup_{n=m}^{+\infty} E_n, \qquad \liminf_n E_n = \bigcup_{m=1}^{+\infty} \bigcap_{n=m}^{+\infty} E_n.$$

We can see that

$$\liminf_n E_n = (\limsup_n E_n^c)^c.$$

A point belongs to $\limsup_n E_n$ iff it belongs to infinitely many terms of the sequence $\{E_n, n \geq 1\}$. A point belongs to $\liminf_n E_n$ iff it belongs to all terms of the sequence from a certain term on. It follows that both sets are independent of the enumeration of the E_n's. In more intuitive language, the event $\limsup_n E_n$ occurs iff the events E_n occur infinitely often. Thus

$$\mathbb{P}\left[\limsup_n E_n\right] = \mathbb{P}[E_n \text{ i.o.}].$$

Using monotone convergence theorem, we can show

$$\mathbb{P}\left[\limsup_n E_n\right] = \lim_{m \longrightarrow +\infty} \mathbb{P}\left[\bigcup_{n=m}^{+\infty} E_n\right], \qquad \mathbb{P}\left[\liminf_n E_n\right] = \lim_{m \longrightarrow +\infty} \mathbb{P}\left[\bigcap_{n=m}^{+\infty} E_n\right].$$

The convergence part of Borel-Cantelli lemma: For arbitrary events $\{E_n\}$,

$$\sum_n \mathbb{P}[E_n] < +\infty \Longrightarrow \mathbb{P}[E_n \text{ i.o.}] = 0.$$

The divergence part of Borel-cantelli lemma: If the events $\{E_n\}$ are independent,

$$\sum_n \mathbb{P}[E_n] = +\infty \Longrightarrow \mathbb{P}[E_n \text{ i.o.}] = 1.$$

The proof relies on showing $\mathbb{P}[\bigcap_{n=m}^{+\infty} E_n^c] = 0$ using the inequality $1 - x \leq e^{-x}$. The above would be true even if the events $\{E_n\}$ are merely pairwise independent. A corollary is that, if the events $\{E_n\}$ are pairwise independent, then

$$\mathbb{P}[\limsup_n E_n] = 0 \text{ or } 1$$

according as $\sum_n \mathbb{P}[E_n] < +\infty$ or $= +\infty$.

The sequence of random variables $X_n \longrightarrow 0$ a.e. iff

$$\mathbb{P}[|X_n| > \epsilon \text{ i.o.}] = 0, \qquad \forall \epsilon > 0.$$

If $X_n \longrightarrow X$ in pr., then there exists a subsequence $\{n_k\}$ such that $X_{n_k} \longrightarrow X$ a.e. That is, convergence in pr. implies convergence in a.e. along a subsequence. We can identify n_k so that $\mathbb{P}[|X_{n_k} - X| > 1/2^k] \leq 1/2^k$. This will lead to $\mathbb{P}[|X_{n_k} - X| > 1/2^k \text{ i.o.}] = 0$.

6 Laws of Large Numbers—C5

The law of large numbers has to do with partial sums $S_n = \sum_{j=1}^n X_j$. The weak or the strong law of large numbers is said to hold for the sequence according as

$$\frac{S_n - \mathbb{E}[S_n]}{n} \longrightarrow 0$$

in pr. or a.e. A natural generalization is

$$\frac{S_n - a_n}{b_n} \longrightarrow 0,$$

where $\{a_n\}$ is a sequence of real numbers and $\{b_n\}$ a sequence of positive numbers tending to infinity. We have used Chebyshev's inequality to show that a sequence of random variables Z_n would satisfy $Z_n \longrightarrow 0$ in pr. when $\mathbb{E}[Z_n^2] \longrightarrow 0$. When applied to $Z_n = S_n/n$, we get

$$\mathbb{E}[S_n^2] = o(n^2) \Longrightarrow \frac{S_n}{n} \longrightarrow 0 \text{ pr.}$$

Now we can calculate $\mathbb{E}[S_n^2]$ more explicitly as

$$\mathbb{E}[S_n^2] = \mathbb{E}\left[\left(\sum_{j=1}^n X_j\right)^2\right] = \sum_{j=1}^n \mathbb{E}[X_j^2] + 2 \sum_{1 \leq j < k \leq n} \mathbb{E}[X_j X_k].$$

So that $\mathbb{E}[S_n^2] = o(n^2)$ can be true, we need to introduce certain assumptions to cause enough cancellation among the "mixed terms" above.

Two random variables are said to be uncorrelated iff both have finite second moments and $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. They are said to be orthogonal iff $\mathbb{E}[XY] = 0$. The random variables of any family are said to be uncorrelated [orthogonal] iff every two of them are. Note that pairwise independence implies uncorrelatedness, provided second moments are finite. If $\{X_n\}$ is a sequence of uncorrelated random variables, then the sequence $\{X_n - \mathbb{E}[X_n]\}$ is orthogonal and hence the "additivity of variance" relationship:

$$\sigma^2(S_n) = \sum_{j=1}^{n} \sigma^2(X_j).$$

Because now $\sigma^2(S_n) = O(n) = o(n^2)$ and hence $\sigma^2(S_n/n) = o(1)$, we have the following: If the X_j's are uncorrelated and their second moments have a common bound, then

$$\frac{S_n - \mathbb{E}[S_n]}{n} \longrightarrow 0$$

will happen in L^2 and hence also in pr. Under the same hypotheses, the convergence can actually be shown to hold a.e. This can be achieved through the "method of subsequences". After showing the convergence of S_{n^2}/n^2 to 0 a.e., one can attempt to show that $|S_k - S_{n^2}|$ is small when $n^2 \leq k < (n+1)^2$.

The most celebrated, as well as the very first case of the strong law of large numbers, due to Borel (1909), is formulated in terms of the so-called normal numbers. Let each real number in $[0, 1]$ be expanded in the usual decimal system:

$$\omega = .x_1 x_2 ... x_n ....$$

Except for the countable set of terminating decimals, for which there are two distinct expansions, this representation is unique. Fix a $k : 0 \leq k \leq 9$, and let $v_k^{(n)}(\omega)$ denote the number of digits among the first n of ω that are equal to k. Then $v_k^{(n)}(\omega)/n$ is the relative frequency of the digit k in the first n places, and the limit, if existing:

$$\lim_{n \longrightarrow +\infty} \frac{v_k^{(n)}(\omega)}{n} = \varphi_k(\omega),$$

may be called the frequency of k in ω. The number ω is called simply normal iff this limit exists for each k and is equal to $1/10$. Except for a Borel set of measure zero, every number in $[0, 1]$ is simply normal. For $\omega = .x_1 x_2 ...$ and a fixed k, we can define $X_n = 1_{x_n(\omega)=k}$. Note $\mathbb{E}[X_n] = 1/10$, $\mathbb{E}[X_n^2] = 1/10$, and the X_n's are uncorrelated. Meanwhile, $\sum_{j=1}^{n} X_j(\omega)/n$ is the relative frequency of the digit k in the first n places of the decimal for ω.

The law of large numbers involves only the first moment, but so far we have operated with the second. In order to drop any assumption on the second moment, we need a new device, that of “equivalent sequences”, due to Khintchine (1894-1959). Two sequences of random variables $\{X_n\}$ and $\{Y_n\}$ are said to be equivalent iff

$$\sum_n \mathbb{P}[X_n \neq Y_n] < +\infty.$$

If $\{X_n\}$ and $\{Y_n\}$ are equivalent, then by Borel-Cantelli lemma,

$$\sum_n (X_n - Y_n) \quad \text{converges a.e.}$$

Indeed, except for a null set, $X_n(\omega)$ and $Y_n(\omega)$ will differ only in a finite number of terms. Furthermore if $a_n \uparrow +\infty$, then

$$\frac{1}{a_n}\sum_{j=1}^{n}(X_j - Y_j) \longrightarrow 0 \quad \text{a.e.}$$

With probability one, the expression $\sum_n X_n$ or $\sum_{j=1}^n X_j/a_n$ converges, diverges to $+\infty$ or $-\infty$, or fluctuates in the same way as $\sum_n Y_n$ or $\sum_{j=1}^n Y_j/a_n$. The next law of large numbers is due to Khintchine. Let $\{X_n\}$ be a pairwise independent and identically distributed random variables with finite mean m. Then,

$$\frac{S_n}{n} \longrightarrow m \quad \text{in pr.}$$

Use truncating by defining $Y_n(\omega) = X_n(\omega)$ or 0 according as $|X_n(\omega)| \leq n$ or not. $\{X_n\}$ and $\{Y_n\}$ are equivalent. For $T_n = \sum_{j=1}^n Y_j$, our job is to prove $T_n/n \longrightarrow m$ in pr.

A strong version due to Kolmogorov does not involve the second moments. Let $\{X_n\}$ be a sequence of independent and identically distributed r.v.'s. It simply states that

$$\mathbb{E}[|X_1|] < +\infty \Longrightarrow \frac{S_n}{n} \longrightarrow \mathbb{E}[X_1] \quad \text{a.e.},$$

$$\mathbb{E}[|X_1|] = +\infty \Longrightarrow \limsup_{n \longrightarrow +\infty} \frac{|S_n|}{n} = +\infty \quad \text{a.e.}$$

There is a reason why convergence or non-convergence would come in an a.e.-fashion when the X_i's are independent. Let $\{X_n\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. For any n, we can define two B.F.'s:

$\mathscr{F}_n$ = the smallest B.F. containing $X_k^{-1}(\mathscr{B}^1)$ for every $k = 1, 2, ..., n$ and all null sets;

$\mathscr{F}'_n$ = the smallest B.F. containing $X_k^{-1}(\mathscr{B}^1)$ for every $k = n+1, n+2, ...$ and null sets.

The union $\bigcup_{n=1}^{+\infty} \mathscr{F}_n$ is a field but not necessarily a B.F. The smallest B.F. containing it, or equivalently, containing every $\mathscr{F}_n$, is denoted by $\mathscr{F}_\infty \equiv \bigvee_{n=1}^{+\infty} \mathscr{F}_n$. On the other hand, the intersection $\bigcap_{n=1}^{+\infty} \mathscr{F}'_n$ is a B.F. denoted also by $\bigwedge_{n=1}^{+\infty} \mathscr{F}'_n$. It will be called the remote field and a member of it a remote event. The latter could be $\sum_{j=1}^{n} X_j/n \longrightarrow m$.

Suppose $\mathscr{F} = \mathscr{F}_\infty$. Then, Kolmogorov's zero-or-one law states that, when the X_n's are independent, each remote event has probability zero or one. Let $\Lambda \in \bigcap_{n=1}^{+\infty} \mathscr{F}'_n$ with $\mathbb{P}[\Lambda] > 0$. Since $\mathscr{F}_n$ and $\mathscr{F}'_n$ are independent fields, Λ is independent of every set M in any $\mathscr{F}_n$. Hence,

$$\mathbb{P}[\Lambda \cap M] = \mathbb{P}[\Lambda]\mathbb{P}[M].$$

Define $\mathbb{P}_\Lambda$ so that $\mathbb{P}_\Lambda(M) = \mathbb{P}[\Lambda \cap M]/\mathbb{P}[\Lambda]$. It is a probability distribution that coincides with $\mathbb{P}$ on $\bigvee_{n=1}^{+\infty} \mathscr{F}_n$. Since $\mathscr{F} = \mathscr{F}_\infty$, an earlier result (Theorem 2.2.3 of Chung 2001) would lead us to $\mathbb{P}_\Lambda = \mathbb{P}$. So even for Λ itself,

$$\mathbb{P}_\Lambda[\Lambda] \equiv \frac{\mathbb{P}[\Lambda \cap \Lambda]}{\mathbb{P}[\Lambda]} = \mathbb{P}[\Lambda].$$

The only possibility is that $\mathbb{P}[\Lambda] = 1$.

7 Applications of Convergence Results—C5

The law of large numbers have numerous applications. Let $\{X_n\}$ be a sequence of independent and identically distributed random variables with a common distribution function F. For each $\omega \in \Omega$, let $F_n(x, \omega)$ be the proportion of the $X_j(\omega)$'s for $j = 1, ..., n$ that are below x. The function $F_n(\cdot, \omega)$ is the empirical distribution based on n samples from F. For each x, $F_n(x, \cdot)$ is a random variable. Note that

$$F_n(x, \omega) = \frac{1}{n} \sum_{j=1}^{n} 1_{X_j(\omega) \leq x}.$$

For each x, the sequence $\{1_{X_j \leq x}\}$ is totally independent since $\{X_j\}$ is. They have the common Bernoulli distribution with $\mathbb{P}[1_{X_j \leq x} = 1] = \mathbb{P}[X_j \leq x] = F(x)$. Thus, $\mathbb{P}[1_{X_j \leq x}] = F(x)$. The strong law of numbers would lead to

$$F_n(x, \omega) \longrightarrow F(x) \quad \text{a.e.}$$

How much better it would be to make a global statement about the functions $F_n(\cdot, \omega)$ and $F(\cdot)$. Already for each x, there exists a null set $N(x)$ such that the above convergence holds for $\omega \in \Omega \setminus N(x)$. So the convergence would hold simultaneously for all x in any given countable set Q, such as the set of rational numbers. This can be further strengthened to

$$\sup_{-\infty < x < +\infty} |F_n(x, \omega) - F(x)| \longrightarrow 0 \quad \text{a.e.}$$

Let J be the countable set of jumps of F. For each $x \in J$, note that

$$F_n(x,\omega) - F_n(x-,\omega) = \frac{1}{n}\sum_{j=1}^{n} 1_{X_j(\omega)=x}.$$

Again by the law of large numbers,

$$F_n(x,\omega) - F_n(x-,\omega) \longrightarrow F(x) - F(x-) \quad \text{a.e.}$$

The uniform convergence of F_n to F will follow from a technical lemma.

The next application is to renewal theory. Let $\{X_n\}$ again be a sequence of independent and identically distributed random variables. We shall further assume that they are positive and not identically zero, a.e. It follows that the common mean is strictly positive but may be $+\infty$. Now the successive random variables are interpreted as "lifespans" of certain objects undergoing a process of renewal, or the "return periods" of certain recurrent phenomena. This raises questions such as: given an epoch in time, how many renewals have there been before it? how long ago was the last renewal? how soon will the next be?

Let $N(t,\omega)$ be the number of renewals up to and including time t. Note

$$N(t,\omega) = n \Longleftrightarrow S_n(\omega) \leq t < S_{n+1}(\omega).$$

Equivalently, we have

$$\{\omega : N(t,\omega) \leq m\} = \{\omega : S_m(\omega) \geq t\}.$$

The family of random variables $\{N(t)\}$ indexed by $t \in [0,+\infty)$ may be called a renewal process. If the common distribution function F of the X_n's is the exponential $F(x) = 1-e^{-\lambda x}$, then $\{N(t)\}$ is just the simple Poisson process. We can use contradiction to prove that

$$\lim_{t\longrightarrow+\infty} N(t) = +\infty \quad \text{a.e.,}$$

namely that the total number of renewals becomes infinite with time. Write $m \equiv \mathbb{E}[X_1] > 0$. Suppose for the moment that it is finite. According to the strong law of large numbers, $S_n/n \longrightarrow m$ a.e. Therefore,

$$\lim_{t\longrightarrow+\infty} \frac{S_{N(t,\omega)}(\omega)}{N(t,\omega)} = m \quad \text{a.e.}$$

We can then confirm that

$$\lim_{t\longrightarrow+\infty} \frac{N(t)}{t} = \frac{1}{m} \quad \text{a.e.}$$

and $\lim_{t\longrightarrow+\infty}\mathbb{E}[N(t)]/t = 1/m$. Even when $m = +\infty$, the results last if we treat $1/m$ as 0. The first convergence comes from the observation that

$$\frac{S_{N(t,\omega)}(\omega)}{N(t,\omega)} \leq \frac{t}{N(t,\omega)} \leq \frac{S_{N(t,\omega)+1}(\omega)}{N(t,\omega)+1}\frac{N(t,\omega)+1}{N(t,\omega)}.$$

The second convergence is not as easy as might have been thought. By introducing a Bernoulli random variable X'_n that is below X_n, we can show that $\mathbb{E}[(N(t)/t)^2] \leq \mathbb{E}[(N'(t)/t)^2] = O(1)$. For each t, we can also show $\mathbb{E}[N(t)] < +\infty$.

One should expect $\mathbb{E}[S_{N(t)}]$ to be near $m\mathbb{E}[N(t)]$ when t is large. The exact statement is

$$\mathbb{E}[X_1 + \cdots + X_{N(t)}] = \mathbb{E}[X_1] \cdot \mathbb{E}[N(t)].$$

This is a striking generalization of the additivity of expectations when the number of terms and the summands are both random. This follows from the more general Wald's equation. Let $\{X_n\}$ be a sequence of independent and identically distributed random variables with finite mean. Let $\mathscr{F}_k$ be the Borel field generated by $\{X_j, 1 \leq j \leq k\}$. Suppose a positive integer random variable N satisfies

$$\{N \leq k\} \in \mathscr{F}_k,$$

meaning that it is a stopping time and $\mathbb{E}[N] < +\infty$, then

$$\mathbb{E}[S_N] = \mathbb{E}[N] \cdot \mathbb{E}[X_1].$$

Since $S_0 = 0$ as usual, we have

$$\mathbb{E}[S_N] = \mathbb{E}\left[\sum_{j=1}^{N} X_j\right] = \sum_{j=1}^{+\infty}\mathbb{E}[X_j 1_{N\geq j}] = \sum_{j=1}^{+\infty}\left\{\mathbb{E}[X_j] - \mathbb{E}[X_j 1_{N\leq j-1}]\right\}.$$

Since $1_{N\leq j-1}$ and X_j are independent, the last term is equal to $\mathbb{E}[X_j]\mathbb{P}[N \leq j-1]$. Thus,

$$\mathbb{E}[S_N] = \sum_{j=1}^{+\infty}\mathbb{E}[X_j]\mathbb{P}[N \geq j] = \mathbb{E}[X_1]\sum_{j=1}^{+\infty}\mathbb{P}[N \geq j] = \mathbb{E}[N] \cdot \mathbb{E}[X_1].$$

This Wald's equation would lead to $\mathbb{E}[S_{N(t)+1}] = \mathbb{E}[X_1]\mathbb{E}[N(t)+1]$. Note $N(t) \leq k-1$ is equivalent to $t < S_k$ and hence whether $\{N(t)+1 \leq k\}$ only depends on $X_1, ..., X_k$.

Another example is a noted triumph of the ideas of probability theory applied to classical analysis. It is S. Bernstein's proof of Weierstrass' theorem on the approximation of continuous functions by polynomials. Let f be a continuous function on $[0,1]$, and define the Bernstein polynomial $\{p_n\}$ as follows:

$$p_n(x) = \sum_{k=0}^{n}\frac{n!}{k!(n-k)!}x^k(1-x)^{n-k} \cdot f\left(\frac{k}{n}\right).$$

Then p_n converges uniformly to f in $[0,1]$. For each x, consider a sequence of Bernoulli random variables $\{X_n(x)\}$ with success probability x. For $S_n(x) = \sum_{k=1}^n X_n(x)$, we have $p_n(x) = \mathbb{E}[f(S_n(x)/n)]$. The law of large numbers has $S_n/n \longrightarrow x$ a.e. and hence in pr. So we have $\mathbb{E}[f(S_n(x)/n)] \longrightarrow f(x)$ at all individual x's. The uniformity of the convergences requires some further work.

8 Weak Convergence—B25

Distribution functions F_n are said to converge weakly to a distribution function F if

$$\lim_n F_n(x) = F(x),$$

for every continuity point x of F; this is expressed by writing $F_n \Longrightarrow F$. If μ_n and μ are probability measures on $(R^1, \mathscr{B}^1)$ corresponding to F_n and F, then $F_n \Longrightarrow F$ iff

$$\lim_n \mu_n(A) = \mu(A),$$

for every A of the form $A = (-\infty, x]$ for which $\mu(\{x\}) = 0$. In this case the distributions themselves are said to converge weakly, which is expressed by writing $\mu_n \Longrightarrow \mu$. Thus $F_n \Longrightarrow F$ and $\mu_n \Longrightarrow \mu$ are only different expressions of the same fact.

Let F_n be the distribution function corresponding to a unit mass at n: $F_n = 1_{[n,+\infty)}$. Then $\lim_n F_n(x) = 0$ for every x, so the limit function F will have to satisfy $F(x) = 0$ everywhere. But $F_n \Longrightarrow F$ does not hold, because F is not even a distribution function.

Let μ_n be the binomial distribution with $p = \lambda/n$. For $k = 0, 1, \ldots$ and $n \geq k$,

$$\begin{aligned}\mu_n(\{k\}) &= [n!/(k!(n-k)!)] \cdot (\lambda/n)^k \cdot (1-\lambda/n)^{n-k} \\ &= [\lambda^k (1-\lambda/n)^n / k!] \times \{[1/(1-\lambda/n)^k] \cdot \textstyle\prod_{i=0}^{k-1}(1 - i/n)\}.\end{aligned}$$

As $n \longrightarrow +\infty$, the first factor goes to $e^{-k} \cdot \lambda^k/k!$ while second factor on the right goes to 1 for fixed k. So $\mu_n(\{k\}) \longrightarrow \mu(\{k\})$ where μ is the Poisson distribution with parameter λ. By the series form of Scheffe's theorem (If $\sum_m x_{nm} = \sum_m x_m < +\infty$, the terms being positive, and if $\lim_n x_{nm} = x_m$ for each m, then $\lim_n \sum_m |x_{nm} - x_m| = 0$. If y_m is bounded, then $\lim_n y_m x_{nm} = \sum_m y_m x_m$), the convergence holds for every set of natural numbers. Since these numbers support μ_n and μ, we do have weak convergence.

Let μ_n correspond to a mass of $1/n$ at each of $0/n, 1/n, \ldots, (n-1)/n$; let μ be the Lebesgue measure confined to the unit interval. We have

$$F_n(x) = \frac{\lfloor nx \rfloor + 1}{n} \longrightarrow F(x) = x, \qquad \forall 0 \leq x < 1,$$

and so $F_n \Longrightarrow F$. However, $\mu_n(A) \longrightarrow \mu(A)$ does not hold for every Borel set A: If it is the set of rationals, $\mu_n(A) = 1$ does not converge to $\mu(A) = 0$. Still, $\mu_n \Longrightarrow \mu$.

If μ_n is a unit mass at x_n and μ is a unit mass at x, then $\mu_n \Longrightarrow \mu$ iff $x_n \longrightarrow x$. If $x_n > x+\epsilon$ for infinitely many n, then at the continuity point $x+\epsilon$ there is no $F_n(x+\epsilon) \longrightarrow F(x+\epsilon) = 1$.

Let X_n and X be random variables with distribution functions F_n and F. If $F_n \Longrightarrow F$, then X_n is said to converge in distribution to X, written $X_n \Longrightarrow X$. Note $X_n \Longrightarrow X$ iff

$$\lim_n \mathbb{P}_n[X_n \leq x] = \mathbb{P}[X \leq x],$$

for every x such that $\mathbb{P}[X = x] = 0$. Yes, the random variables X_n can be defined on entirely different probability spaces. Let Ω_n be the space of n-tuples of 0's and 1's, let $\mathscr{F}_n$ consist of all subsets of Ω_n, and let $\mathbb{P}_n$ assign probability $(\lambda/n)^k(1-\lambda/n)^{n-k}$ to each ω consisting of k 1's and $n-k$ 0's (there are $n!/(k! \cdot (n-k)!)$ of them). Let $X_n(\omega)$ be the number of 1's in ω; then X_n, a random variable on $(\Omega_n, \mathscr{F}_n, \mathbb{P}_n)$, represents the number of successes in n Bernoulli trials having probability λ/n of success at each. Let X be a random variable on some $(\Omega, \mathscr{F}, \mathbb{P})$ having the Poisson distribution with parameter λ. Then $X_n \Longrightarrow X$.

Now suppose all random variables live on the same probability space. Note that

$$\mathbb{P}[X \leq x-\epsilon] - \mathbb{P}[|X_n - X| \geq \epsilon] \leq \mathbb{P}[X_n \leq x] \leq \mathbb{P}[X \leq x+\epsilon] + \mathbb{P}[|X_n - x| \geq \epsilon].$$

From this, we get that $X_n \longrightarrow X$ p.r. will lead to $X_n \Longrightarrow X$. The converse is not true. If X and Y are independent and assume the values 0 and 1 with probability 1/2 each, and if $X_n = Y$, then $X_n \Longrightarrow X$ but X_n cannot converge to X in probability because $\mathbb{P}[|X - Y| = 1] = 1/2$. The situation will flip if X is a constant a. From $X_n \Longrightarrow a$,

$$\mathbb{P}[|X_n - a| \geq \epsilon] \leq \mathbb{P}[X_n \leq a-\epsilon] + 1 - \mathbb{P}[X_n \leq a+\epsilon],$$

but $X_n \Longrightarrow a$ demands that $\mathbb{P}[X_n \leq a-\epsilon] \longrightarrow 0$ and $\mathbb{P}[X_n \leq a+\epsilon] \longrightarrow 1$.

The asymptotic properties of a random variable should remain unaffected if it is altered by the addition of a random variable that goes to 0 in probability. If $X_n \Longrightarrow X$ and $X_n - Y_n \Longrightarrow 0$, then $Y_n \Longrightarrow X$. Suppose $y' < x < y''$ and $\mathbb{P}[X = y'] = P[X = y''] = 0$. If $y' < x-\epsilon < x < x+\epsilon < y''$, then

$$\mathbb{P}[X_n \leq y'] - \mathbb{P}[|X_n - Y_n| \geq \epsilon] \leq \mathbb{P}[Y_n \leq x] \leq \mathbb{P}[X_n \leq y''] + \mathbb{P}[|X_n - Y_n| \geq \epsilon].$$

This can be used to show

$$\mathbb{P}[X \leq y'] \leq \liminf_n \mathbb{P}[Y_n \leq x] \leq \limsup_n \mathbb{P}[Y_n \leq x] \leq \mathbb{P}[X \leq y''].$$

If $F_n \Longrightarrow F$ and $F_n \Longrightarrow G$, then $F = G$. Note F and G agree at their common points of continuity, hence at all but countably many points, and hence by right continuity at all points. Also, if $\lim_n F_n(d) = F(d)$ for d in a set D dense in $\Re^1$, then $F_n \Longrightarrow F$. Indeed, if F is continuous at x, there are D points d' and d'' such that $d' < x < d''$ and $F(d'') - F(d') < \epsilon$, and it follows that the limits superior and inferior of $F_n(x)$ are within ϵ of $F(x)$.

Skorohod's theorem makes possible very simple and transparent proofs of many important results. Suppose that μ_n and μ are probability measures on $(\Re^1, \mathscr{B}^1)$ and $\mu_n \Longrightarrow \mu$. There exist random variables Y_n and Y on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$ such that Y_n has distribution μ_n, Y has distribution μ, and $Y_n(\omega) \longrightarrow Y(\omega)$ for each ω. In the proof, take $\Omega = (0,1)$, let $\mathscr{F}$ consist of the Borel subsets of $(0,1)$, and take the Lebesgue measure for $\mathbb{P}$. Consider the distribution functions F_n and F corresponding to μ_n and μ. For $0 < \omega < 1$, put $Y_n(\omega) = \inf\{x : \omega \leq F_n(x)\}$ and $Y(\omega) = \inf\{x : \omega \leq F(x)\}$. Since F is increasing, $\{x : \omega \leq F(x)\}$ is an interval stretching to $+\infty$. Since F is right-continuous, this interval is closed on the left. Therefore, $\{x : \omega \leq F(x)\} = [Y(\omega), +\infty)$, and

$$Y(\omega) \leq y \quad \text{if and only if} \quad \omega \leq F(y).$$

So $\mathbb{P}[Y(\omega) \leq y] = \mathbb{P}[\omega \leq F(y)] = F(y)$. Similarly, $\mathbb{P}[Y_n(\omega) \leq y] = F_n(y)$. To show $Y_n(\omega) \longrightarrow Y(\omega)$, we rely on the facts that they are essentially inverses of F_n and F and that $F_n \Longrightarrow F$. This can be achieved at continuity points of Y.

The mapping theorem is very useful. Suppose that $h : \Re^1 \longrightarrow \Re^1$ is measurable and that the set D_h of its discontinuities is measurable. If $\mu_n \Longrightarrow \mu$ and $\mu(D_h) = 0$, then $\mu \circ h^{-1} \Longrightarrow \mu \circ h^{-1}$. Consider random variables Y_n and Y as constructed above. Since $Y_n(\omega) \longrightarrow Y(\omega)$, if $Y(\omega) \notin D_h$, then $h(Y_n(\omega)) \longrightarrow h(Y(\omega))$. Now $h(Y_n) \longrightarrow h(Y)$ a.e. By an earlier result, $h(Y_n) \Longrightarrow h(Y)$. But $h(Y_n)$ has the distribution $\mu_n \circ h^{-1}$ and $h(Y)$ has the distribution $\mu \circ h^{-1}$. In the language of random variables, if $X_n \Longrightarrow X$ and $\mathbb{P}[X \in D_h] = 0$, then $h(X_n) \Longrightarrow h(X)$. Take $X = a$. If $X \Longrightarrow a$ and h is continuous at a, then $h(X_n) \Longrightarrow h(a)$.

The boundary ∂A of A consists of the points that are limits of sequences in A and are also limits of sequences of A^c; alternatively, ∂A is the closure of A minus its interior. A set A is a μ-continuity set if it is a Borel set and $\mu(\partial A) = 0$. The following are equivalent:

(i) $\mu_n \Longrightarrow \mu$;

(ii) $\int f d\mu_n \longrightarrow \int f d\mu$ for every bounded and continuous function f;

(iii) $\mu_n(A) \longrightarrow \mu(A)$ for every μ-continuity set A.

(iv) $\limsup_n \mu_n(C) \leq \mu(C)$ for any closed set C.

(v) $\liminf_n \mu_n(G) \geq \mu(G)$ for any open set G.

Use Skorohod, mapping, and bounded convergence theorems to prove (i) to (ii). For $f = 1_A$, we have $D_f = \partial A$. This helps (ii) to (iii). From $\partial(-\infty, +\infty] = \{x\}$, we can get (iii) to (i). For (ii) to (i), consider $x < y$ and f so that $f(t) = 1$ when $t \leq x$, $f(t) = 0$ when $t \geq y$, and $f(t) = (y-t)/(y-x)$ for $x \leq t \leq y$. We can get $F_n(x) \leq \int f d\mu_n$ and yet $\int f d\mu \leq F(y)$. The concept of weak convergence would be nearly useless if $\mu_n(A) \longrightarrow \mu(A)$ were not allowed to fail when $\mu(\partial A) > 0$. Since $F(x) - F(x-) = \mu(\{x\}) = \mu(\partial(-\infty, x])$, it is natural in the original definition to allow $F_n(x) \longrightarrow F(x)$ to fail when x is not a continuity point of F.

We can also show that (ii) implies (iv), which is equivalent to (v). Let C be a closed set in $\Re^1$. The distance $\text{dist}(x, C) = \inf\{|x - y| : y \in C\}$ from x to C is continuous in x. Let $\varphi_j(t) = 1$ when $t \leq 0$, $1 - jt$ when $0 \leq t \leq 1/j$, and 0 when $t \geq 1/j$. Then $f_j(x) = \varphi_j(\text{dist}(x, C))$ is continuous and bounded by 1. Also, $f_j(x) \downarrow 1_C(x)$ due to C's closedness. With (ii),

$$\limsup_n \mu_n(C) \leq \lim_n \int f_j d\mu_n = \int f_j d\mu.$$

As $j \longrightarrow +\infty$, $\int f_j d\mu \downarrow \int 1_C d\mu = \mu(C)$. To see that (iv) and (v) together imply (iii), note

$$\mu(\text{int}A) \leq \liminf_n \mu_n(\text{int}A) \leq \liminf_n \mu_n(A) \leq \limsup_n \mu_n(A) \leq \limsup_n \mu_n(\text{cl}A) \leq \mu(\text{cl}A).$$

One of the most frequently used results in analysis is the Helly selection theorem. For every sequence $\{F_n\}$ of distribution functions there exists a subsequence $\{F_{n_k}\}$ and an increasing, right-continuous function F such that $\lim_k F_{n_k}(x) = F(x)$ at continuity points x of F. The diagonal method will yield a sequence $\{n_k\}$ along which the limit $G(r) = \lim_k F_{n_k}(r)$ exists for all rationals r. Define $F(x) = \inf\{G(r) : x < r\}$. F can be shown to be a distribution function for a subprobability measure. It is important to have a condition which ensures that for some subsequence the limit F is a distribution function.

A sequence of probability measures μ_n on $(\Re^1, \mathscr{B}^1)$ is said to be tight if for each $\epsilon > 0$ there exists a finite interval $(a, b]$ such that $\mu_n(a, b] > 1 - \epsilon$ for all n. In terms of the corresponding distribution functions F_n, the condition is that for each $\epsilon > 0$ there exist x and y such that $F_n(x) < \epsilon$ and $F_n(y) > 1 - \epsilon$ for all n. If μ_n is a unit mass at n, $\{\mu_n\}$ is not tight in this sense—the mass of μ_n "escapes to infinity".

Tightness is a condition preventing this escape of mass. It is a necessary and sufficient condition that for every subsequence $\{\mu_{n_k}\}$ there exist a further subsequence $\{\mu_{n_{k(j)}}\}$ and a probability measure μ such that $\mu_{n_{k(j)}} \Longrightarrow \mu$ as $j \longrightarrow +\infty$. For sufficiency, we can identify $(a, b]$ with a and b both being continuity points of F for any $\epsilon > 0$ and ensure that $\mu(a, b] > 1 - \epsilon$. For necessity, we use contraposition. The existence of $\mu_{n_k}(-k, k] \leq 1 - \epsilon$ makes it impossible that $\mu_{n_{k(j)}}$ converges to some μ and yet $\mu(a, b] > 1-\epsilon$ for some continuity points a and b. As a corollary, if $\{\mu_n\}$ is a tight sequence of probability measures, and if each subsequence that converges weakly at all converges weakly to the same μ, then $\mu_n \Longrightarrow \mu$.

If μ_n is a unit mass at x_n, then $\{\mu_n\}$ is tight iff $\{x_n\}$ is bounded. The above results reduce in this case to standard facts about the real line. Tightness of sequences of probability measures is analogous to boundedness of sequences of real numbers. For normal distributions μ_n with mean m_n and standard deviation σ_n, the sequence will be tight iff the means and variances are bounded. When applying Skorohod's representation theorem and Fatou's lemma, we can show that if $X_n \Longrightarrow X$ then $\mathbb{E}[|X|] \leq \liminf_n \mathbb{E}[|X_n|]$.

The random variables X_n are said to be uniformly integrable if

$$\lim_{\alpha \longrightarrow +\infty} \sup_n \mathbb{E}[X_n \cdot 1_{\{|X_n| \geq \alpha\}}] = 0.$$

When α is large enough, every $\mathbb{E}[|X_n| \cdot 1_{\{|X_n| \geq \alpha\}}]$ would be less than 1. So $\mathbb{E}[|X_n|] \leq \alpha + 1$. If $X_n \Longrightarrow X$ and the X_n's are uniformly integrable, then X is integrable and $\mathbb{E}[X_n] \longrightarrow \mathbb{E}[X]$. To prove, we can construct the Y_n's and Y on a common probability space using Skorohod's theorem. By Fatou's lemma, Y is integrable. Define $Y_n^{(\alpha)} = Y_n$ or 0 depending on whether $|Y_n| < \alpha$. Do the same for Y. If $\mathbb{P}[|Y| = \alpha] = 0$, then $Y_n^{(\alpha)}$ would converge to $Y^{(\alpha)}$ a.e. By the bounded convergence theorem,

$$\mathbb{E}[Y_n^{(\alpha)}] \longrightarrow \mathbb{E}[Y^{(\alpha)}].$$

Since $\mathbb{E}[Y_n] - \mathbb{E}[Y_n^{(\alpha)}] = \mathbb{E}[Y_n \cdot 1_{\{|Y| \geq \alpha\}}]$ and $\mathbb{E}[Y] - \mathbb{E}[Y^{(\alpha)}] = \mathbb{E}[Y \cdot 1_{\{|Y| \geq \alpha\}}]$,

$$\limsup_n |\mathbb{E}[Y_n] - \mathbb{E}[Y]| \leq \sup_n \mathbb{E}[|Y_n| \cdot 1_{\{|Y_n| \geq \alpha\}}] + \mathbb{E}[|Y| \cdot 1_{|Y| \geq \alpha}].$$

The rest follows from uniform integrality and that $\mathbb{P}[|Y| = \alpha] = 0$ for all but countable α's.

9 Characteristic Functions—C6

For any random variable X with the probability measure μ and distribution function F, its characteristic function f on $\Re^1$ is defined in the followsing:

$$f(t) = \mathbb{E}[e^{itX}] = \int_\Omega e^{itX(\omega)} \mathbb{P}(d\omega) = \int_{\Re^1} e^{itx} \mu(dx) = \int_{-\infty}^{+\infty} e^{itx} dF(x).$$

Its real and imaginary parts are, respectively,

$$\mathbf{R}f(t) = \int \cos(xt)\mu(dx), \qquad \mathbf{I}f(t) = \int \sin(xt)\mu(dx).$$

Note that $|f(t)| \leq 1 = f(0)$ and $f(-t) = \overline{f(t)}$ where $\overline{z}$ denotes the complex conjugate of z. Also, f is uniformly continuous in $\Re^1$. Since $f(t+h) - f(t) = \int (e^{i(t+h)x} - e^{itx})\mu(dx)$,

$$|f(t+h) - f(t)| \leq \int |e^{itx}||e^{ihx} - 1|\mu(dx) = \int |e^{ihx} - 1|\mu(dx).$$

The last integrand is bounded by 2 and tends to 0 as $h \longrightarrow 0$, for each x. Hence, the integral converges to 0 by bounded convergence. Since it does not involve t, the convergence is uniform. In addition, if we write f_X for the characteristic function of X, then

$$f_{aX+b}(t) = f_X(at)e^{itb}, \qquad\qquad f_{-X}(t) = f_X(-t) = \overline{f_X(t)}.$$

To see these, note that

$$f_{aX+b}(t) = \mathbb{E}[e^{it(aX+b)}] = \mathbb{E}[e^{i(at)X}]e^{itb}.$$

A convex combination $\sum_{n=1}^{+\infty} \lambda_n f_n$ of characteristic functions remains one: it is simply the characteristic function corresponding to $\sum_{n=1}^{+\infty} \lambda_n \mu_n$. Moreover, $\prod_{n=1}^{n} f_n$ is a chacteristic function. There exist independent random variables X_j with probability distributions μ_j. Letting $S_n = \sum_{j=1}^{n} X_j$, we have

$$f_{S_n}(t) = \mathbb{E}[e^{itS_n}] = \mathbb{E}[e^{it\sum_{j=1}^{n} X_j}] = \mathbb{E}\left[\prod_{j=1}^{n} e^{itX_j}\right] = \prod_{j=1}^{n} \mathbb{E}[e^{itX_j}] = \prod_{j=1}^{n} f_j(t).$$

The convolution $F_1 * F_2$ of two distribution functions F_1 and F_2 is defined as

$$(F_1 * F_2)(x) = \int_{-\infty}^{+\infty} F_1(x-y)dF_2(y).$$

Let X_1 and X_2 be independent random variables with distribution functions F_1 and F_2. Then $X_1 + X_2$ has the distribution function $F_1 * F_2$—Fubini's theorem can help to verify

$$\mathbb{P}[X_1 + X_2 \leq x] = \mathbb{E}[1_{X_1+X_2 \leq x}] = (F_1 * F_2)(x).$$

Because + among random variables is, the binary convolution operation $*$ is commutative and associative. The convolution $p_1 * p_2$ of probability density functions p_1 and p_2 satisfies

$$(p_1 * p_2)(x) = \int_{-\infty}^{+\infty} p_1(x-y)p_2(y)dy.$$

Using Fubini's theorem, we can show that $p_1 * p_2$ is the density of the convolution of two absolutely continuous distribution functions F_1 and F_2 with densities p_1 and p_2. That is,

$$\int_{-\infty}^{x} (p_1 * p_2)(u)du = (F_1 * F_2)(x).$$

For arbitrary subsets A and B of $\Re^1$, we denote their vector sum and difference by $A+B$ and $A-B$, respectively:

$$A \pm B = \{x \pm y : \; x \in A, y \in B\}.$$

If we denote by $\mu_1 * \mu_2$ the probability measure that corresponds to $F_1 * F_2$, then

$$(\mu_1 * \mu_2)(B) = \int_{\Re^1} \mu_1(B-y)\mu_2(dy).$$

For $B = (-\infty, x]$, the right-hand side will become $\int_{-\infty}^{+\infty} F_1(x-y)dF_2(y)$. For any Borel mearuable g that is integrable with respect to $\mu_1 * \mu_2$, we have

$$\int_{\Re^1} g(u)(\mu_1 * \mu_2)(du) = \int_{\Re^1}\int_{\Re^1} g(x+y)\mu_1(dx)\mu_2(dy).$$

When $g(x) = 1_{x \in B}$, we have $g_y(x) \equiv g(x+y) = 1_{x+y \in B} = 1_{x \in B-y}$. The above can be achieved by first considering simple functions and then passing to the limit. Consequently,

$$\int e^{itu}(\mu_1 * \mu_2)(du) = \int e^{it(x+y)}\mu_1(dx)\mu_2(dy) = \int e^{itx}\mu_1(dx)\int e^{ity}\mu_2(dy).$$

Addition of a finite number of independent random variables corresponds to convolution of their distribution functions and multiplication of characteristic functions. So if f is a c.f., so is $|f|^2 = f\overline{f}$. A few well known cases are as follows:

(1) Point mass at a: d.f. δ_a; ch.f. e^{iat}.

(2) Symmetric Bernoullian distribution with mass 1/2 each at +1 and −1:

$$\text{d.f. } \frac{\delta_1 + \delta_{-1}}{2}; \qquad \text{ch.f. } \cos(t).$$

(3) Bernoullian distribution with success probability p and $q \equiv 1-p$:

$$\text{d.f. } q\delta_0 + p\delta_1; \qquad \text{ch.f. } q + pe^{it} = 1 + p(e^{it}-1).$$

(4) Binomial distribution for n trials with success probability p:

$$\text{d.f. } \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} p^k q^{n-k}\delta_k; \qquad \text{ch.f. } (q+pe^{it})^n.$$

(5) Geometric distribution with success probability p:

$$\text{d.f. } \sum_{n=0}^{+\infty} q^n p \delta_n; \qquad \text{ch.f. } \frac{p}{1-qe^{it}}.$$

(6) Poisson distribution with parameter λ:

$$\text{d.f. } \sum_{n=0}^{+\infty} e^{-\lambda}\frac{\lambda^n}{n!}\delta_n; \qquad \text{ch.f. } e^{\lambda(\exp(it)-1)}.$$

(7) Exponential distribution with rate λ:

$$\text{p.d. } \lambda e^{-\lambda x} \text{ in } [0,+\infty); \qquad \text{ch.f. } \frac{1}{1-it/\lambda}.$$

(8) Uniform distribution in $[-a,+a]$:

$$\text{p.d. } \frac{1}{2a} \text{ in } [-a,a]; \qquad \frac{\sin(at)}{at}(= 1 \text{ for } t=0).$$

(9) Triangular distribution in $[-a,a]$:

$$\text{p.d. } \frac{a-|x|}{a^2} \text{ in } [-a,a]; \qquad \text{ch.f. } \frac{2-2\cos(at)}{a^2t^2} \equiv \left(\frac{\sin(at/2)}{at/2}\right)^2.$$

(10) Reciprocal of the above:

$$\text{p.d. } \frac{1-\cos(ax)}{\pi a x^2} \text{ in } (-\infty,+\infty); \qquad \text{ch.f. } \left(1-\frac{|t|}{a}\right) \vee 0.$$

(11) Normal distribution $N(m,\sigma)$ with mean m and standard deviation σ:

$$\text{p.d. } \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] \text{ in } (-\infty,+\infty); \qquad \text{ch.f. } \exp\left(imt-\frac{\sigma^2t^2}{2}\right).$$

(12) Cauchy distribution with parameter $a>0$:

$$\text{p.d. } \frac{a}{\pi(a^2+x^2)} \text{ in } (-\infty,+\infty); \qquad \text{ch.f. } e^{-a|t|}.$$

Convolution with the normal kernel is particularly effective. Let n_δ be the density of the normal distribution $N(0,\delta)$. For any bounded measurable function f, put

$$f_\delta(x) = (f * n_\delta)(x) = \int_{-\infty}^{+\infty} f(x-y)n_\delta(y)dy = \int_{-\infty}^{+\infty} n_\delta(x-y)f(y)dy.$$

The integrals above converge. Let C_B^∞ denote the class of functions on $\Re^1$ which have bounded derivatives of all orders; C_U the class of bounded and uniformly continuous functions on $\Re^1$. For each $\delta>0$, we have $f_\delta \in C_B^\infty$. Furthermore if $f \in C_U$, then $f_\delta \longrightarrow f$ uniformly.

The following is an important analytic result:

$$\lim_{T\longrightarrow+\infty}\int_0^T \frac{\sin(\alpha x)}{x}dx = \frac{\pi}{2}\text{sgn}(\alpha).$$

Indeed, we have

$$\lim_{T\longrightarrow+\infty}\int_0^{+\infty}\frac{\sin(x)}{x}dx=\lim_{T\longrightarrow+\infty}\int_0^T\sin(x)\left[\int_0^{+\infty}e^{-xu}du\right]dx=\int_0^{\infty}\left[\int_0^{+\infty}e^{-xu}\sin(x)dx\right]du,$$

which is equal to $\int_0^{+\infty}du/(1+u^2)$ and further to $\pi/2$. The following formulas related to the "Dirichlet integrals" are also useful:

$$\int_0^{\pi}\frac{\sin(x)}{x}dx\geq\operatorname{sgn}(\alpha)\int_0^y\frac{\sin(\alpha x)}{x}dx\geq 0.$$

$$\int_0^{+\infty}\frac{1-\cos(\alpha x)}{x^2}dx=\frac{\pi}{2}|\alpha|.$$

We want to answer the question of how to find the corresponding distribution function F or probability measure μ when given a characteristic function f. The inversion formula is of theoretical importance. By the above properties and Fubini's theorem, if $x_1<x_2$, then

$$\mu((x_1,x_2))+\frac{1}{2}\mu(\{x_1\})+\frac{1}{2}\mu(\{x_2\})=\lim_{T\longrightarrow+\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-itx_1}-e^{-itx_2}}{it}f(t)dt.$$

One useful equation has

$$\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-itx_1}-e^{-itx_2}}{it}\left[\int_{-\infty}^{+\infty}e^{itx}\mu(dx)\right]dt=\int_{-\infty}^{+\infty}\left[\int_{-T}^{T}\frac{e^{it(x-x_1)}-e^{it(x-x_2)}}{2\pi it}\right]\mu(dx).$$

The left-hand side can be $F(x_2)-F(x_1)$ if x_1 and x_2 are points of continuity of F. If

$$I(T,x,x_1,x_2)=\frac{1}{\pi}\int_0^T\frac{\sin(tx-tx_1)}{t}dt-\frac{1}{\pi}\int_0^T\frac{\sin(tx-tx_2)}{t}dt,$$

then it will follow from the above that

$$\lim_{T\longrightarrow+\infty}I(T,x,x_1,x_2)=\begin{cases}-1/2-(-1/2)=0 & \text{for } x<x_1,\\ 0-(-1/2)=1/2 & \text{for } x=x_1,\\ 1/2-(1-1/2)=1 & \text{for } x_1<x<x_2,\\ 1/2-0=1/2 & \text{for } x=x_2,\\ 1/2-1/2=0 & \text{for } x>x_2.\end{cases}$$

The uniqueness theorem states that one ch.f. can be the result of only one p.m. or d.f. An important particular case: if $f\in L^1(-\infty,+\infty)$, then F is continuously differentiable and

$$F'(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}e^{-ixt}f(t)dt.$$

The derivative F' being continuous, we have $F(x) = \int_{-\infty}^{x} F'(u)du$. We can now state in a more symmetric form the inverse formula. If $f \in L^1$, then $p \in L^1$, where

$$p(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-ixt} f(t)dt, \qquad\qquad f(t) = \int_{-\infty}^{+\infty} e^{itx} p(x)dx.$$

For each x_0, we have

$$\mu(\{x_0\}) = \lim_{T \longrightarrow +\infty} \frac{1}{2T}\int_{-T}^{T} e^{-itx_0} f(t)dt.$$

We also have

$$\lim_{T \longrightarrow +\infty} \frac{1}{2\pi}\int_{-T}^{T} |f(t)|^2 dt = \sum_{x \in \Re^1} \mu(\{x\})^2.$$

So μ is atomless (F is continuous) if and only if the limit on the left-hand side is zero. A random variable X is called symmetric iff X and $-X$ have the same distribution. Its distribution μ satisfies that $\mu(B) = \mu(-B)$ for any $B \in \mathscr{B}$. This is true iff the ch.f. is real-valued for all t; namely, $f(t) = \overline{f(t)}$.

10 Central Limit Theory—B26, B27

Let μ_n, μ be probability measures with characteristic functions f_n, f. A necessary and sufficient condition for $\mu_n \Longrightarrow \mu$ is that $f_n(t) \longrightarrow f(t)$ for each t.

For each t, e^{itx} has bounded modulus and is continuous in x. The necessity follows from a property of weak convergence. For sufficiency, we apply Fubini's theorem to obtain

$$\frac{1}{u}\int_{-u}^{u}(1 - f_n(t))dt = \int_{-\infty}^{+\infty}\left[\frac{1}{u}\int_{-u}^{u}(1 - e^{itx})dt\right]\mu_n(dx) = 2\int_{-\infty}^{+\infty}\left(1 - \frac{\sin ux}{ux}\right)\mu_n(dx),$$

which is greater than $2\int_{|x| \geq 2/u}(1 - 1/|ux|)\mu_n(x)$ which is in turn greater than $\mu_n(x : |x| \geq 2/u)$. Since f is continuous at the origin and $f(0) = 1$, there is for positive ϵ a u for which $\int_{-u}^{u}(1 - f(t))dt/u < \epsilon$. Since f_n converges to f, the bounded convergence theorem implies that there exists an n_0 such that $\int_{-u}^{u}(1 - f_n(t))dt/u < 2\epsilon$ for $n \geq n_0$. If $a = 2/u$ in the above, then $\mu_n(x : |x| \geq a) < 2\epsilon$ for $n \geq n_0$. Increasing a if necessary will ensure that the inequality also holds for the finitely many n preceding n_0. Therefore, $\{\mu_n\}$ is tight.

Now $\mu_n \Longrightarrow \mu$ will follow if it is shown that each subsequence $\{\mu_{n_k}\}$ that converges weakly at all converges weakly to μ. But if $\mu_{n_k} \Longrightarrow \nu$, then by the necessity ν has a characteristic function $\lim_k f_{n_k}(t) = f(t)$. By the uniqueness result, $\nu = \mu$.

Let $\{X_n\}$ be a sequence of independent random variables with the common d.f. F, and $S_n = \sum_{j=1}^{n} X_j$. If F has a finite mean m, then $S_n/n \longrightarrow m$ in probability. This amounts to showing the ch.f. of S_n/n converges to e^{imt}. But by expanding to the first order,

$$\mathbb{E}[e^{it(S_n/n)}] = \mathbb{E}[e^{i(t/n)S_n}] = \left[f\left(\frac{t}{n}\right)\right]^n = \left(1 + im\frac{t}{n} + o\left(\frac{t}{n}\right)\right)^n,$$

which converges to e^{imt} as $n \longrightarrow +\infty$ at each fixed t. This weak law of large numbers can be strengthened by both the strong law of large numbers and the central limit theorem.

If F has mean m and finite variance $\sigma^2 > 0$, then $(S_n - mn)/(\sigma\sqrt{n}) \longrightarrow Z$ in distribution, where Z is the standard normal distribution. We may suppose $m = 0$ by considering the random variables $X_j - m$ whose second moment is σ^2. Now up to the second order,

$$\mathbb{E}\left[\exp\left(it\frac{S_n}{\sigma\sqrt{n}}\right)\right] = \left[f\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n = \left\{1 + \frac{i^2\sigma^2}{2}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + o\left(\frac{t^2}{\sigma^2 n}\right)\right\}^n,$$

which is equal to $\{1 - t^2/(2n) + o(t^2/n)\}^n$ that converges to $e^{-t^2/2}$, the ch.f. of Z's d.f. In general, central limit theorem refers to a result that asserts the convergence in distribution of a normed sum of random variables $(S_n - a_n)/b_n$ to Z.

The above Lindeberg-Levy theorem is a special case of the Lindeberg-Lyapounov theorems. For each n, let there be r_n independent random variables $\{X_{nk}, 1 \leq k \leq r_n\}$ on row n where $r_n \longrightarrow +\infty$ as $n \longrightarrow +\infty$. We have dealt with the special case where $X_{nk} = X_k$ and $r_n = n$. Suppose that the means are 0 and the variances are finite; write

$$\mathbb{E}[X_{nk}] = 0, \qquad \sigma_{nk}^2 = \mathbb{E}[X_{nk}^2], \qquad s_n^2 = \sum_{k=1}^{n} \sigma_{nk}^2.$$

Assume $s_n^2 > 0$ for large n. Lindeberg condition states

$$\lim_{n \longrightarrow +\infty} \sum_{k=1}^{r_n} \frac{1}{s_n^2} \int_{|X_{nk}| \geq \epsilon s_n} X_{nk}^2 \, d\mathbb{P} = 0.$$

The theorem asserts that $S_n/s_n \longrightarrow Z$ in distribution. This theorem contains the previous one because when $s_n^2 = n\sigma^2$, the condition becomes

$$\lim_{n \longrightarrow +\infty} \frac{1}{\sigma^2} \int_{|X_1| \geq \epsilon\sigma\sqrt{n}} X_1^2 \, d\mathbb{P} = 0,$$

which holds because $\{|X_1| \geq \epsilon\sigma\sqrt{n}\} \downarrow \emptyset$. Replacing X_{nk} by X_{nk}/s_n shows that there is no loss of generality by assuming $s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2 = 1$. The objective is to show

$$\prod_{k=1}^{r_n} f_{nk}(t) = \prod_{k=1}^{r_n} \left(1 - \frac{1}{2}t^2\sigma_{nk}^2\right) + o(1) = \prod_{k=1}^{r_n} e^{-t^2\sigma_{nk}^2/2} + o(1) = e^{-t^2/2} + o(1).$$

The process utilizes the inequality that

$$\left|\prod_{k=1}^{m} z_k - \prod_{k=1}^{m} w_k\right| \le \sum_{k=1}^{m} |z_k - w_k|,$$

when $z_1, ..., z_m$ and $w_1, ..., w_m$ are complex numbers of modulus at most 1. It also invokes some of the following results. Integration by parts shows that

$$\int_0^x (x-s)^n e^{is} ds = \frac{x^{n+1}}{n+1} + \frac{i}{n+1}\int_0^x (x-s)^{n+1} e^{is} ds,$$

and it follows by induction that

$$e^{ix} = \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is} ds.$$

These will also lead to

$$e^{ix} = \sum_{k=0}^{n} \frac{(ix)^k}{k!} + \frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1} (e^{is}-1) ds.$$

Combine the latest two, and we can obtain

$$\left| e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} \right| \le \min\left\{ \frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!} \right\}.$$

For $n = 0, 1, 2$, the inequality specializes to

$$|e^{ix} - 1| \le \min\{|x|, 2\},$$

$$|e^{ix} - (1+ix)| \le \min\left\{\frac{1}{2}x^2, 2|x|\right\},$$

$$\left|e^{ix} - \left(1 + ix - \frac{1}{2}x^2\right)\right| \le \min\left\{\frac{1}{6}|x|^3, x^2\right\}.$$

If X has an absolute moment $m^{(n)}$ of order n, it follows that

$$\left| f(t) - \sum_{k=0}^{n} \frac{(it)^k}{k!} m^{(k)} \right| \le \mathbb{E}\left[\min\left\{ \frac{|tX|^{n+1}}{(n+1)!}, \frac{2|tX|^n}{n!} \right\} \right].$$

For any t satisfying $\lim_n |t|^n m^{(n)}/n! = 0$, we must have

$$f(t) = \sum_{k=0}^{+\infty} \frac{(it)^k}{k!} m^{(k)}.$$

Suppose that $|X_{nk}|^{2+\delta}$ are integral for some $\delta > 0$ and that Lyapounov's condition

$$\lim_{n \longrightarrow +\infty} \sum_{k=1}^{r_n} \frac{1}{s_n^{2+\delta}} \mathbb{E}\left[|X_{nk}|^{2+\delta}\right] = 0$$

holds. Then Lindeberg's condition follows because the sum in the former condition is below

$$\sum_{k=1}^{r_n} \frac{1}{s_n^2} \int_{|X_{nk}| \geq \epsilon s_n} \frac{|X_{nk}|^{2+\delta}}{\epsilon^\delta s_n^\delta} \, d\mathbb{P} \leq \frac{1}{\epsilon^\delta} \sum_{k=1}^{r_n} \frac{1}{s_n^{2+\delta}} \, \mathbb{E}\left[|X_{nk}|^{2+\delta}\right].$$

Thus, we will again have $S_n/s_n \longrightarrow Z$ in distribution under Lyapounov's condition.

11 Radon-Nikodym to Conditional Expectation—B32, B33, B34

If f is a positive function on a measurable space $(\Omega, \mathscr{F}, \mu)$, then $\nu(A) = \int_A f d\mu$ defines another measure on $\mathscr{F}$. In a sense, ν has density f with respect to μ. For each $A \in \mathscr{F}$, $\mu(A) = 0$ implies $\nu(A) = 0$. The converse is true as well when μ and ν both are σ-finite.

An additive function φ from $\mathscr{F}$ to the reals satisfies

$$\varphi\left(\bigcup_n A_n\right) = \sum_n \varphi(A_n)$$

when $A_1, A_2, ...$ is a sequence of disjoint sets. A set function differs from a measure in that the values $\varphi(A)$ may be negative but must be finite. If μ_1 and μ_2 are finite measures, then $\varphi(A) = \mu_1(A) - \mu_2(A)$ is an additive set function. It will turn out that the general additive set function has this form. A special case is $\varphi(A) = \int_A f d\mu$ where f is integrable. If $E_n \uparrow E$ or $E_n \downarrow E$, then $\varphi(E_n) \longrightarrow \varphi(E)$.

Hahn decomposition: For any additive set function φ, there exist disjoint sets A^+ and A^- such that $A^+ \cup A^- = \Omega$, $\varphi(E) \geq 0$ for all E in A^+, and $\varphi(E) \leq 0$ for all E in A^-. For $\varphi(A) = \int_A f d\mu$, we can simply take $A^+ = \{f \geq 0\}$ and $A^- = \{f \leq 0\}$. To prove, let $\alpha = \sup\{\varphi(A) : A \in \mathscr{F}\}$. If there exists a set A^+ satisfying $\varphi(A^+) = \alpha$, we are done. The rest of the proof attempts to construct such a set. Choose sets A_n so that $\varphi(A_n) \longrightarrow \alpha$. Let B_{ni} for $1 \leq i \leq 2^n$ be the sets of the form $\bigcap_{k=1}^n A'_k$ where A'_k is either A_k or A_k^c. Let C_n be the union of those B_{ni}'s for which $\varphi(B_{ni}) > 0$. We have

$$\varphi(A_m) \leq \varphi(C_m) \leq \varphi(C_m \cup C_{m+1}) \leq \cdots \leq \varphi\left(\bigcup_{n=m}^{+\infty} C_n\right).$$

Construct $A^+ = \bigcap_{m=1}^{+\infty} \bigcup_{n=m}^{+\infty} C_m$, and we have

$$\varphi(A^+) = \lim_{m \longrightarrow +\infty} \varphi\left(\bigcup_{n=m}^{+\infty} C_n\right) \geq \lim_{m \longrightarrow +\infty} \varphi(A_m) = \alpha.$$

If $\varphi^+(A) = \varphi(A \cap A^+)$ and $\varphi^-(A) = -\varphi(A \cap A^-)$, then φ^+ and φ^- are finite measures. Thus,

$$\varphi(A) = \varphi^+(A) - \varphi^-(A)$$

represents the set function φ as the difference of two finite measures having disjoint supports. If $E \subseteq A$, then $\varphi(E) \leq \varphi^+(E) \leq \varphi^+(A)$, and there is equality if $E = A \cap A^+$. Therefore, $\varphi^+(A) = \sup_{E \subseteq A} \varphi(E)$. Similarly, $\varphi^-(A) = -\inf_{E \subseteq A} \varphi(E)$. The measures φ^+ and φ^- are called the upper and lower variations of φ, and the measure $|\varphi| \equiv \varphi^+ + \varphi^-$ is called the total variation. This is the Jordan decomposition.

The measure ν is absolutely continuous with respect to μ if $\mu(A) = 0$ implies $\nu(A) = 0$ for each $A \in \mathscr{F}$. This is denoted as $\nu \ll \mu$. When ν is finite, this is equivalent to the existence of $\delta > 0$ for any $\epsilon > 0$ so that $\nu(A) < \epsilon$ whenever $\mu(A) < \delta$. If $\nu(A) = \int_A f d\mu$, certainly $\nu \ll \mu$. The Radon-Nikodym theorem goes in the opposite direction. If μ and ν are σ-finite measures such that $\nu \ll \mu$, then there exists a positive f, a density, such that $\nu(A) = \int_A f d\mu$ for all $A \in \mathscr{F}$. For two such densities f and f', $\mu(\{f \neq f'\}) = 0$. The f is often denoted by $d\nu/d\mu$, the Radon-Nikodym derivative. A proof for the finite-measure case would suffice. For this, we need the following lemma: If μ and ν are finite measures without mutually exclusive supports, then there exist a set A and $\epsilon > 0$ such that $\mu(A) > 0$ and $\epsilon\mu(E) \leq \nu(E)$ for all $E \subseteq A$. Indeed, let $A_n^+ \cup A_n^-$ be a Hahn decomposition of $\nu - \mu/n$; put $M = \bigcup_n A_n^+$ and $M^c = \bigcap_n A_n^-$. We can show $\nu(M^c) = 0$ and hence $\mu(M) > 0$. So $\mu(A_n^+) > 0$ for some n. Take $A = A_n^+$ and $\epsilon = 1/n$.

To prove the Radon-Nikodym theorem, let $\mathscr{G}$ be the class of positive functions g such that $\int_E g d\mu \leq \nu(E)$. It is closed under the formation of finite maxima and increasing passages to the limit. Let $\alpha = \sup\{\int g d\mu : g \in \mathscr{G}\}$. Choose $g_n \in \mathscr{G}$ so that $\int g_n d\mu > \alpha - 1/n$. If $f_n = \max(g_1, ..., g_n)$ and $f = \lim_n f_n$, then f lies in $\mathscr{G}$ and $\int f d\mu = \lim_n \int f_n d\mu \geq \lim_n \int g_n d\mu = \alpha$. Thus, f is an element of $\mathscr{G}$ for which $\int f d\mu$ is maximal. Define ν_{ac} by $\nu_{ac}(E) = \int_E f d\mu$ and ν_s by $\nu_s(E) = \nu(E) - \nu_{ac}(E)$. Using the lemma, we can show that ν_s is singular. That is,

$$\nu(E) = \nu_{ac}(E) + \nu_s(E) = \int_E f d\mu + \nu_s(E)$$

does constitute a Lebesgue decomposition. But since $\nu \ll \mu$, the only possibility is $\nu_S = 0$ and hence f is the desired derivative.

For conditional probability, we first consider the discrete case. Let $B_1, B_2, ...$ be a finite or countable partition of Ω into $\mathscr{F}$-sets, and let $\mathscr{G}$ consist of all the unions of the B_i. For $A \in \mathscr{F}$, consider the function with values

$$f(\omega) = \mathbb{P}[A|B_i] \equiv \frac{\mathbb{P}[A \cap B_i]}{\mathbb{P}[B_i]} \qquad \text{if } \omega \in B_i, i = 1, 2, ...$$

The partition $\{B_i\}$ or equivalently the σ-field $\mathscr{G}$ can be regarded as an experiment, and to learn which B_i it is that contains ω is to learn the outcome of the experiment. For this reason the function or random variable f is called the conditional probability of A given $\mathscr{G}$ and denoted by $\mathbb{P}[A||\mathscr{G}]$. Thus $\mathbb{P}[A||\mathscr{G}]$ is the function whose value on B_i is the ordinary conditional probability $\mathbb{P}[A|B_i]$. In case $\mathbb{P}[B_i] = 0$, the function will be taken to have any constant value on B_i; the value is arbitrary but must be the same over all of the set B_i. A specific such function is often called a version of the conditional probability; any two versions are equal except on a set of probability zero.

In general, one can imagine an observer who knows for each G in $\mathscr{G}$ whether $\omega \in G$ or not. Thus the σ-field $\mathscr{G}$ can in principle be identified with an experiment or observation. It is natural to try and define conditional probabilities $\mathbb{P}[A||\mathscr{G}]$ with respect to the experiment $\mathscr{G}$. To do this, fix an $A \in \mathscr{F}$ and define a finite measure ν on $\mathscr{G}$ by

$$\nu(G) = \mathbb{P}[A \cap G], \qquad G \in \mathscr{G}.$$

Then $\mathbb{P}[G] = 0$ implies $\nu(G) = 0$. By the Radon-Nikodym theorem, there a function f measurable $\mathscr{G}$ and integrable with respect to $\mathbb{P}$, such that $\mathbb{P}[A \cap G] = \nu(G) = \int_G f d\mathbb{P}$ for all $G \in \mathscr{G}$. Denote this f by $\mathbb{P}[A||\mathscr{G}]$. It is a random variable with two properties:

(i) $\mathbb{P}[A||\mathscr{G}]$ is measurable $\mathscr{G}$ and integrable.

(ii) For any $G \in \mathscr{G}$, it satisfies $\mathbb{P}[A \cap G] = \int_G \mathbb{P}[A||\mathscr{G}] \cdot d\mathbb{P}$.

There will in general be many such random variables $\mathbb{P}[A||\mathscr{G}]$, but any two of them are equal with probability 1. A specific such random variable is called a version of the conditional probability. In case $\mathscr{G}$ is generated by a partition $B_1, B_2, ...$, the function f defined through $\mathbb{P}[A|B_i]$ earlier is measurable $\mathscr{G}$. Any G in $\mathscr{G}$ is a disjoint union $G = \bigcup_k B_{i_k}$ and $\mathbb{P}[A \cap G] = \sum_k \mathbb{P}[A|B_{i_k}] \cdot \mathbb{P}[B_{i_k}]$. Thus the general definition is an extension of the discrete case.

If $A \in \mathscr{G}$, then 1_A satisfies both (i) and (ii). Here, to know the outcome of $\mathscr{G}$ viewed as an experiment is to know whether or not A has occurred. If $\mathscr{G}$ is $\{0, \Omega\}$, we have $\mathbb{P}[A||\mathscr{G}]_\omega = \mathbb{P}[A]$ for all ω. The observer learns nothing from the experiment $\mathscr{G}$. Therefore, 1_A is a version of $\mathbb{P}[A||\mathscr{F}]$ and $\mathbb{P}[A||\{\emptyset, \Omega\}]$ is identically $\mathbb{P}[A]$. For any $\mathscr{G}$, the latter always satisfies (i) and the former always (ii). Condition (i) becomes more stringent as $\mathscr{G}$ decreases, and condition

(ii) becomes more stringent as $\mathscr{G}$ increases. The two conditions work in opposite directions and between them delimit the class of versions of $\mathbb{P}[A||\mathscr{G}]$.

The set A is by definition independent of the σ-field $\mathscr{G}$ if it is independent of each G in $\mathscr{G}$: $\mathbb{P}[A \cap G] = \mathbb{P}[A] \cdot \mathbb{P}[G]$. This being the same thing as $\mathbb{P}[A \cap G] = \int_G \mathbb{P}[A] \cdot d\mathbb{P}$, A is independent of $\mathscr{G}$ iff $\mathbb{P}[A||\mathscr{G}] = \mathbb{P}[A]$ with probability 1.

The σ-field $\sigma(X)$ generated by a random variable X consists of the sets $\{\omega : X(\omega) \in B\}$ for $B \in \mathscr{B}^1$. The conditional probability of A given X is defined as $\mathbb{P}[A||\sigma(X)]$ and denoted by $\mathbb{P}[A||X]$. The definition applies without change to random vector, or equivalently, to a finite set of random variables. It can be adapted to arbitrary sets of random variables as well. For any such set $(X_t, t \in T)$, the σ-field $\sigma(X_t, t \in T)$ it generates is the smallest σ-field with respect to which each X_t is measurable. It is generated by the collection of sets of the form $\{\omega : X_t(\omega) \in B\}$ for $t \in T$ and $B \in \mathscr{B}^1$. The conditional probability $\mathbb{P}[A||X_t, t \in T]$ of A with respect to this set of random variables is by definition the conditional probability $\mathbb{P}[A||\sigma(X_t, t \in T)]$ of A with respect to the σ-field $\sigma(X_t, t \in T)$. In this notation, the property of Markov chains becomes

$$\mathbb{P}[X_{n+1} = j||X_0, ..., X_n] = \mathbb{P}[X_{n+1} = j||X_n].$$

For a fixed $\mathscr{G}$, the conditional probability $\mathbb{P}[\cdot||\mathscr{G}]$ behaves just like an ordinary probability almost surely. With probability 1, $\mathbb{P}[\emptyset||\mathscr{G}] = 0$, $\mathbb{P}[\Omega||\mathscr{G}] = 1$, $0 \leq \mathbb{P}[A||\mathscr{G}] \leq 1$ for each $A \in \mathscr{F}$; if $A_1, A_2, ...$ is a finite or countable sequence of disjoint sets,

$$\mathbb{P}\left[\bigcup_n A_n||\mathscr{G}\right] = \sum_n \mathbb{P}[A_n||\mathscr{G}].$$

For $\int_G \mathbb{P}[A||\mathscr{G}]d\mathbb{P} = \mathbb{P}[A \cap G] \geq 0$ for each $G \in \mathscr{G}$ and $\mathbb{P}[A||\mathscr{G}]$ being measurable $\mathscr{G}$, it must be the case that $\mathbb{P}[A||\mathscr{G}]$ is positive except on a set of $\mathbb{P}$-measure 0. The other inequality can be proved the same way. If the A_n's are disjoint and if G lies in $\mathscr{G}$, it follows that

$$\int_G \left(\sum_n \mathbb{P}[A_n||\mathscr{G}]\right) d\mathbb{P} = \sum_n \int_G \mathbb{P}[A_n||\mathscr{G}] \cdot d\mathbb{P} = \sum_n \mathbb{P}[A_n \cap G] = \mathbb{P}\left[\left(\bigcup_n A_n\right) \cap G\right].$$

Thus $\sum_n \mathbb{P}[A_n||\mathscr{G}]$ satisfies the functional equation for $[\bigcup_n A_n||\mathscr{G}]$, and so must coincide with it except perhaps on a set of $\mathbb{P}$-measure 0.

Let X be a random variable on $(\Omega, \mathscr{F}, \mathbb{P})$, and let $\mathscr{G}$ be a σ-field in $\mathscr{F}$. There exists a function $\mu(B, \omega)$, defined for $B \in \mathscr{B}^1$ and $\omega \in \Omega$, with these two properties:

(i) For each $\omega \in \Omega$, $\mu(\cdot, \omega)$ is a probability measure on $\mathscr{B}^1$.

(ii) For each $B \in \mathscr{B}^1$, $\mu(B, \cdot)$ is a version of $\mathbb{P}[X \in B||\mathscr{G}]$.

The probability measure $\mu(\cdot,\cdot)$ is a conditional distribution of X given $\mathscr{G}$. If $\mathscr{G} = \sigma(Y)$, it is a conditional distribution of X given Y.

Suppose X is an integrable random variable on $(\Omega, \mathscr{F}, \mathbb{P})$ and that $\mathscr{G}$ is a σ-field in $\mathscr{F}$. There exists a random variable $\mathbb{E}[X||\mathscr{G}]$, called the conditional expected value of X given $\mathscr{G}$, having these two properties:

(i) $\mathbb{E}[X||\mathscr{G}]$ is measurable $\mathscr{G}$ and integrable.

(ii) $\mathbb{E}[X||\mathscr{G}]$ satisfies the functional equation

$$\int_G \mathbb{E}[X||\mathscr{G}] \cdot d\mathbb{P} = \int_G X \cdot d\mathbb{P}, \qquad \forall G \in \mathscr{G}.$$

The proof the existence can start with a positive X. Define measure μ on $\mathscr{G}$ by $\nu(G) = \int_G X d\mathbb{P}$. It is finite because X is integrable, and it is absolutely continuous with respect to $\mathbb{P}$. By the Radon-Nikodym theorem there is a $\mathscr{G}$-measurable function f such that $\nu(G) = \int_G f d\mathbb{P}$. This f has properties (i) and (ii). For a general X, then $\mathbb{E}[X^+||\mathscr{G}] - \mathbb{E}[X^-||\mathscr{G}]$ clearly has the required properties. Any two versions of the conditional expectation are equal with probability 1. Note $\mathbb{E}[X|\{\emptyset, \Omega\}] = \mathbb{E}[X]$ and $\mathbb{E}[X|\mathscr{F}] = X$ with probability 1. As $\mathscr{G}$ increases, (i) becomes weaker and (ii) becomes stronger.

Suppose $B_1, B_2, ...$ is a finite or countable partition of Ω generating the σ-field $\mathscr{G}$. Then

$$\mathbb{E}[X||\mathscr{G}]_\omega = \frac{1}{\mathbb{P}[B_i]} \int_{B_i} X d\mathbb{P}, \qquad \forall \omega \in B_i \text{ with } \mathbb{P}[B_i] > 0.$$

Also, $\mathbb{E}[1_A||\mathscr{G}] = \mathbb{P}[A||\mathscr{G}]$ with probability 1 because after all, $\int_G 1_A d\mathbb{P} = \mathbb{P}[A \cap G]$. More generally, $\mathbb{E}[X||\mathscr{G}] = \sum_i \alpha_i \mathbb{P}[A_i||\mathscr{G}]$ with probability 1 for a simple function $X = \sum_i \alpha_i 1_{A_i}$. Suppose X, Y, X_n are integrable,

(i) If $X = a$ with probability 1, then $\mathbb{E}[X||\mathscr{G}] = a$.

(ii) For constants a and b, $\mathbb{E}[aX + bY||\mathscr{G}] = a\mathbb{E}[X||\mathscr{G}] + b\mathbb{E}[Y||\mathscr{G}]$.

(iii) If $X \leq Y$ with probability 1, then $\mathbb{E}[X||\mathscr{G}] \leq \mathbb{E}[Y||\mathscr{G}]$.

(iv) $|\mathbb{E}[X||\mathscr{G}]| \leq \mathbb{E}[|X|\ ||\mathscr{G}]$.

(v) If $\lim_n X_n = X$ with probability 1, $|X_n| \leq Y$ for some integrable Y, then $\lim_n \mathbb{E}[X_n||\mathscr{G}] = \mathbb{E}[X||\mathscr{G}]$ with probability 1. To prove, consider $Z_n = \sup_{k \geq n} |X_k - X|$. Now $Z_n \downarrow 0$ with probability 1 and also,

$$|\mathbb{E}[X_n||\mathscr{G}] - \mathbb{E}[X||\mathscr{G}]| = |\mathbb{E}[X_n - X||\mathscr{G}]| \leq \mathbb{E}[|X_n - X|\ ||\mathscr{G}] \leq \mathbb{E}[Z_n||\mathscr{G}].$$

It suffices to show that $\mathbb{E}[Z_n||\mathscr{G}] \downarrow 0$ with probability 1. Since the sequence is already decreasing with a limit Z, the problem is to prove that $Z = 0$ with probability 1 or, Z is positive and yet $\mathbb{E}[Z] = 0$. But $0 \leq Z_n \leq 2Y$, and so property (ii) of conditional expectations and the dominated convergence theorem would give

$$\mathbb{E}[Z] = \int \mathbb{E}[Z||\mathscr{G}]d\mathbb{P} \leq \int \mathbb{E}[Z_n||\mathscr{G}]d\mathbb{P} = \mathbb{E}[Z_n] \longrightarrow 0.$$

For an observer with the information in $\mathscr{G}$, a $\mathscr{G}$-measurable X is effectively a constant. In particular, if Y and XY are integrable, then $\mathbb{E}[XY||\mathscr{G}] = X\mathbb{E}[Y||\mathscr{G}]$ with probability 1. We can first show that $1_G\mathbb{E}[Y||\mathscr{G}]$ is a version of $\mathbb{E}[1_GY||\mathscr{G}]$ for $G \in \mathscr{G}$. It is certainly measurable $\mathscr{G}$, and for any $G' \in \mathscr{G}$, we do have

$$\int_{G'} 1_G\mathbb{E}[Y||\mathscr{G}]d\mathbb{P} = \int_{G\cap G'} \mathbb{E}[Y||\mathscr{G}]d\mathbb{P} = \int_{G\cap G'} Y d\mathbb{P} = \int_{G'} 1_G Y d\mathbb{P}.$$

The result will hold if X is a simple function. For a general X that is measurable $\mathscr{G}$, there exist simple functions X_n measurable $\mathscr{G}$ such that $|X_n| \leq |X|$ and $\lim_n X_n = X$. Since $|X_nY| \leq |XY|$ and $|XY|$ is integrable,

$$X\mathbb{E}[Y||\mathscr{G}] = \lim_n X_n\mathscr{E}[Y||\mathscr{G}] = \lim_n \mathscr{E}[X_nY|\mathscr{G}] = \mathbb{E}[XY|\mathscr{G}],$$

with probability 1. Note X has not been assumed integrable.

If X is integrable and the σ-fields $\mathscr{G}_1$ and $\mathscr{G}_2$ satisfy $\mathscr{G}_1 \subseteq \mathscr{G}_2$, then

$$\mathbb{E}\left[\mathbb{E}[X||\mathscr{G}_2]\;||\mathscr{G}_1\right] = \mathbb{E}[X||\mathscr{G}_1].$$

Note the left-hand side is measurable $\mathscr{G}_1$. So the key is for $G \in \mathscr{G}_1 \subseteq \mathscr{G}_2$,

$$\int_G \mathbb{E}\left[\mathbb{E}[X||\mathscr{G}_2]\;||\mathscr{G}_1\right] \cdot d\mathbb{P} = \int_G \mathbb{E}[X||\mathscr{G}_2] \cdot d\mathbb{P} = \int_G X d\mathbb{P}.$$

Jensen's inequality for conditional expectation states that, if φ is a convex function on the line and X and $\varphi(X)$ are both integrable, then with probability 1,

$$\varphi(\mathbb{E}[X||\mathscr{G}]) \leq \mathbb{E}[\varphi(X)||\mathscr{G}].$$

For each x_0 take a support line through $(x_0, \varphi(x_0))$: $\varphi(x_0) + A(x_0)\cdot(x - x_0) \leq \varphi(x)$ where the slope $A(x_0)$ can be taken as the right-hand derivative of φ, so that it is increasing in x_0. Replacing x_0 with $\mathbb{E}[X|\mathscr{G}]$, we have

$$\varphi(\mathbb{E}[X||\mathscr{G}]) + A(\mathbb{E}[X||\mathscr{G}])\cdot(X - \mathbb{E}[X||\mathscr{G}]) \leq \varphi(X).$$

Suppose $\mathbb{E}[X||\mathscr{G}]$ is bounded. Then all three terms are integrable, and taking expected values with respect to $\mathscr{G}$ and using the just proved result on the middle term would do. In general, let $G_n = \{|\mathbb{E}[X||\mathscr{G}]| \leq n\}$. Then $\mathbb{E}[1_{G_n}X||\mathscr{G}] = 1_{G_n}\mathbb{E}[X||\mathscr{G}]$ is bounded, and so the inequality holds: $\varphi(1_{G_n}\mathbb{E}[X||\mathscr{G}]) \leq \mathbb{E}[\varphi(1_{G_n}X)||\mathscr{G}]$. The left-hand side $\varphi(1_{G_n}\mathbb{E}[X||\mathscr{G}])$ converges to $\varphi(\mathbb{E}[X||\mathscr{G}])$ by the continuity of φ; meanwhile, the right-hand side

$$\mathbb{E}[\varphi(1_{G_n}X)||\mathscr{G}] = \mathbb{E}[1_{G_n}\varphi(X) + 1_{G_n^c}\varphi(0)||\mathscr{G}] = 1_{G_n}\mathbb{E}[\varphi(X)||\mathscr{G}] + 1_{G_n^c}\varphi(0) \longrightarrow \mathbb{E}[\varphi(X)||\mathscr{G}].$$

Note the inequality for $\varphi(x) = |x|$ will revert back to the (iv) before.

Let $\mu(\cdot, \omega)$ be a conditional distribution of a random variable X with respect to $\mathscr{G}$ in the earlier sense that $\mu(B, \omega) = \mathbb{P}[X \in B|\mathscr{G}]_\omega$ for any $B \in \mathscr{B}^1$. If $\varphi : \Re^1 \longrightarrow \Re^1$ is a Borel function for which $\varphi(X)$ is integrable, then $\int_{\Re^1} \varphi(x)\mu(dx, \omega)$ is a version of $\mathbb{E}[\varphi(X)|\mathscr{G}]_\omega$. We start with $\varphi = 1_B$ for some $B \in \mathscr{B}^1$, then follow it with a simple function. For any positive φ, we resort to an increasing sequence of simple functions and the monotone convergence theorem. The most general case is treated through the decomposition into positive and negative parts.

12 Large-deviation Bounds—CT2, CT3, CT11

We focus on discrete random variables. Let X be one with alphabet $\mathcal{X}$ and probability mass function $p(x) = \mathbb{P}[X = x]$ for $x \in \mathcal{X}$. The entropy $H(X)$ of X is defined by

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{1}{p(x)} \right),$$

with the log based on 2 and the understanding that $0 \log(1/0) = 0$. The entropy can be interpreted as the expected value of the random variable $\log(1/p(X))$—$H(X) = \mathbb{E}_p[\log(1/p(X))]$. Note that entropy is a functional of the distribution of X; it does not depend on the actual values taken by the random variable X. Surely, $H(X) \geq 0$. When X is Bernoulli with parameter p, $H(X) = p\log(1/p) + (1-p)\log(1/(1-p))$ is also denoted as $H(p)$. Note $H(p)$ is symmetric around $1/2$ where it reaches its peak of 1, $H(0) = H(1) = 0$, and is concave.

The relative entropy or Kullback-Leibler distance between two probability mass functions $p(x)$ and $q(x)$ is

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) = \mathbb{E}_p \left[\frac{p(X)}{q(X)} \right],$$

where we use the convention that $0 \log(0/0) = 0$, $0 \log(0/q) = 0$ for $q > 0$, and $p \log(p/0) = +\infty$ for $p > 0$. Especially, if there is a symbol $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$, then $D(p||q) = +\infty$. The relative entropy is always positive and is 0 iff $p = q$. Indeed, let

$A = \{x : p(x) > 0\}$. Then

$$\begin{aligned} -D(p||q) &= -\textstyle\sum_{x\in A} p(x)\log(p(x)/q(x)) = \sum_{x\in A} p(x)\log(q(x)/p(x)) \\ &= \mathbb{E}_p[\log(q(X)/p(X))] \le \log\left(\mathbb{E}_p[q(X)/p(X)]\right) \\ &= \log\left(\textstyle\sum_{x\in A} p(x)(q(x)/p(x))\right) = \log\left(\sum_{x\in A} q(x)\right) \le \log\left(\sum_{x\in\mathcal{X}} q(x)\right) = \log(1) = 0, \end{aligned}$$

where the inequality follows from Jensen's inequality and the concavity of $\log(\cdot)$. Since the concavity is strict, the equality can be had only if $p(x) = q(x)$ for every $x \in \mathcal{X}$. However, the relative entropy is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality.

We can also show that $H(X) \le \log(|\mathcal{X}|)$ where $|\mathcal{X}|$ denotes the cardinality of $\mathcal{X}$, with equality iff X has a uniform distribution over $\mathcal{X}$. Let $u(x) = 1/|\mathcal{X}|$ be the uniform probability mass function over $\mathcal{X}$, and let $p(x)$ be the probability mass function for X. Then

$$D(p||u) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{u(x)}\right) = \log(|\mathcal{X}|) - H(X),$$

which is positive and equal 0 only when $p = u$.

In information theory, the analog of the law of large numbers is the asymptotic equipartition property (AEP). The former states that for i.i.d. random variables, $\sum_{i=1}^n X_i/n$ is close to its expected value $\mathbb{E}_p[X]$ for large vales of n; the latter states that $\log(1/p(X_1, ..., X_n))/n$ is close to the entropy $H \equiv H(X) = \mathbb{E}_p[\log(1/p(X))]$. Thus, the probability $p(X_1, ..., X_n)$ assigned to an observed sequence will be close to 2^{-nH}. This enables us to divide the set of all sequences into two sets, the typical set, where the sample entropy is close to the true entropy, and the nontypical set, which contains the other sequences. Any property that is proved for the typical sequences will be true with high probability and will determine the average behavior of a large sample. We summarize this by saying, "Almost all events are almost equally surprising." That is,

$$\mathbb{P}_p\left[(X_1, ..., X_n) : p(X_1, ..., X_n) = 2^{-n(H\pm\epsilon)}\right] \simeq 1,$$

if $X_1, ..., X_n$ are i.i.d. with distribution p. Indeed,

$$\frac{1}{n}\log\left(\frac{1}{p(X_1, ..., X_n)}\right) = \frac{1}{n}\sum_{i=1}^n \log\left(\frac{1}{p(X_i)}\right) \longrightarrow \mathbb{E}_p\left[\log\left(\frac{1}{p(X)}\right)\right] = H(X),$$

where the convergence is in probability by the weak law of large numbers.

The typical set $A_\epsilon^{(n)}$ is the set of sequences $(x_1, ..., x_n) \in \mathcal{X}^n$ with the property

$$2^{-n(H(X)+\epsilon)} \le p(x_1, ..., x_n) \le 2^{-n(H(X)-\epsilon)}.$$

As a consequence of AEP, we can show that $A_\epsilon^{(n)}$ has the following properties:

(i) If $(x_1, ..., x_n) \in A_\epsilon^{(n)}$, then $H(X) - \epsilon \leq \log(1/p(x_1, ..., x_n))/n \leq H(X) + \epsilon$.

(ii) $\mathbb{P}_p \left[A_\epsilon^{(n)}\right] > 1 - \epsilon$ for n sufficiently large.

(iii) $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$, where $|A|$ denotes the cardinality of A.

(iv) $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(x)-\epsilon)}$ for n sufficiently large.

Thus, the typical set has probability nearly 1, all elements of the typical set are nearly equiprobable, and the number of elements in the typical set is nearly 2^{nH}. Note (i) comes from $A_\epsilon^{(n)}$'s definition; also, (ii) comes from AEP. When n is large enough,

$$\mathbb{P}_p \left[A_\epsilon^{(n)}\right] = \mathbb{P}_p \left[\left|\frac{1}{n} \log \left(\frac{1}{p(X_1, ..., X_n)}\right) - H(X)\right| \leq \epsilon\right] > 1 - \epsilon.$$

For (iii), note that

$$1 = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \geq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} p(\mathbf{x}) \geq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} = 2^{-n(H(X)+\epsilon)} |A_\epsilon^{(n)}|,$$

where the second inequality follows from $A_\epsilon^{(n)}$'s definition. For sufficiently large n,

$$1 - \epsilon < \mathbb{P}_p \left[A_\epsilon^{(n)}\right] = \sum_{\mathbf{x} \in A_\epsilon^{(n)}} p(\mathbf{x}) \leq \sum_{\mathbf{x} \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)} |A_\epsilon^{(n)}|,$$

where the second inequality follows from $A_\epsilon^{(n)}$'s definition. Thus we have (iv) as well.

The method of types can go further than AEP in deriving large-deviation results. Let $X_1, X_2, ..., X_n$ be a sequence of n symbols from an alphabet $\mathcal{X} \equiv \{a_1, a_2, ..., a_{|\mathcal{X}|}\}$. The type $P_{\mathbf{x}}$ or empirical probability distribution of a sequence $\mathbf{x} \equiv (x_1, x_2, ..., x_n)$ is the relative proportion of occurrences of each symbol of $\mathcal{X}$. That is, $P_{\mathbf{x}}(a) = N(a|\mathbf{x})/n$ for all $a \in \mathcal{X}$, where $N(a|\mathbf{x})$ is the number of times the symbol a occurs in the sequence $\mathbf{x} \in \mathcal{X}^n$.

Let $\mathcal{P}_n$ denote the set of types with denominator n. For example, if $\mathcal{X} = \{0, 1\}$, the set of possible types with denominator n is

$$\mathcal{P}_n = \left\{(P(0), P(1)) : \left(\frac{0}{n}, \frac{n}{n}\right), \left(\frac{1}{n}, \frac{n-1}{n}\right), ..., \left(\frac{n}{n}, \frac{0}{n}\right)\right\}.$$

If $P \in \mathcal{P}_n$, the set of sequences of length n and type P is the type class of P, denoted $T(P)$:

$$T(P) = \{\mathbf{x} \in \mathcal{X}^n : P_{\mathbf{x}} = P\}.$$

Let $\mathcal{X} = \{1, 2, 3\}$, a ternary alphabet. Let $\mathbf{x} = 11321$, Then the type $P_{\mathbf{x}}$ is

$$P_{\mathbf{x}}(1) = \frac{3}{5}, \qquad P_{\mathbf{x}} = \frac{1}{5}, \qquad P_{\mathbf{x}} = \frac{1}{5}.$$

Also, $T(P_{\mathbf{x}})$ contains all 5-symbol sequences with three 1's, one 2, and one 3. Its size is

$$|T(P)| = \frac{5!}{3!1!1!} = 20.$$

The essential power of the method of types arises from the following result, which shows that the number of types is at most polynomial in n: $|\mathcal{P}| \leq (n+1)^{|\mathcal{X}|}$. Indeed, there are $|\mathcal{X}|$ components in the vector that specifies $P_{\mathbf{x}}$. The numerator in each component can take on only $n+1$ values. So there are at most $(n+1)^{|\mathcal{X}|}$ choices for the type vector.

Since the number of sequences is exponential in n, it follows that at least one type has exponentially many sequences in its type class. In fact, the largest type class has essentially the same number of elements as the entire set of sequences, to first order in the exponent. Now we assume that the sequence $X_1, X_2, ..., X_2$ is drawn i.i.d. according to a distribution Q. All sequences with the same type have the same probability. Let $Q^n(\mathbf{x}) = \prod_{i=1}^n Q(x_i)$ denote the product distribution associated with Q. Then

$$Q^n(\mathbf{x}) = 2^{-n(H(P_{\mathbf{x}})+D(P_{\mathbf{x}}||Q))}.$$

To prove this, we note that

$$\begin{aligned} Q^n(\mathbf{x}) &= \textstyle\prod_{i=1}^n Q(x_i) = \prod_{a\in\mathcal{X}} Q(a)^{N(a|\mathbf{x})} = \prod_{a\in\mathcal{X}} Q(a)^{nP_{\mathbf{x}}(a)} = \prod_{a\in\mathcal{X}} 2^{nP_{\mathbf{x}}(a)\log Q(a)} \\ &= 2^{n(P_{\mathbf{x}}(a)\log Q(a) - P_{\mathbf{x}}(a)\log P_{\mathbf{x}}(a) + P_{\mathbf{x}}(a)\log P_{\mathbf{x}}(a))} = 2^{n\sum_{a\in\mathcal{X}}(-P_{\mathbf{x}}(a)\log(P_{\mathbf{x}}(a)/Q(a)) + P_{\mathbf{x}}\log P_{\mathbf{x}}(a))}, \end{aligned}$$

which is finally equal to $2^{n(-D(P_{\mathbf{x}}||Q)-H(P_{\mathbf{x}}))}$. If $P_{\mathbf{x}} = Q$ or equivalently $\mathbf{x} \in T(Q)$, then $Q^n(\mathbf{x}) = 2^{-nH(Q)}$. There is an estimate of the size of type class $T(P)$ for any type $P \in \mathcal{P}_n$:

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(p)} \leq |T(P)| \leq 2^{nH(P)}.$$

Indeed, there is the exact formula

$$|T(P)| = \frac{n!}{(nP(a_1))!(nP(a_2))!\cdots(nP(a_{|\mathcal{X}|}))!}.$$

So the proof could resort to Stirling's formula. We take an alternative approach. Note

$$1 \geq P^n(T(P)) = \sum_{\mathbf{x}\in T(P)} P^n(\mathbf{x}) = \sum_{\mathbf{x}\in T(P)} 2^{-nH(P)} = |T(P)|2^{-nH(P)}.$$

Thus, $|T(P)| \leq 2^{nH(P)}$. For the lower bound, we first prove that the type class $T(P)$ has the highest probability among all type classes under the probability distribution P:

$$P^n(T(P)) \geq P^n(T(Q)), \qquad \forall Q \in \mathcal{P}_n.$$

This is because

$$P^n(T(P))/P^n(T(Q)) = |T(P)|\prod_{a\in\mathcal{X}} P(a)^{nP(a)}/[|T(Q)|\prod_{a\in\mathcal{X}} P(a)^{nQ(a)}]$$
$$= n!/((nP(a_1))!\cdots(nP(a_{|\mathcal{X}|})!)\prod_{a\in\mathcal{X}} P(a)^{nP(a)}/[(n!/(nQ(a_1))!\cdots(nQ(a_{|\mathcal{X}|})!)\prod_{a\in\mathcal{X}} P(a)^{nQ(a)}],$$

which is equal to $\prod_{a\in\mathcal{X}}[(nQ(a))!/((nP(a))!)]P(a)^{n(P(a)-Q(a))}$. By separately considering the cases $m \geq n$ and $m < n$, we can show that $m!/n! \geq n^{m-n}$. Therefore, This is because

$$P^n(T(P))/P^n(T(Q)) \geq \prod_{a\in\mathcal{X}}(nP(a))^{nQ(a)-nP(a)}P(a)^{nP(a)-nQ(a)} = \prod_{a\in\mathcal{X}} n^{nQ(a)-nP(a)}$$
$$= n^{n(\sum_{a\in\mathcal{X}} Q(a)-\sum_{a\in\mathcal{X}} P(a))} = n^{n(1-1)} = 1.$$

Now we can continue to pursue the lower bound, since

$$1 = \sum_{Q\in\mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q\in\mathcal{P}_n} \max_Q P^n(T(Q)) = \sum_{Q\in\mathcal{P}_n} P^n(T(P)) \leq (n+1)^{|\mathcal{X}|}P^n(T(P))$$
$$= (n+1)^{|\mathcal{X}|}\sum_{\mathbf{x}\in T(P)} P^n(\mathbf{x}) = (n+1)^{|\mathcal{X}|}\sum_{\mathbf{x}\in T(P)} 2^{-nH(P)} = (n+1)^{|\mathcal{X}|}|T(P)|2^{-nH(P)}.$$

For any $P \in \mathcal{P}_n$ and distribution Q, the probability of the type class $T(P)$ under Q^n is $2^{-nD(P||Q)}$ to first order in the exponent. More precisely,

$$\frac{1}{(n+1)^{|\mathcal{X}|}}2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}.$$

We have

$$Q^n(T(P)) = \sum_{\mathbf{x}\in T(P)} Q^n(\mathbf{x}) = \sum_{\mathbf{x}\in T(P)} 2^{-n(D(P||Q)+H(P))} = |T(P)|2^{-n(D(P||Q)+H(P))},$$

while previously we have derived two bounds for $|T(P)|$. Since the probability of each type class depends exponentially on the relative entropy distance between the type P and the distribution Q, type classes that are far from the true distribution have exponentially smaller probability. For $\epsilon > 0$, define a typical set T_Q^ϵ of sequences for the distribution Q^n as

$$T_Q^\epsilon = \{\mathbf{x} : D(P_{\mathbf{x}}||Q) \leq \epsilon\}.$$

Then the probability that $\mathbf{x}$ is not typical is

$$1 - Q^n(T_Q^\epsilon) = \sum_{P:D(P||Q)>\epsilon} Q^n(T(P)) \leq \sum_{P:D(P||Q)>\epsilon} 2^{-nD(P||Q)} \leq \sum_{P:D(P||Q)>\epsilon} 2^{-n\epsilon},$$

which is below $(n+1)^{|\mathcal{X}|}2^{-n\epsilon} = 2^{-n[\epsilon-|\mathcal{X}|\log((n+1)/n)]}$. The latter would go to 0 as $n \longrightarrow +\infty$. Hence, the probability of the typical set T_Q^ϵ goes to 1 as $n \longrightarrow +\infty$. This is similar to AEP. Let $X^n \equiv (X_1, ..., X_n)$ be i.i.d. from P. Since $\mathbb{P}_P[D(P_{X^n}||P) > \epsilon] \leq 2^{-n[\epsilon-|\mathcal{X}|\log((n+1)/n)]}$,

$$\sum_{n=1}^{+\infty} \mathbb{P}_P[D(P_{X^n}||P) > \epsilon] < +\infty.$$

By Borel-Cantelli lemma, $D(P_{X^n}||P) \longrightarrow 0$ with probability 1.

Let E be a subset of the set of probability mass functions. If E contains a relative entropy neighborhood of Q, then by the just proved result, $Q^n(E \cap \mathcal{P}_n) \longrightarrow 1$. On the other hand, if E does not contain Q or a neighborhood of it, then $Q^n(E \cap \mathcal{P}_n) \longrightarrow 0$ exponentially fast. We can use the method of types to calculate the exponent and obtain Sanov's theorem:

$$Q^n(E \cap \mathcal{P}_n) \leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)},$$

where $P^* = \mathrm{argmax}_{P \in E} D(P||Q)$ is the distribution in E that is closest to Q in relative entropy. If the set E is the closure of its interior, then

$$\frac{1}{n} \log \left(\frac{1}{Q^n(E \cap \mathcal{P}_n)} \right) \longrightarrow D(P^*||Q).$$

For the upper bound, note that

$$\sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} \leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P^*||D)} \leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}.$$

When E is the closure of its interior, $E \cap \mathcal{P}_n$ will be nonempty for all $n \geq n_0$ for some n_0, as $\bigcup_n \mathcal{P}_n$ is dense in the set of probabilities. We can find a sequence of distributions P_n such that $P_n \in E \cap \mathcal{P}_n$ and $D(P_n||Q) \longrightarrow D(P^*||Q)$. For each $n \geq n_0$,

$$\sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \geq Q^n(T(P_n)) \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P_n||Q)}.$$

Consequently,

$$\limsup \frac{1}{n} \log \left(\frac{1}{Q^n(E \cap \mathcal{P}_n)} \right) \leq \limsup \left(\frac{|\mathcal{X}| \log(n+1)}{n} + D(P_n||D) \right) = D(P^*||Q).$$

Combining this with the upper bound would establish the result.

References

Chung, K.L. 2001. A Course in Probability Theory, 3rd Edition. Academic Press (C).

Billingsley, P. 1995. Probability and Measure, 3rd Edition. John Wiley & Sons (B).

Cover, T.M. and J.A. Thomas. 2006. Elements of Information Theory, 2nd Edition. Wiley (CT).