

# Chapter 1

## Preliminaries

### 1.1 Probability Axioms

Kolmogorov's Axioms, 1933.

**Definition 1 (Probability Space)** A probability space is a triple  $(\Omega, \mathcal{F}, \Pr)$  where

1.  $\Omega$  is a non-empty set;
2.  $\mathcal{F}$  is a  $\sigma$ -algebra of subsets of  $\Omega$ , that is,
  - (a)  $\Omega \in \mathcal{F}$ ;
  - (b)  $\mathcal{F}$  is closed under complements;
  - (c)  $\mathcal{F}$  is closed under countable unions.
3.  $\Pr$  is a probability measure on  $\mathcal{F}$ , that is,
  - (a)  $\Pr : \mathcal{F} \rightarrow [0, 1]$ ;
  - (b)  $\Pr(\Omega) = 1$ ;
  - (c) If  $\{A_i\}_{i=1}^{\infty} \subset \mathcal{F}$  and  $A_i \cap A_j = \emptyset$  for  $i \neq j$  then

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i).$$

Some examples:

1.  $\Omega = \{1, 2, 3, 4\}$ ,  $\mathcal{F} = 2^{\Omega}$ , that is, the power set of  $\Omega$ . We define  $\Pr$  by defining it on the atoms, so we need only give  $p_k = \Pr(\{k\}) \geq 0$ ,  $\sum_{k=1}^4 p_k = 1$ .
2.  $\Omega = [0, 1]^2$ ,  $\mathcal{F}$  is the Borel subsets of  $\Omega$  (the smallest  $\sigma$ -algebra containing the metric topology of  $\Omega$ ), and  $\Pr$  is Lebesgue measure.

3. Let  $S$  be any finite, non-empty set.  $\Omega = \{0, 1\}^S$ . Each  $\omega \in \Omega$  is a function from  $S$  into  $\{0, 1\}$ , or alternatively, a finite sequence of 0's and 1's of length the cardinality of  $S$ .  $\mathcal{F}$  is again the power set of  $\Omega$ . What are the interesting ways to describe Pr?

a) For each  $A \subset S$  let

$$\mu(A) = \Pr(\{\omega : \omega(x) = 1, x \in A, \omega(x) = 0, x \in A^c\}).$$

The  $\mu(A)$  may be chosen arbitrarily (so long as they are consistent with the axioms) and these will determine Pr.

b)  $\rho(A) = \Pr(\{\omega : \omega(x) = 1, x \in A\})$ .  $\rho$  is called a **correlation function**. For example, we may have

$$\rho(A) = \left(\frac{1}{3}\right)^{\#A}.$$

The question then is does  $\rho$  determine  $\mu$  and does  $\mu$  determine  $\rho$ ? Clearly since  $\mu$  determines Pr then  $\mu$  determines  $\rho$ . In particular,

$$\rho(A) = \sum_{B: B \subset A} \mu(B).$$

But  $\rho$  also determines Pr, as we shall see later, under the **Inclusion - Exclusion Principle**.

Examples 1, 2, and 3 often arise as models of the stationary state of a more complicated model, namely time evolution, either stochastic or deterministic. In more detail:

1. might be the steady state of a random walk on 4 points,  $X = \{1, 2, 3, 4\}$ , that is, the random walk with reflecting barriers at 1 and 4. What then is  $\Omega$  for such a random walk? The points of  $\Omega$  should describe the entire motion or trajectory: each  $\omega$  is a sequence  $\omega = (\omega_0, \omega_1, \omega_2, \dots)$  where  $\omega_j \in X$ . The notation is  $\Omega = X^W$ , where  $W$  is the whole numbers, that is  $W = \{0, 1, 2, \dots\}$ . Note that  $\Omega$  is uncountable.

$\mathcal{F}$  will be the smallest  $\sigma$ -algebra containing all of the (Borel) cylinder sets:

$$\{\omega \in \Omega : \omega_0 = x_0, \dots, \omega_n = x_n\}$$

for all  $n \in W$ .

There are theorems which guarantee that a consistent assignment of probabilities to these cylinder sets determines exactly one probability measure on  $(\Omega, \mathcal{F})$ .

2. may arise from studying the motion of a deterministic particle on the unit square, like a billiard ball, or from random motion on the unit square, like Brownian motion.
3. may arise from studying birth and death evolution on a finite lattice.

## 1.2 Simple Functions and Combinatorial Probability

See Loève.

Suppose we are given a probability space  $(\Omega, \mathcal{F}, \Pr)$ .

**Definition 2 (Indicator of an event)** Let  $A \in \mathcal{F}$ . The **indicator of  $A$** ,  $I_A$ , is the function  $I_A : \Omega \rightarrow \{0, 1\}$  such that

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \in A^c \end{cases}$$

**Proposition 3 (Properties of Indicators)** 1.  $I_{A \cap B}(\omega) = I_A(\omega)I_B(\omega)$ ;

2.  $I_{A \cup B}(\omega) = I_A(\omega) + I_B(\omega) - I_{A \cap B}(\omega)$ ;

3.  $I_{A^c}(\omega) = 1 - I_A(\omega)$ .

**Definition 4 (Partition)** A **partition of  $\Omega$**  is a collection of pairwise disjoint elements of  $\mathcal{F}$  whose union is  $\Omega$ .

**Definition 5 (Simple Function)** Let  $f : \Omega \rightarrow (-\infty, \infty)$ .  $f$  is called a **simple function** if and only if there exists a finite partition  $\{A_j\}_{j=1}^n \subset \mathcal{F}$  and  $\alpha_j \in (-\infty, \infty)$  such that

$$f(\omega) = \sum_{j=1}^n \alpha_j I_{A_j}(\omega)$$

for every  $\omega \in \Omega$ .

Note that a simple function has a finite range. If we let  $R$  be the the range of a the simple function  $f$  then we can write

$$f = \sum_{r \in R} r I_{f^{-1}(r)}.$$

**Definition 6 (Simple random variable)** A **simple random variable** is a simple function defined on a probability space.

**Definition 7 (Expectation of a Simple Function)** The **expected value** of a simple function  $f$ , denoted by  $E[f]$  is

$$E[f] = \sum_{j=1}^n \alpha_j \Pr(A_j).$$

Note that it follows from the remark following the definition of simple function that  $E[f]$  is well-defined, that is, is independent of the partition used to represent  $f$ , since it is clear that

$$\sum_{j=1}^n \alpha_j \Pr(A_j) = \sum_{r \in R} r \Pr(f = r)$$

by appropriate regrouping of the left hand side.

**Theorem 8 (Algebra of Expectations)** *If  $f$  and  $g$  are simple functions and  $\alpha \in (-\infty, \infty)$  then  $f+g$ ,  $f \cdot g$  and  $\alpha f$  are simple functions and*

$$\begin{aligned} \mathbb{E}[f + g] &= \mathbb{E}[f] + \mathbb{E}[g]; \\ \mathbb{E}[\alpha f] &= \alpha \mathbb{E}[f]. \end{aligned}$$

**Proof:** That  $f + g$ ,  $f \cdot g$  and  $\alpha f$  are simple functions is clear. Let  $\{A_j\}_{j=1}^n$  be a partition for  $f$  and let  $\{B_k\}_{k=1}^m$  be a partition for  $g$ , and let  $\alpha_j$  and  $\beta_k$  be the corresponding constants. Then  $\{A_j \cap B_k\}_{j=1, k=1}^{n, m}$  is a partition for both  $f$  and  $g$ . Therefore

$$\begin{aligned} \mathbb{E}[f + g] &= \sum_{j=1}^n \sum_{k=1}^m (\alpha_j + \beta_k) \Pr(A_j \cap B_k) \\ &= \sum_{j=1}^n \sum_{k=1}^m \alpha_j \Pr(A_j \cap B_k) + \sum_{j=1}^n \sum_{k=1}^m \beta_k \Pr(A_j \cap B_k) \\ &= \sum_{j=1}^n \alpha_j \Pr(A_j) + \sum_{k=1}^m \beta_k \Pr(B_k) \\ &= \mathbb{E}[f] + \mathbb{E}[g]. \end{aligned}$$

The formula for  $\mathbb{E}[\alpha f]$  is even easier to prove. **QED**

**Corollary 9** 1.  $\mathbb{E}[I_A] = \Pr(A)$ ;

2.  $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$ ;

This may be extended to  $n$  events as follows. Let  $I_k$  be the indicator of  $A_k$ . Let  $S \subset \{1, 2, \dots, n\}$ . We want

$$\Pr \left( \left( \bigcap_{k \in S} A_k \right) \cap \left( \bigcap_{k \in S^c} A_k^c \right) \right)$$

in terms of

$$\Pr \left( \bigcap_{j \in T} A_j \right)$$

for  $T \subset \{1, 2, \dots, n\}$ .

Consider the following.

$$\prod_{k \in S} I_k \prod_{a \in S^c} (1 - I_a) = \prod_{k \in S} I_k \sum_{T \subset S^c} (-1)^{|T|} \prod_{t \in T} I_t = \sum_{T \subset S^c} (-1)^{|T|} \prod_{t \in T \cup S} I_t. \quad (1.1)$$

We make the convention that

$$\prod_{t \in \emptyset} I_t := 1 = I_\Omega \quad (1.2)$$

which, in view of Corollary 9, forces the convention

$$\bigcap_{t \in \emptyset} A_t := \Omega \quad (1.3)$$

Taking expectations throughout (1.1) we get

$$\Pr\left(\left(\bigcap_{k \in S} A_k\right) \cap \left(\bigcap_{k \in S^c} A_k^c\right)\right) = \sum_{T \subset S^c} (-1)^{|T|} \Pr\left(\bigcap_{t \in T \cup S} A_t\right), \quad (1.4)$$

the **Inclusion-Exclusion formula**. If we take  $S = \emptyset$  in the Inclusion-Exclusion formula and use (1.3) we see

$$\Pr\left(\bigcap_{k=1}^n A_k^c\right) = 1 - \sum_k \Pr(A_k) + \sum_{k \neq t} \Pr(A_k \cap A_t) - \dots$$

Using DeMorgan's Laws, we conclude that

$$\Pr\left(\bigcup_{k=1}^n A_k\right) = \sum_k \Pr(A_k) - \sum_{k \neq t} \Pr(A_k \cap A_t) + \dots,$$

One application is to the **Rencontre Problem**. Let  $\Sigma_n$  be the set of permutations of  $n$  objects. We might as well take these objects to be  $\{1, 2, \dots, n\}$  and  $\Sigma_n = S_n$ . Let  $\Omega = S_n$ , let  $\mathcal{F}$  be the set of all subsets of  $\Omega$ , and suppose that each element of  $\Omega$  has probability  $1/n!$ . This determines  $\Pr$ . A concrete realization is that we have  $n$  persons, and we have a letter for each person. If each person chooses a letter a random, then the event that person  $k$  gets the letter intended for him translates into the event  $A_k = \{\sigma : \sigma(k) = k\}$ . We want to find the probability,  $p_n$ , that at least one person gets the correct letter. We have

$$p_n = \Pr(A_1 \cup A_2 \cup \dots \cup A_n).$$

It is straightforward to compute  $\Pr(A_k)$ ,  $\Pr(A_k \cap A_t)$ , and so on, and we see that

$$\begin{aligned} \Pr(A_1 \cup A_2 \cup \dots \cup A_n) &= \sum_{j=1}^n (-1)^{j+1} \binom{n}{j} \frac{(n-j)!}{n!} \\ &= \sum_{j=1}^n \frac{(-1)^{j+1}}{j!} \end{aligned}$$

which converges to  $1 - e^{-1}$  as  $n \rightarrow \infty$ !

### 1.3 Independence

Suppose that an experiment is repeated  $n$  times.

Let  $n_A$  be the number of outcomes having attribute A. Let  $n_B$  be the number of outcomes having attribute B.

Intuitively A and B are independent if for  $n$  large, the fraction of times A occurs,  $n_A/n$ , equals the fraction of times A occurs where B occurs,  $n_{A \cap B}/n_B$ . Equivalently,

$$\frac{n_A}{n} = \frac{n_{A \cap B}/n}{n_B/n}. \quad (1.5)$$

If we believe in the frequency  $n_A/n$  being approximately  $\Pr(A)$  for large  $n$ , then (1.5) would become

$$\Pr(A) \Pr(B) = \Pr(A \cap B). \quad (1.6)$$

For this reason we make the following definition.

**Definition 10 (Pairwise Independence)** *Two events  $A$  and  $B$  in  $(\Omega, \mathcal{F}, \Pr)$  are said to be **independent** if and only if (1.6) holds.*

**Definition 11 (Mutual Independence)** *Suppose that  $\mathcal{G} \subset \mathcal{F}$ . We say that the elements of  $\mathcal{G}$  are **mutually independent** if for any finite subset  $\{G_1, \dots, G_n\}$  of  $\mathcal{G}$  we have*

$$\Pr(G_1 \cap G_2 \cap \dots \cap G_n) = \prod_{k=1}^n \Pr(G_k).$$

Using these definitions we shall justify the frequency interpretation which suggested this definition. We shall prove a **Weak Law of Large Numbers** and a **Strong Law of Large Numbers**.

First, a comment and an example.

Pairwise independence of several events does not imply mutual independence. For example: Let  $\Omega = \{\text{all sequences } \omega = (\omega_1, \omega_2, \omega_3), \omega_j \in \{-1, +1\}\}$  with  $\Pr(\{\omega : \omega_j = -1\}) = \Pr(\{\omega : \omega_j = +1\}) = 1/2$ ,  $\mathcal{F} = 2^\Omega$ . Let

$$\begin{aligned} A_1 &= \{\omega : \omega_1 = \omega_2\}; \\ A_2 &= \{\omega : \omega_2 = \omega_3\}; \\ A_3 &= \{\omega : \omega_3 = \omega_1\}. \end{aligned}$$

Then

$$\begin{aligned} \frac{1}{2} &= \Pr(A_j) \quad j = 1, 2, 3; \\ \frac{1}{4} &= \Pr(A_i \cap A_j) \quad i \neq j; \\ \frac{1}{8} &\neq \Pr(A_1 \cap A_2 \cap A_3). \end{aligned}$$

**Definition 12 (Independent simple random variables)** *Let  $f$  and  $g$  be two simple random variables,*

$$\begin{aligned} f &= \sum_k b_k I_{B_k}; \\ g &= \sum_j a_j I_{A_j}. \end{aligned}$$

*$f$  and  $g$  are independent if and only if*

$$\Pr(\{\omega : f(\omega) = b_k\} \cap \{\omega : g(\omega) = a_j\}) = \Pr(\{\omega : f(\omega) = b_k\}) \Pr(\{\omega : g(\omega) = a_j\})$$

*for all  $j$  and  $k$ .*

**Definition 13 (Mutually independent simple random variables)** Let  $\{f_\alpha\}$  be a collection of simple random variables defined on  $(\Omega, \mathcal{F}, \Pr)$ . We say that these simple random variables are mutually independent if for any subset  $\{f_1, \dots, f_n\}$  and any set of real numbers  $\{x_1, \dots, x_n\}$  we have

$$\Pr\left(\bigcap_{k=1}^n \{\omega : f_k(\omega) = x_k\}\right) = \prod_{k=1}^n \Pr(\{\omega : f_k(\omega) = x_k\})$$

**Theorem 14 (Chebychev's Inequality)** Let  $X$  be a non-negative simple random variable and let  $b$  be a positive real number. Then

$$\Pr(\{\omega : X(\omega) \geq b\}) \leq \frac{E[X]}{b}.$$

**Proof:** Let  $A_b = \{\omega : X(\omega) \geq b\}$ . Then  $X \geq XI_{A_b}$ . Therefore,  $E[X] \geq E[XI_{A_b}]$ . But  $XI_{A_b} \geq bI_{A_b}$ . Therefore  $E[X] \geq bE[I_{A_b}] = b\Pr(\{\omega : X(\omega) \geq b\})$  as desired. **QED**

Note, once we have defined expectation of arbitrary random variables then the same proof holds with the hypothesis of "simple random variable" deleted and replaced by "random variable".

**Theorem 15 (Weak Law of Large Numbers)** Suppose that  $(\Omega, \mathcal{F}, \Pr)$  is a probability space and  $A_1, A_2, A_3, \dots$  is an infinite sequence of pairwise independent events, each with probability  $p$ . Let  $I_n$  be the indicator of  $A_n$  and let  $S_n = I_1 + \dots + I_n$ . Then for any  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \Pr\left(\left\{\left|\frac{S_n}{n} - p\right| > \epsilon\right\}\right) = 0.$$

**Proof:** First we need the following lemma about simple random variables.

**Lemma 16** If  $f$  and  $g$  are simple random variables which are independent then

$$E[fg] = E[f]E[g].$$

**Proof:** It suffices to consider  $f = I_A$  and  $g = I_B$  where  $A$  and  $B$  are independent events. Then

$$\begin{aligned} E[fg] &= E[I_A I_B] \\ &= E[I_{A \cap B}] \\ &= \Pr(A \cap B) \\ &= \Pr(A) \Pr(B) \\ &= E[f]E[g] \end{aligned}$$

which proves the lemma.

Now to prove the theorem. According to Chebychev's Inequality,

$$\begin{aligned} \Pr\left(\left\{\left|\frac{S_n}{n} - p\right| \geq \epsilon\right\}\right) &\leq \frac{1}{\epsilon^2} E\left[\left(\frac{S_n}{n} - p\right)^2\right] \\ &= \frac{1}{n^2 \epsilon^2} E\left[\left(\sum_{j=1}^n (I_j - p)\right)^2\right] \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n^2 \epsilon^2} \mathbb{E} \left[ \sum_{j=1}^n (I_j - p)^2 \right] \\
&\quad + \frac{2}{n^2 \epsilon^2} \mathbb{E} \left[ \sum_{j=1}^n \sum_{k=1}^{j-1} (I_j - p)(I_k - p) \right] \\
&= \frac{1}{n^2 \epsilon^2} \sum_{j=1}^n \mathbb{E}[(I_j - p)^2] \\
&\quad + \frac{2}{n^2 \epsilon^2} \sum_{j=1}^n \sum_{k=1}^{j-1} \mathbb{E}[I_j - p] \mathbb{E}[I_k - p] \\
&= \frac{1}{n^2 \epsilon^2} np(1-p)
\end{aligned}$$

which converges to 0 as  $n \rightarrow \infty$ . **QED**

**Theorem 17 (Poisson Approximation to the Binomial Probability Law)** *Suppose that  $r \rightarrow \infty$ ,  $p \rightarrow 0$  and  $rp \rightarrow \lambda > 0$ , and  $q = 1 - p$ . Then*

$$\binom{r}{k} p^k q^{r-k} \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}.$$

**Proof:**

$$\binom{r}{k} p^k (1-p)^{r-k} = \frac{1}{k!} \frac{r!}{r^k (r-k)!} (rp)^k (1-p)^{-k} \left(1 - \frac{rp}{r}\right)^r.$$

Since  $k$  is fixed,

$$\begin{aligned}
\lim_{p \rightarrow 0} (1-p)^{-k} &= 1, \\
\lim_{r \rightarrow \infty} \frac{r!}{r^k (r-k)!} &= 1,
\end{aligned}$$

while

$$(1-p)^r \rightarrow e^{-\lambda}$$

under the conditions on  $p$  and  $r$ . **QED**

**Theorem 18 [Weierstrass Approximation Theorem]** *Let  $f$  be a continuous function on  $[0, 1]$ . Then there exists a sequence of polynomials,  $P_n(x)$  such that*

$$\lim_{n \rightarrow \infty} \left( \max_{0 \leq x \leq 1} |f(x) - P_n(x)| \right) = 0,$$

that is,  $P_n$  converges to  $f$  uniformly on  $[0, 1]$ .

**Proof:** We shall prove this by constructing the polynomials explicitly, using **Bernstein polynomials**. Choose  $f$  and let

$$P_n(x) = \sum_{k=0}^n \binom{n}{k} x^k (1-x)^{n-k} f(k/n).$$

Consider the binomial probability model with probability of success equal to  $x$ , and  $n$  trials. Let  $A_k$  be the event of a success on the  $k^{\text{th}}$  trial, and let  $I_k$  be the indicator of  $A_k$ . Then  $S_n = \sum_{k=1}^n I_k$  is the number of successes in  $n$  trials.

$$\Pr(\{S_n = k\}) = \binom{n}{k} x^k (1-x)^{n-k}$$

and so  $P_n(x) = E[f(S_n/n)]$ . Therefore

$$\begin{aligned} 0 &\leq |P_n(x) - f(x)| \\ &= |E[f(S_n/n) - f(x)]| \\ &\leq E[|f(S_n/n) - f(x)|]. \end{aligned}$$

For any simple random variable  $g$  and any event  $A$  let  $E[g; A] = E[gI_A]$ . Then we may write

$$\begin{aligned} 0 &\leq |P_n(x) - f(x)| \\ &\leq E[|f(S_n/n) - f(x)|; \{|n^{-1}S_n - x| < \epsilon\}] + E[|f(S_n/n) - f(x)|; \{|n^{-1}S_n - x| \geq \epsilon\}]. \end{aligned}$$

Now let  $\delta(\epsilon) = \max\{|f(t) - f(s)| : 0 \leq s \leq t \leq 1, |t - s| < \epsilon\}$ .  $\delta(\epsilon) \rightarrow 0$  as  $\epsilon \rightarrow 0$  by the uniform continuity of  $f$  on  $[0, 1]$ . Let  $M = \max_{[0,1]} |f|$ . Then we may continue with

$$\begin{aligned} |P_n(x) - f(x)| &\leq \delta(\epsilon) + 2M \Pr(\{|n^{-1}S_n - x| \geq \epsilon\}) \\ &\leq \delta(\epsilon) + 2M \frac{x(1-x)}{n\epsilon^2} \\ &\leq \delta(\epsilon) + \frac{2M}{4n\epsilon^2} \end{aligned}$$

which is independent of  $x$ . So for all  $\epsilon > 0$ ,

$$\limsup_{n \rightarrow \infty} \left( \max_{0 \leq x \leq 1} |f(x) - P_n(x)| \right) \leq \delta(\epsilon),$$

so

$$\limsup_{n \rightarrow \infty} \left( \max_{0 \leq x \leq 1} |f(x) - P_n(x)| \right) = 0,$$

which proves the theorem. **QED**

(See Feller, Volume I for experimental results on 10,000 tosses of a fair coin.)

## 1.4 Some Useful Facts About Any Probability Space

**Proposition 19** *If  $\{A_k\}_{k=1}^{\infty}$  is any sequence of events in  $\mathcal{F}$  then*

$$\Pr\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \Pr(A_k).$$

**Proof:** Let  $A = A_1 \cup A_2 \cup \dots$ . Then  $I_A \leq I_{A_1} + I_{A_2} + \dots$ . Take the expected value of both sides. **QED**

**Proposition 20 (Continuity of Pr)** If  $\{A_k\}_{k=1}^{\infty}$  is a monotone increasing sequence of events in  $\mathcal{F}$ , define  $\lim_{k \rightarrow \infty} A_k = \bigcup_{k=1}^{\infty} A_k$ . If  $\{A_k\}_{k=1}^{\infty}$  is a monotone decreasing sequence of events in  $\mathcal{F}$ , define  $\lim_{k \rightarrow \infty} A_k = \bigcap_{k=1}^{\infty} A_k$ . In either case,

$$\lim_{k \rightarrow \infty} \Pr(A_k) = \Pr(\lim_{k \rightarrow \infty} A_k).$$

**Proof:** Suppose the sets are monotone increasing. Put  $B_1 = A_1$  and  $B_k = A_k \cap A_{k-1}^c$  for  $k \geq 2$ . Note that the  $B_k$  are pairwise disjoint events,

$$\begin{aligned} A_k &= \bigcup_{n=1}^k B_n, \\ \lim_{k \rightarrow \infty} A_k &= \bigcup_{k=1}^{\infty} B_k. \end{aligned}$$

Therefore

$$\begin{aligned} \lim_{k \rightarrow \infty} \Pr(A_k) &= \lim_{k \rightarrow \infty} \Pr\left(\bigcup_{n=1}^k B_n\right) \\ &= \lim_{k \rightarrow \infty} \sum_{n=1}^k \Pr(B_n) \\ &= \sum_{n=1}^{\infty} \Pr(B_n) \\ &= \Pr\left(\bigcup_{n=1}^{\infty} B_n\right) \\ &= \Pr\left(\lim_{k \rightarrow \infty} A_k\right) \end{aligned}$$

as desired. For decreasing sequences, use Boole's identities. **QED**

**Definition 21 (Infinitely often)** If  $\{A_n\}_{n=1}^{\infty}$  is any sequence of events in  $\mathcal{F}$  one defines

$$\limsup_{n \rightarrow \infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

Note that  $\omega \in \limsup_{n \rightarrow \infty} A_n$  if and only if  $\omega \in A_k$  for infinitely many  $k$ . This motivates the notation  $A_n$  **infinitely often**, written  $A_n \text{i.o.}$ , for  $\limsup_{n \rightarrow \infty} A_n$ . This is most frequently seen as  $\Pr(A_n \text{i.o.})$ . (The *liminf* of  $A_n$  is defined as

$$\liminf_{n \rightarrow \infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k.$$

This is **not**  $A_n$  *finitely often*!)

An example of limsup is found in simple random walk on  $Z^d$ . Let  $A_n$  be the event of being at the starting point after  $n$  moves. Then as we shall show later on,

$$\Pr(A_n \text{i.o.}) = \begin{cases} 1 & \text{if } d \in 1, 2 \\ 0 & \text{if } d \geq 3 \end{cases}$$

**Theorem 22 (Borel-Cantelli Lemma)** Let  $\{A_n\}_{n=1}^\infty$  be an arbitrary sequence of events in  $(\Omega, \mathcal{F}, \Pr)$ . Then

1. If

$$\sum_{n=1}^{\infty} \Pr(A_n) < \infty$$

then  $\Pr(A_n \text{ i.o.}) = 0$ .

2. If

$$\sum_{n=1}^{\infty} \Pr(A_n) = \infty$$

and the events  $A_j$  are mutually independent then  $\Pr(A_n \text{ i.o.}) = 1$ .

**Proof:**

1.

$$\begin{aligned} \Pr(A_n \text{ i.o.}) &= \Pr\left(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\right) \\ &= \lim_{n \rightarrow \infty} \Pr\left(\bigcup_{k=n}^{\infty} A_k\right) \text{ by continuity} \\ &\leq \lim_{n \rightarrow \infty} \sum_{k=n}^{\infty} \Pr(A_k) \text{ by Proposition 19} \\ &= 0 \text{ by hypothesis.} \end{aligned}$$

2. From calculus we know that  $1 - x \leq \exp(-x)$  for all  $x \in (-\infty, \infty)$ . From the continuity of probability measures we know that

$$\Pr(A_n \text{ i.o.}) = \lim_{n \rightarrow \infty} \left( \lim_{m \rightarrow \infty} \Pr\left(\bigcup_{k=n}^m A_k\right) \right).$$

Let  $A_n \text{ f.o.} = (A_n \text{ i.o.})^c$ . From Boole's identities we see that

$$\begin{aligned} \Pr(A_n \text{ f.o.}) &= \lim_{n \rightarrow \infty} \left( \lim_{m \rightarrow \infty} \Pr\left(\bigcap_{k=n}^m A_k^c\right) \right) \\ &= \lim_{n \rightarrow \infty} \left( \lim_{m \rightarrow \infty} \prod_{k=n}^m \Pr(A_k^c) \right) \\ &\quad \text{since } A_k \text{ ind. iff } A_k^c \text{ ind.} \\ &= \lim_{n \rightarrow \infty} \left( \lim_{m \rightarrow \infty} \prod_{k=n}^m (1 - \Pr(A_k)) \right) \\ &\leq \lim_{n \rightarrow \infty} \left( \lim_{m \rightarrow \infty} \prod_{k=n}^m \exp(-\Pr(A_k)) \right) \\ &= 0 \text{ by hypothesis.} \end{aligned}$$

Therefore  $\Pr(A_n \text{ i.o.}) = 1$ . **QED**

There is an intimate connection between “limits” and “infinitely often”. Let  $\{X_n\}_{n=1}^\infty$  be a sequence of random variables defined on  $(\Omega, \mathcal{F}, \Pr)$ , and let  $A_n = \{\omega : |X_n(\omega)| \geq \epsilon\}$ . Then for each  $\omega$ ,  $X_n(\omega) \rightarrow 0$  as  $n \rightarrow \infty$  if and only if finitely many of the  $A_n(\epsilon)$  occur for any  $\epsilon > 0$ . Therefore,

$$\Pr(\{\omega : X_n(\omega) \rightarrow 0\}) = 1 \text{ iff } \Pr(A_n(\epsilon)\text{i.o.}) = 0 \forall \epsilon > 0.$$

So, the Borel-Cantelli Lemma can be used to prove theorems about limits of sequences of random variables.

For example, now we can prove a version of the Strong Law of Large Numbers.

**Theorem 23 (Strong Law of Large Numbers for Events)** *Let  $(\Omega, \mathcal{F}, \Pr)$  be a probability space and let  $\{A_k\}_{k=1}^\infty$  be a sequence of mutually independent events in  $\mathcal{F}$ . Let  $I_k$  be the indicator of  $A_k$ , and let  $S_n = I_1 + \cdots + I_n$ . Suppose that  $\Pr(A_k) = p$  for all  $k$ . Then*

$$\Pr(\{\omega : \lim_{n \rightarrow \infty} n^{-1} S_n(\omega) = p\}) = 1.$$

**Proof:** Choose  $\epsilon > 0$  and let  $A_n(\epsilon) = \{\omega : |n^{-1} S_n - p| \geq \epsilon\}$ . If we can show that

$$\sum_{n=1}^{\infty} \Pr(A_n(\epsilon)) < \infty$$

then we are done. Now from Chebychev’s Inequality we have

$$\Pr(A_n(\epsilon)) \leq \frac{p(1-p)}{n\epsilon^2},$$

but

$$\sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

so we must try something else.

Observe that

$$\Pr(A_{n^2}(\epsilon)) \leq \frac{p(1-p)}{n^2\epsilon^2},$$

and

$$\sum_{n=1}^{\infty} \frac{1}{n^2} < \infty,$$

so by the first part of the Borel-Cantelli Lemma we have proven that

$$\Pr(\{\omega : \lim_{n \rightarrow \infty} n^{-2} (S_{n^2}(\omega) - n^2 p) = 0\}) = 1.$$

Now we will show that for any  $\omega$  for which  $n^{-2} S_{n^2}(\omega) \rightarrow p$  we have  $n^{-1} S_n(\omega) \rightarrow p$ . Let  $k$  be a positive integer. Let  $n_k$  be the unique positive integer such that  $n_k^2 \leq k < (n_k + 1)^2$ . Then

$$\frac{S_{n_k^2}}{n_k} \leq \frac{S_k}{k} \leq \frac{S_{n_k^2} + 2n_k + 1}{k},$$

so that we have

$$\frac{S_{n_k^2} n_k^2}{(n_k + 1)^2 n_k^2} \leq \frac{S_k}{k} \leq \frac{S_{n_k^2}}{n_k^2} + \frac{2n_k + 1}{n_k^2}.$$

Now,  $n_k \rightarrow \infty$  as  $k \rightarrow \infty$ , so the result follows from the pinching theorem. **QED**

## 1.5 Random Digits:

Emile Borel (1909) showed that the digits in the binary expansion of a number picked at random from  $[0, 1]$  are independent random variables. More precisely, if we let

- $\Omega = [0, 1]$ ;
- $\Omega' = \Omega - \{x \in [0, 1] : x = k/2^n, k, n \in \mathbb{Z}\}$ ;
- $\mathcal{F}$  is the Borel subsets of  $\Omega$ ;
- $\text{Pr}$  is Lebesgue measure on  $\Omega$ ;

then every  $\omega \in \Omega'$  has a unique binary representation:

$$\omega = \sum_{k=1}^{\infty} \frac{\omega_k}{2^k}$$

and  $\text{Pr}(\Omega') = 1$ . For each  $\omega \in \Omega'$  let  $I_k(\omega) = \omega_k$ . For the other  $\omega$ , set  $I_k(\omega) = 0$ .

**Theorem 24 (Borel, 1909)** *The  $I_k$  are mutually independent random variables with  $\mathbb{E}[I_k] = 1/2$ . Equivalently, if  $A_k = \{\omega : I_k(\omega) = 1\}$ , then the  $A_k$  are mutually independent events, each with probability  $1/2$ .*

**Proof:** Since the complement of  $\Omega'$  has probability zero, we ignore it. The theorem then follows by dividing  $[0, 1]$  into halves, quarters, eighths, etc, and computing lengths of intervals. **QED**

## Chapter 2

# General Theory

### 2.1 Simple Applications of Carathèodory's Extension Theorem

**Theorem 25 (Carathèodory's Extension Theorem)** *Let  $\mu$  be a measure on a field  $\mathcal{F}_0$  of subsets of  $\Omega$ , and suppose that  $\mu$  is sigma-finite on  $\mathcal{F}_0$ . Then  $\mu$  has a unique extension to a measure on  $\sigma(\mathcal{F}_0)$ .*

**Proof:** See Ash, *Real Analysis and Probability*, pages 19-20. **QED**

Let  $\Omega = (-\infty, \infty)$  and let  $\mathcal{F}_0$  be the set of all finite unions of intervals. Then we can extend to Lebesgue measure the measure  $\mu$  defined by  $\mu([a, b]) = b - a$ .

More generally, we can do the following:

**Theorem 26 (Distribution Functions and Measures)** *Let  $F : (-\infty, \infty) \rightarrow (-\infty, \infty)$  be such that*

1.  $F$  is non-decreasing;
2.  $F$  is right continuous;
3.  $F(\infty) = 1$ ;  $F(-\infty) = 0$ .

*Let  $\mathcal{L}$  be the algebra of all intervals of the form  $(a, b]$ ,  $(-\infty, b]$ ,  $(a, \infty)$  and  $(-\infty, \infty)$ , and define  $\mu((a, b]) = F(b) - F(a)$ ,  $\mu((-\infty, b]) = F(b)$ ,  $\mu((a, \infty)) = 1 - F(a)$  and  $\mu((-\infty, \infty)) = 1$ . Then there is a unique extension of  $\mu$  to a probability measure on the Borel subsets of the real line.*

**Remark:** The crux of the proof is to show that  $\mu$  is countably additive so that the Carathèodory's Extension Theorem applies.

**Proof:** Let  $I = (a, b]$ . It suffices to show that for  $I_k = (a_k, b_k]$  such that

$$I = \bigcup_{k=1}^{\infty} I_k$$

that

$$F(b) - F(a) \leq \sum_{k=1}^{\infty} (F(b_k) - F(a_k)). \tag{2.1}$$

Since  $F$  is right continuous, given  $\epsilon > 0$  we can find  $b'_k > b_k$  such that  $F(b'_k) - F(b_k) < 2^{-k}\epsilon$  for each  $k$ . Let  $G_k = (a_k, b'_k)$  and  $I'_k = (a_k, b'_k]$ . For any  $\delta > 0$  the sets  $G_k$  are an open cover of the compact set  $[a + \delta, b]$ , so for each  $\delta$  there is a finite sub-cover such that

$$[a + \delta, b] \subset \bigcup_{k=1}^N G_k \subset \bigcup_{k=1}^N I'_k.$$

Therefore,

$$\begin{aligned} F(b) - F(a + \delta) &\leq \sum_{k=1}^N (F(b'_k) - F(a_k)) \\ &\leq \sum_{k=1}^{\infty} (F(b_k) - F(a_k)) + \sum_{k=1}^{\infty} (F(b'_k) - F(b_k)) \\ &\leq \sum_{k=1}^{\infty} (F(b_k) - F(a_k)) + \epsilon. \end{aligned}$$

Letting  $\delta \rightarrow 0^+$  and using the right-continuity of  $F$  we see that

$$F(b) - F(a) \leq \sum_{k=1}^{\infty} (F(b_k) - F(a_k)) + \epsilon.$$

Since  $\epsilon$  is arbitrary, (2.1) holds. **QED**

## 2.2 Random variables:

**Lemma 27** *Let  $X : \Omega_1 \rightarrow \Omega_2$  be a function, and let  $\mathcal{B}$  be a sigma algebra in  $\Omega_1$ , and let  $\mathcal{A}$  be a collection of subsets of  $\Omega_2$ . If  $X^{-1}(A) \in \mathcal{B}$  for every  $A \in \mathcal{A}$  then  $X^{-1}(A) \in \mathcal{B}$  for every  $A \in \sigma(\mathcal{A})$ . More generally, if  $\mathcal{A}$  is any collection of subsets of  $\Omega_2$  then  $X^{-1}(\sigma(\mathcal{A})) = \sigma(X^{-1}(\mathcal{A}))$ .*

**Proof:** The lemma follows from the observation that inverse mappings commute with all set operations. **QED**

**Definition 28 (Abstract Random Variables)** *Let  $(\Omega_1, \mathcal{F}_1)$  and  $(\Omega_2, \mathcal{F}_2)$  be measure spaces. Let  $X : (\Omega_1, \mathcal{F}_1) \rightarrow (\Omega_2, \mathcal{F}_2)$ .  $X$  is said to be an **(abstract) random variable** (measurable function) if and only if  $X^{-1}(A) \in \mathcal{F}_1$  for all  $A \in \mathcal{F}_2$ .  $X$  is said to be a real random variable if and only if  $(\Omega_2, \mathcal{F}_2) = ((-\infty, \infty), \mathcal{B}(-\infty, \infty))$ .*

We can see from the preceding lemma that for real valued random variables it suffices to assume that  $X^{-1}((-\infty, t]) \in \mathcal{F}_1$  for all real  $t$ . See Rudin *Real and Complex Analysis* Chapter 1.

**Theorem 29 (Properties of Distribution Functions)** *Let  $X$  be a real valued random variable on  $(\Omega, \mathcal{F}, \Pr)$  and define*

$$F(x) = \Pr(\{\omega : X(\omega) \leq x\}).$$

*Then  $F : (-\infty, \infty) \rightarrow [0, 1]$  and*

1.  $F$  is non-decreasing;
2.  $F$  is right continuous;
3.  $F(-\infty) = 0$  and  $F(\infty) = 1$ .

Furthermore, if there is some function  $F : (-\infty, \infty) \rightarrow [0, 1]$  with the preceding three properties then there is some probability space  $(\Omega, \mathcal{F}, \Pr)$  and a real valued random variable  $X$  defined on  $(\Omega, \mathcal{F}, \Pr)$  so that for every real number  $x$ ,  $F(x) = \Pr(\{\omega : X(\omega) \leq x\})$ .

**Proof:** Suppose  $X$  is given. Clearly  $F : (-\infty, \infty) \rightarrow [0, 1]$ .  $F$  is non-decreasing because  $A \subset B$  implies  $\Pr(A) \leq \Pr(B)$  for any events  $A$  and  $B$ .  $F$  is right continuous and has the indicated limits because probability measures are continuous. Then  $\leq$  ensures right-continuity. If  $\leq$  were replaced with  $<$  then  $F$  would be left continuous.

Suppose  $F$  is given. Apply Theorem 26 to construct  $(\Omega, \mathcal{F}, \Pr)$  and let  $X(\omega) = \omega$ . **QED**

Note that we may just as well have vector valued random variables

$$X : (\Omega, \mathcal{F}) \rightarrow (R^d, \mathcal{B}(R^d)),$$

or complex valued random variables.

Also, suppose that  $X : \Omega_1 \rightarrow \Omega_2$  and that  $(\Omega_1, \mathcal{F}_1, \Pr_1)$  is a probability space. We can induce a probability structure on  $\Omega_2$  by taking

- $\mathcal{F}_2 = \{B \subset \Omega_2 : X^{-1}(B) \in \mathcal{F}_1\}$ ;
- $\Pr_2(B) = \Pr_1(X^{-1}(B))$ .

**Definition 30 (Distribution Function)**  $F : (-\infty, \infty) \rightarrow [0, 1]$  is called a **distribution function** if and only if

1.  $F$  is non-decreasing;
2.  $F$  is right continuous;
3.  $F(-\infty) = 0$  and  $F(\infty) = 1$ .

**Definition 31 (Expected Value)** Let  $(\Omega, \mathcal{F}, \Pr)$  be a probability space and let  $f \in \mathcal{L}^1(\Omega, \mathcal{F}, \Pr)$ . We define the expected value of  $f$ , denoted  $E[f]$  by

$$E[f] = \int_{\Omega} f(\omega) d\Pr(\omega).$$

Note that this definition is consistent with the definition given earlier for simple random variables.

Some useful notation: If  $A \in \mathcal{F}$  then

$$E[f; A] \equiv E[fI_A] = \int_A f(\omega) \Pr(\omega).$$

An example: Let  $\Omega = (-\infty, \infty)$ ,  $\mathcal{F} = \mathcal{B}$  and  $\mu_F$  is a probability measure determined by the distribution function  $F$ . Then

$$\int_{(-\infty, \infty)} f(x) d\mu_F(x) \equiv \int_{-\infty}^{\infty} f(x) dF(x)$$

is called the Lebesgue-Stieltjes integral of  $f$  with respect to  $F$ .

## 2.3 Changing variables

Consider the following:

$\Phi : (\Omega_1, \mathcal{F}_1, \Pr) \rightarrow (\Omega_2, \mathcal{F}_2, \Pr_2)$  and  $f : (\Omega_2, \mathcal{F}_2, \Pr_2) \rightarrow ((-\infty, \infty), \mathcal{B})$  where the probability structure on  $\Omega_2$  is induced by  $\Phi$ . If  $f$  is measurable with respect to  $\mathcal{F}_2$  then

$$\int_{\Omega_2} f(\omega_2) d\Pr_2(\omega_2) = \int_{\Omega_1} f \circ \Phi(\omega_1) d\Pr_1(\omega_1)$$

in the sense that either both sides exist and are equal or neither side exists.

**Definition 32 (Distribution function of a random variable)** We say that random variable  $X$  on  $(\Omega, \mathcal{F}, \Pr)$  has distribution function  $F$  if and only if for all real  $x$ ,  $F(x) = \Pr(\{\omega : X(\omega) \leq x\})$ .

**Definition 33 ( $k^{\text{th}}$  moment)** The  $k^{\text{th}}$  moment of a random variable  $X$  on  $(\Omega, \mathcal{F}, \Pr)$  with distribution function  $F$ , denoted  $E[X^k]$ , is

$$E[X^k] = \int_{-\infty}^{\infty} x^k dF(x) = \int_{\Omega} X^k(\omega) d\Pr(\omega)$$

provided

$$\int_{-\infty}^{\infty} |x|^k dF(x) < \infty.$$

Recall from Jensen's Inequality that if  $\Phi$  is a convex function then  $E[\Phi(X)] \geq \Phi(E[X])$ .

**Definition 34 (Characteristic Function)** The characteristic function of a random variable  $X$  with distribution function  $F$ , denoted  $\phi_X(\lambda)$ , is defined by

$$\phi_X(\lambda) = \int_{-\infty}^{\infty} e^{i\lambda x} dF(x) = E[e^{i\lambda X}],$$

for all  $\lambda \in (-\infty, \infty)$ .

**Theorem 35** The characteristic function of any random variable is a uniformly continuous function.

**Proof:** Let  $X$  be a random variable on  $(\Omega, \mathcal{F}, \Pr)$  with distribution function  $F$ . Let  $\phi_X = \phi$  and let  $h$  be a real number. Then

$$\begin{aligned} \phi(\lambda + h) - \phi(\lambda) &= E[e^{i(\lambda+h)X} - e^{i\lambda X}] \\ &= E[e^{i\lambda X} (e^{ihX} - 1)], \end{aligned}$$

so

$$\begin{aligned} |\phi(\lambda + h) - \phi(\lambda)| &\leq E[|e^{i(\lambda+h)X} - e^{i\lambda X}|] \\ &= E[|e^{ihX} - 1|], \end{aligned}$$

which is independent of  $\lambda$ . Since  $|e^{ihX} - 1| \leq 2$  and we are in a probability space we can apply Lebesgue's Dominated Convergence Theorem and the pinching theorem to conclude that

$$\begin{aligned} 0 &\leq \lim_{h \rightarrow 0} |\phi(\lambda + h) - \phi(\lambda)| \\ &\leq \lim_{h \rightarrow 0} \mathbf{E}[|e^{ihX} - 1|] \\ &\leq \mathbf{E}[\lim_{h \rightarrow 0} |e^{ihX} - 1|] \\ &= 0, \end{aligned}$$

independent of  $\lambda$ . **QED**

**Theorem 36** *Let  $X$  be a non-negative random variable on  $(\Omega, \mathcal{F}, \Pr)$  with distribution function  $F$  and let  $r > 0$ . Then*

$$\mathbf{E}[X^r] = \int_0^\infty r x^{r-1} (1 - F(x)) dx.$$

**Proof:** We could use integration by parts for Riemann-Stieltjes integrals or Fubini's theorem. We shall use that latter (see Theorem 67):

$$\begin{aligned} \int_0^\infty r x^{r-1} (1 - F(x)) dx &= \int_0^\infty r x^{r-1} \Pr(\{\omega : X(\omega) > x\}) dx \\ &= \int_0^\infty r x^{r-1} \mathbf{E}[I_{\{\omega : X(\omega) > x\}}] dx \\ &= \mathbf{E}\left[\int_0^\infty r x^{r-1} I_{\{\omega : X(\omega) > x\}} dx\right] \\ &= \mathbf{E}\left[\int_0^{X(\omega)} r x^{r-1} dx\right] \\ &= \mathbf{E}[X^r] \end{aligned}$$

Remember that Fubini's Theorem works so long as  $I_{\{\omega : X(\omega) > x\}}$  is measurable with respect to  $(-\infty, \infty) \times \Omega$ . We also need to be a little careful about the case  $0 < r < 1$  with convergence near to 0. **QED**

## 2.4 Independence: General Case

Let  $X_1$  and  $X_2$  be two random variables on  $(\Omega, \mathcal{F}, \Pr)$ .

**Definition 37 (Joint Distribution Function)** *The joint distribution function of  $X_1$  and  $X_2$ , denoted  $F(x_1, x_2)$  is defined by*

$$F(x_1, x_2) = \Pr(\{\omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2\})$$

*for all ordered pairs  $(x_1, x_2)$  of real numbers. The distribution functions of  $X_1$  and  $X_2$  separately are sometimes called the **marginal distribution functions** of  $F$ .*

There are two commonly used definitions of what it means for  $X_1$  and  $X_2$  to be independent random variables, and there is a theorem which says that these two definitions are equivalent.

**Definition 38** Let  $F$  be the joint distribution function of  $X_1$  and  $X_2$ , and let  $F_1$  and  $F_2$  be the associated marginal distribution functions. We say that  $X_1$  and  $X_2$  are **independent** if

$$F(x_1, x_2) = F_1(x_1)F_2(x_2)$$

for all ordered pairs  $(x_1, x_2)$  of real numbers.

**Definition 39**  $X_1$  and  $X_2$  are **independent random variables** if and only if

$$\Pr(\{\omega : X_1(\omega) \in A, X_2(\omega) \in B\}) = \Pr(\{\omega : X_1(\omega) \in A\}) \Pr(\{\omega : X_2(\omega) \in B\})$$

for all pairs of  $A$  and  $B$  of Borel subsets of  $(-\infty, \infty)$ .

**Theorem 40** These two definitions of independence are equivalent.

**Proof:** It is clear that Definition 39 implies Definition 38. For the reverse implication we use an extension argument.

Suppose that  $X_1$  and  $X_2$  satisfy the conditions of Definition 38. Fix a real number  $t_1$ . Then

$$\Pr(\{\omega : X_1(\omega) \leq t_1, X_2(\omega) \in B\}) = \Pr(\{\omega : X_1(\omega) \leq t_1\}) \Pr(\{\omega : X_2(\omega) \in B\}) \quad (2.2)$$

for all half lines  $B$ . Then (2.2) holds for all finite intersections and differences of half lines for any fixed value of  $t_1$ , so (2.2) holds in the semi-algebra of intervals of the form  $(a, b]$ , and therefore, in the field of finite unions of elements of this semi-algebra. If  $\Pr(\{\omega : X_1(\omega) \leq t_1\}) = 0$  there is nothing left to prove. If not, divide both sides of (2.2) by  $\Pr(\{\omega : X_1(\omega) \leq t_1\})$  and then the condition of Definition 39 holds for any half line  $A$  from the Carathéodory Extension Theorem. Now fix any Borel set  $B$  and repeat the entire process. **QED**

There is still a third possible definition of independence. It is more elegant to define independence of  $\sigma$ -fields, which we do as follows.

**Definition 41 (Independence of  $\sigma$ -fields)** Let  $(\Omega, \mathcal{F}, \Pr)$  be a probability space and let  $\mathcal{F}_1$  and  $\mathcal{F}_2$  be subset of  $\mathcal{F}$  which are also sigma algebras on  $\Omega$ . We say that  $\mathcal{F}_1$  and  $\mathcal{F}_2$  are **independent** if and only if

$$\Pr(A \cap B) = \Pr(A) \Pr(B)$$

for all  $A \in \mathcal{F}_1$  and  $B \in \mathcal{F}_2$ .

**Definition 42** Let  $X : \Omega \rightarrow (-\infty, \infty)$ .  $\sigma(X)$  denotes the smallest  $\sigma$ -field on  $\Omega$  in which  $X$  is measurable. That is

$$\sigma(X) = \{A \subset \Omega : X^{-1}(B) = A \text{ for some } B \in \mathcal{B}(R)\}.$$

It is clear then that if  $X_1$  and  $X_2$  are random variables on  $(\Omega, \mathcal{F}, \Pr)$  then  $X_1$  and  $X_2$  are independent if and only if  $\sigma(X_1)$  and  $\sigma(X_2)$  are independent.

**Definition 43 (Mutual independence)** Let  $\{\mathcal{F}_j\}_{j=1}^n$  be a set of sub- $\sigma$  fields of  $(\Omega, \mathcal{F}, \Pr)$ . We say that  $\{\mathcal{F}_j\}_{j=1}^n$  are **(mutually) independent** if and only if

$$\Pr\left(\bigcap_{j=1}^n A_j\right) = \prod_{j=1}^n \Pr(A_j)$$

for all  $A_j \in \mathcal{F}^j$ . If  $\{\mathcal{F}_j\}$  is an infinite collection of sub- $\sigma$  fields we say that these  $\sigma$ -fields are independent if every finite subset of them is independent.

**Theorem 44** Suppose that  $\mathcal{F}_0, \mathcal{F}_1, \dots$  is a infinite sequence of independent sub- $\sigma$  algebras of  $(\Omega, \mathcal{F}, \Pr)$ . Let  $\mathcal{G}$  be the smallest  $\sigma$  algebra containing each of the  $\mathcal{F}_j$  for  $j \geq 1$ . Then  $\mathcal{G}$  and  $\mathcal{F}_0$  are independent.

**Proof:** Let  $A \in \mathcal{F}_0$  and  $B \in \mathcal{G}$ . If  $B = \bigcap_{k \geq 1} A_k$  and  $A_k \in \mathcal{F}_k$  and only finitely many of the  $A_k$  are different from  $\Omega$  then it is clear that

$$\Pr(A \cap B) = \Pr(A) \Pr(B). \quad (2.3)$$

The collection of all such  $B$  is a semi-algebra, and finite unions of such  $B$  form a field. These unions may always be taken as disjoint unions, so (2.3) holds for all  $B$  in this field. So by the Carathéodory Extension Theorem, (2.3) holds for all  $B \in \mathcal{G}$ , which is what we needed to show. **QED**

Some notation. If  $\{\mathcal{F}_i, i \in I\}$  is a collection of sub-sigma algebras in  $(\Omega, \mathcal{F}, \Pr)$ . We let

$$\bigvee_{i \in I} \mathcal{F}_i$$

denote the smallest sigma algebra containing each of the  $\mathcal{F}_i$ .

**Corollary 45** Suppose that  $\mathcal{F}_0, \mathcal{F}_1, \dots$  is a infinite sequence of independent sub- $\sigma$  algebras of  $(\Omega, \mathcal{F}, \Pr)$  and  $I$  and  $J$  are disjoint subsets of the non-negative integers. If  $I$  is finite then the sigma algebras  $\{\mathcal{F}_i, i \in I\}$  and  $\bigvee_{j \in J} \mathcal{F}_j$  are independent sigma algebras.

**Theorem 46** Suppose that  $\mathcal{F}_0, \mathcal{F}_1, \dots$  is a infinite sequence of independent sub- $\sigma$  algebras of  $(\Omega, \mathcal{F}, \Pr)$  and  $I$  and  $J$  are disjoint subsets of the non-negative integers. Then  $\mathcal{H} \equiv \bigvee_{j \in J} \mathcal{F}_j$  and  $\mathcal{G} \equiv \bigvee_{i \in I} \mathcal{F}_i$  are independent sigma algebras.

**Proof:** By the preceding corollary,  $\mathcal{H}$  and any finite number of the  $\mathcal{F}_i, i \in I$  are independent. Hence by the definition of mutual independence of sigma algebras,  $\mathcal{H}$  and  $\{\mathcal{F}_i, i \in I\}$  are independent. The theorem now follows from Theorem 44. **QED**

**Corollary 47** Under the hypotheses of Theorem 46, if  $X_1$  and  $X_2$  are measurable with respect to  $\mathcal{G}$  and  $\mathcal{H}$  respectively, then  $X_1$  and  $X_2$  are independent.

As an application,  $X$  might be a Borel measurable function of the independent random variables  $X_i$ , each measurable on  $\mathcal{F}_i, i \in I$ , and  $Y$  a Borel measurable function of the independent random variables  $Y_j$  measurable on the  $\mathcal{F}_j, j \in J$ . Then  $X$  and  $Y$  are independent random variables.

## 2.5 Two Important Theorems

**Theorem 48 (Expected value of a product)** Let  $X_1, X_2, \dots, X_n$  be mutually independent random variables for which  $E[X_j]$  is defined for  $j = 1, \dots, n$ . Then

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

and

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

**Proof:** Let  $Y$  and  $Z$  be two independent non-negative random variables. Let  $0 \leq \phi_n \leq Y$  and  $0 \leq \psi_n \leq Z$  be Lebesgue ladder functions for  $Y$  and  $Z$  (see Rudin *Real and Complex Analysis, Third Edition*, Theorem 1.17). Since  $Y$  and  $Z$  are independent, so are  $\phi_n$  and  $\psi_n$  for any  $n$ . We know that this theorem is true for simple random variables (Lemma 16), so  $E[\phi_n \psi_n] = E[\phi_n]E[\psi_n]$  for each  $n$ . Let  $n \rightarrow \infty$  and apply the Lebesgue Dominated Convergence Theorem to conclude that  $E[YZ] = E[Y]E[Z]$ . Now for the general case of two random variables, first use the positive/negative part decomposition and then the real/imaginary part decomposition. For more than two random variables use the corollary to Theorem 46 and a proof by induction. **QED**

**Theorem 49** [Kolmogorov's Zero-One Law] Let  $(\Omega, \mathcal{F}, \Pr)$  be a given probability space and let  $\{X_j\}_{j=1}^{\infty}$  be a sequence of independent random variables on  $(\Omega, \mathcal{F}, \Pr)$ . Let  $\mathcal{F}_n$  be the sigma algebra generated by  $X_n$ . Define the  $n^{\text{th}}$  tail field  $\mathcal{T}_n$  to be the sigma algebra generated by  $X_{n+1}, X_{n+2}, \dots$ , and the tail sigma algebra,  $\mathcal{T}$  to be

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \mathcal{T}_n.$$

Then  $\mathcal{T}$  is trivial. That is,  $F \in \mathcal{T}$  implies  $\Pr(F) = 0$  or  $\Pr(F) = 1$ .

**Proof:** For every  $n$ , it follows from the corollary to Theorem 46 that  $\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n, \mathcal{T}_n$  are independent sigma algebras, and so  $\mathcal{T}, \mathcal{F}_1, \mathcal{F}_2, \dots$  are independent sigma algebras. Hence by Theorem 46,  $\mathcal{T}$  and  $\bigvee_{n=1}^{\infty} \mathcal{F}_n$  are independent sigma algebras. However,  $\mathcal{T} \subset \bigvee_{n=1}^{\infty} \mathcal{F}_n$ , so  $\mathcal{T}$  is independent of itself, proving the theorem, since  $\Pr(A) = \Pr(A) \Pr(A)$  implies that  $\Pr(A) = 0$  or  $\Pr(A) = 1$ . **QED**

Kolmogorov's Zero-One Law leads to the theory of sums of independent random variables. For example:

**Corollary 50** Let  $\{X_n\}_{n=1}^{\infty}$  be a sequence of independent real valued random variables defined on  $(\Omega, \mathcal{F}, \Pr)$ . Then the series

$$\sum_{n=1}^{\infty} X_n(\omega)$$

converges with probability 0 or 1. That is,

$$\Pr \left( \left\{ \omega : \sum_{n=1}^{\infty} X_n(\omega) \text{ converge} \right\} \right)$$

is either 0 or 1.

**Proof:** Let

$$A = \left\{ \omega : \sum_{n=1}^{\infty} X_n(\omega) \text{ converges} \right\}.$$

It is clear that  $A \in \mathcal{T}_n$  for every  $n$ , so  $A \in \mathcal{T}$ . **QED**

## 2.6 Two Applications of Kolmogorov's Zero-One Law

Steinhaus Random Sign Problem  $\approx 1930$ .

Let  $\{a_n\}_{n=1}^{\infty}$  be a sequence of real numbers, and let  $\epsilon_n$  be a sequence of independent, identically distributed random variables with

$$\Pr(\{\omega : \epsilon_n(\omega) = 1\}) = \Pr(\{\omega : \epsilon_n(\omega) = -1\}) = 1/2.$$

The problem is does

$$\sum_{n=1}^{\infty} a_n \epsilon_n(\omega)$$

converge or not. The Kolmogorov Zero-One Law says that it either converges with probability 1 or 0. In fact, it converges with probability 1 if and only if the  $a_n$  are square-summable, a fact which can be proved using Martingale Theory. (See below.)

**Sum of IID Random Variables** Suppose that  $\{X_n\}_{n=1}^{\infty}$  is a sequence of independent, identically distributed random variables on  $(\Omega, \mathcal{F}, \Pr)$ . For each positive integer  $N$  put  $S_N = X_1 + \cdots + X_N$ . What does Kolmogorov's Zero-One Law say about  $N^{-1}S_N$  as  $N \rightarrow \infty$ ?

We know that  $A \equiv \limsup_{N \rightarrow \infty} N^{-1}S_N$  is measurable with respect to  $\mathcal{T}$  because for each positive integer  $k$

$$A = \limsup_{N \rightarrow \infty} N^{-1}(X_k + \cdots + X_N).$$

The same is true for  $B \equiv \liminf_{N \rightarrow \infty} N^{-1}S_N$ . Therefore the set of  $\omega$  on which  $N^{-1}S_N$  converges is a subset of  $\mathcal{T}$ , and, therefore, has probability 0 or probability 1.

## Chapter 3

# The Strong Law of Large Numbers

Now we shall follow Kolmogorov's development up to the Strong Law of Large Numbers ( $\approx 1930$ ).

### 3.1 Preliminaries

The basic technique is to use a variance calculation, because of the following fundamental lemma.

**Definition 51 (Variance)** *Suppose that  $X$  is a random variable for which  $E[X^2] < \infty$ . Then  $E[|X|] < \infty$  and the **variance** of  $X$ , denoted  $\text{Var}[X]$ , is defined by*

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2.$$

Observe that if  $\text{Var}[X] = 0$  then  $X = E[X]$  with probability 1.

**Lemma 52 [Variance of a sum]** *Suppose that  $X_1$  and  $X_2$  are two independent random variables on the probability space  $(\Omega, \mathcal{F}, \text{Pr})$  and that  $X_1$  and  $X_2$  have finite variances. Then  $X_1 + X_2$  has a finite variance and  $\text{Var}[X_1 + X_2] = \text{Var}[X_1] + \text{Var}[X_2]$ .*

**Proof:** Without loss of generality we may assume that  $E[X_1] = E[X_2] = 0$ . Then  $E[X_1 + X_2] = 0$ ,  $E[(X_1 + X_2)^2] < \infty$ , and

$$\begin{aligned} E[(X_1 + X_2)^2] &= E[X_1^2] + E[X_2^2] + E[X_1 X_2] \\ &= E[X_1^2] + E[X_2^2] + E[X_1]E[X_2] \text{ Theorem 49} \\ &= \text{Var}[X_1] + \text{Var}[X_2], \end{aligned}$$

as desired. **QED**

We begin by proving Kolmogorov's inequality, an improvement over Chebychev's Inequality.

**Lemma 53 (Kolmogorov's Inequality)** *Let  $\{X_n\}_{n=1}^{\infty}$  be a sequence of independent random variables defined on  $(\Omega, \mathcal{F}, \text{Pr})$ , with  $E[X_n] = 0$  and  $\text{Var}[X_n] = \sigma_n^2 < \infty$ . For each positive integer  $N$  let  $S_N = X_1 + \cdots + X_N$ , and let  $M_N = \max\{|S_1|, \dots, |S_N|\}$ . Then for any  $a > 0$  we have*

$$\text{Pr}(\{\omega : M_N(\omega) \geq a\}) \leq \frac{1}{a^2} \sum_{n=1}^N \sigma_n^2.$$

**Proof:** The proof uses the idea of a stopping time.

Fix  $a > 0$  and let  $A = \{\omega : M_N(\omega) \geq a\}$ . For each positive integer  $k \leq N$  let  $A_k = \{\omega : |S_n(\omega)| < a, n < k; |S_k(\omega)| \geq a\}$ . The  $A_k$  keep track of when the maximum first meets or exceeds  $a$ . The  $A_k$  are disjoint and their union is  $A$ . Let  $I_k$  be the indicator of  $A_k$ , and let  $I$  be the indicator of  $A$ . We then have

$$\begin{aligned} \sum_{n=1}^N \sigma_n^2 &= \text{Var}[S_N] \\ &= \mathbb{E}[S_N^2] \\ &\geq \mathbb{E}[S_N^2 I] \\ &= \mathbb{E}(S_N^2 (I_1 + \cdots + I_N)) \\ &= \mathbb{E}[S_N^2 I_1] + \cdots + \mathbb{E}[S_N^2 I_N] \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}[S_N^2 I_j] &= \mathbb{E}[(S_j + S_N - S_j)^2 I_j] \\ &= \mathbb{E}[S_j^2 I_j] + 2\mathbb{E}[S_j(S_N - S_j)I_j] + \mathbb{E}[(S_N - S_j)^2 I_j] \\ &= \mathbb{E}[S_j^2 I_j] + 2\mathbb{E}[S_j I_j] \mathbb{E}[(S_N - S_j)] + \mathbb{E}[(S_N - S_j)^2 I_j] \\ &= \mathbb{E}[S_j^2 I_j] + \mathbb{E}[(S_N - S_j)^2 I_j] \\ &\geq \mathbb{E}[S_j^2 I_j]. \end{aligned}$$

Therefore

$$\begin{aligned} \sum_{n=1}^N \sigma_n^2 &\geq \mathbb{E}[S_1^2 I_1] + \cdots + \mathbb{E}[S_N^2 I_N] \\ &\geq a^2 \mathbb{E}[I_1 + \cdots + I_N] \\ &= a^2 \Pr(A) \end{aligned}$$

which implies the lemma. **QED**

Before proceeding to the next theorem, some remarks on  $\mathcal{L}_2(\Omega, \mathcal{F}, \Pr)$ , the random variables on  $(\Omega, \mathcal{F}, \Pr)$  with finite second moments. This is a Hilbert space with inner product  $\langle X, Y \rangle \equiv \mathbb{E}[XY]$ . A sequence  $\{X_n\}_{n=1}^\infty$  in  $\mathcal{L}_2(\Omega, \mathcal{F}, \Pr)$  converges in the  $L^2$  sense if and only if it is a Cauchy sequence, that is, for every  $\epsilon > 0$  there is an  $N > 0$  such that  $n > N$  and  $m > N$  implies  $\mathbb{E}[|X_n - X_m|^2] < \epsilon$ . If  $\{X_n\}_{n=1}^\infty$  is an infinite sequence of independent random variables with  $\mathbb{E}[X_n] = 0$  and  $S_N = X_1 + \cdots + X_N$  then the following are equivalent:

1.  $S_N$  converges in the  $L^2$  sense as  $N \rightarrow \infty$ .
2.  $\{S_N\}_{N=1}^\infty$  is a Cauchy sequence in  $\mathcal{L}_2(\Omega, \mathcal{F}, \Pr)$ .
3.  $\{\mathbb{E}[S_N^2]\}_{N=1}^\infty$  is a real valued Cauchy sequence.
4.  $\{\mathbb{E}[X_1^2] + \cdots + \mathbb{E}[X_N^2]\}_{N=1}^\infty$  is a real valued Cauchy sequence.
5.  $\mathbb{E}[X_1^2] + \cdots + \mathbb{E}[X_N^2]$  converges as  $N \rightarrow \infty$ .

**Theorem 54** Let  $\{X_n\}_{n=1}^{\infty}$  be a sequence of independent random variables with  $E[X_n] = 0$  and  $\text{Var}[X_n] < \infty$ . Let  $\sigma_n^2 = \text{Var}(X_n)$ . If  $\sum_{n=1}^{\infty} \sigma_n^2 < \infty$  then  $\sum_{n=1}^{\infty} X_n$  converges with probability 1.

**Remark:** This proves the first part of the Steinhaus Random Signs Problem.

**Proof:** We will show that if  $S_N = X_1 + \cdots + X_N$  then  $\{S_N(\omega)\}_{N=1}^{\infty}$  is a Cauchy sequence for almost every  $\omega \in \Omega$ .

Suppose not. Then for some positive integer  $k$ ,  $|S_j - S_p| > 1/k$  for arbitrarily large  $j$  and  $p$  on a set of positive probability. So it suffices to prove that for all positive integers  $k$ ,

$$P \equiv \Pr(\{\omega : |S_j(\omega) - S_p(\omega)| > 1/k \text{ for arbitrarily large } j, p\}) = 0.$$

The probability in question can be estimated so as to use Kolmogorov's Inequality by using the triangle inequality.

For each positive integer  $q$ ,

$$\Pr(\{\omega : |S_j(\omega) - S_p(\omega)| > 1/k\}) \leq \Pr(\{\omega : |S_j(\omega) - S_q(\omega)| + |S_p(\omega) - S_q(\omega)| > 1/k\})$$

Therefore,

$$\begin{aligned} P &\leq \Pr(\{\omega : |S_j(\omega) - S_p(\omega)| + |S_p(\omega) - S_q(\omega)| > 1/k \text{ for arbitrarily large } j, p\}) \\ &\leq 2 \Pr(\{\omega : |S_j(\omega) - S_q(\omega)| > 1/2k \text{ for arbitrarily large } j\}) \end{aligned}$$

Let  $a = 2k$ . Kolmogorov's Inequality says

$$\Pr(\{\omega : \max_{1 \leq j \leq m} |S_{q+j}(\omega) - S_q(\omega)| > a\}) \leq \frac{1}{a^2} \sum_{j=q+1}^{q+m} \sigma_j^2 \leq \frac{1}{a^2} \sum_{j=q+1}^{\infty} \sigma_j^2$$

so

$$\Pr(\{\omega : \max_{1 \leq j < \infty} |S_{q+j}(\omega) - S_q(\omega)| > a\}) \leq \frac{1}{a^2} \sum_{j=q+1}^{\infty} \sigma_j^2$$

which converges to 0 as  $q \rightarrow \infty$ . Therefore

$$\Pr(\{\omega : |S_j(\omega) - S_q(\omega)| > 1/2k \text{ for arbitrarily large } j\}) = 0$$

which proves the theorem. **QED**

**Lemma 55 (Kronecker's Lemma)** Let  $\{a_k\}_{k=1}^{\infty}$  be a sequence of real numbers and Let  $\{b_k\}_{k=1}^{\infty}$  be a sequence of non-negative real numbers which decreases monotonically to 0. If

$$\sum_{k=1}^{\infty} a_k b_k$$

is convergent then

$$\lim_{N \rightarrow \infty} b_N \sum_{k=1}^N a_k = 0.$$

**Proof:** Let  $s_0 = 0$  and  $s_n = \sum_{k=1}^n a_k b_k$ . Then we have  $a_n = b_n^{-1}(s_n - s_{n-1})$  for each positive integer  $n$ , and

$$\begin{aligned} b_N \sum_{k=1}^N a_k &= b_N \sum_{k=1}^N \frac{s_k - s_{k-1}}{b_k} \\ &= s_N - b_N \sum_{k=0}^{N-1} s_k \left( \frac{1}{b_{k+1}} - \frac{1}{b_k} \right) \end{aligned}$$

where we take  $1/b_0 = 0$ . We know that

$$b_N \sum_{k=0}^{N-1} \left( \frac{1}{b_{k+1}} - \frac{1}{b_k} \right) = 1$$

for every  $N$ . Suppose that  $s_N \rightarrow s$  as  $N \rightarrow \infty$ . Then

$$\begin{aligned} \lim_{N \rightarrow \infty} b_N \sum_{k=1}^N a_k &= \lim_{N \rightarrow \infty} \left( s_N - b_N \sum_{k=0}^{N-1} s_k \left( \frac{1}{b_{k+1}} - \frac{1}{b_k} \right) \right) \\ &= s - s \\ &= 0 \end{aligned}$$

as desired. **QED**

**Theorem 56 (Kolmogorov's Theorem)** Let  $\{X_n\}_{n=1}^{\infty}$  be a sequence of independent random variables on  $(\Omega, \mathcal{F}, \Pr)$  with  $E[X_n] = 0$  for all  $n$ . Let  $\sigma_n^2 = E[X_n^2]$  and  $S_N = X_1 + \cdots + X_N$ . If

$$\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} \leq \infty$$

then

$$\Pr(\{\omega : \lim_{N \rightarrow \infty} N^{-1} S_N(\omega) = 0\}) = 1.$$

**Remark:** This is a strong law of large numbers since the  $X_n$  have mean 0.

**Proof:** The hypotheses guarantee via Theorem 54 that

$$\Pr\left(\left\{\omega : \sum_{n=1}^{\infty} \frac{X_n(\omega)}{n} \text{ exists}\right\}\right) = 1.$$

For each  $\omega$  for which the sum is convergent we apply Kronecker's Lemma, giving us the desired conclusion. **QED**

**Theorem 57 (Kolmogorov's Strong Law of Large Numbers)** Let  $\{X_n\}_{n=1}^{\infty}$  be a sequence of independent identically distributed random variables with  $E[X_n] = \mu$ . Put  $S_N = X_1 + \cdots + X_N$ . Then

$$\Pr\left(\omega : \lim_{N \rightarrow \infty} N^{-1} S_N(\omega) = \mu\right) = 1.$$

Furthermore, if

$$\Pr\left(\omega : \lim_{N \rightarrow \infty} N^{-1} S_N(\omega) = \alpha\right) = a > 0,$$

then  $\alpha = \mu$  and  $a = 1$ .

**Proof:** Without loss of generality we may assume  $\mu = 0$ . The main idea of the proof is truncation.

Let  $B_k = \{\omega : |X_k(\omega)| \leq k\}$ . Let  $I_k$  be the indicator of  $B_k$  and let  $Y_k = X_k I_k$  and  $Z_k = X_k(1 - I_k)$ . Then

$$N^{-1}S_N = N^{-1}(Y_1 + \cdots + Y_N) + N^{-1}(Z_1 + \cdots + Z_N).$$

Note that if  $\Pr(\omega : Z_n(\omega) \neq 0 \text{ i.o.}) = 0$  then

$$N^{-1}(Z_1 + \cdots + Z_N) \rightarrow 0 \tag{3.1}$$

with probability 1.

According to the Borel-Cantelli Lemma,  $\Pr(\omega : Z_n(\omega) \neq 0 \text{ i.o.}) = 0$  is equivalent to

$$\sum_{k=1}^{\infty} \Pr(\{\omega : Z_k(\omega) \neq 0\}) < \infty,$$

since the  $Z_k$  are independent random variables. However,

$$\begin{aligned} \sum_{k=1}^{\infty} \Pr(\{\omega : Z_k(\omega) \neq 0\}) &= \sum_{k=1}^{\infty} \Pr(\{\omega : |X_k(\omega)| > k\}) \\ &= \sum_{k=1}^{\infty} \Pr(\{\omega : |X_1(\omega)| > k\}) \\ &< \infty \end{aligned}$$

since  $E[|X_1|] < \infty$ .

Now we just have to show that  $N^{-1}(Y_1 + \cdots + Y_N) \rightarrow 0$  with probability 1.

First we get a handle on  $\text{Var}(Y_k)$ :

$$\begin{aligned} E[Y_k] &= E[X_k; |X_k| \leq k] \\ &= E[X_1; |X_1| \leq k] \end{aligned}$$

so from Lebesgue's Dominated Convergence Theorem we get  $E[Y_k] \rightarrow 0$  as  $k \rightarrow \infty$ . Therefore

$$\lim_{N \rightarrow \infty} N^{-1} \sum_{k=1}^N E[Y_k] = 0. \tag{3.2}$$

$$\begin{aligned} \text{Var}[Y_k] &\leq E[Y_k^2] \\ &= E[X_k^2; |X_k| \leq k] \\ &= E[X_1^2; |X_1| \leq k] \\ &= \sum_{j=1}^k E[X_1^2; j-1 < X_1 \leq j], \end{aligned}$$

so

$$\begin{aligned}
\sum_{k=1}^{\infty} k^{-2} \text{Var}[Y_k] &\leq \sum_{k=1}^{\infty} k^{-2} \sum_{j=1}^k \mathbb{E}[X_1^2; j-1 < X_1 \leq j] \\
&= \sum_{j=1}^{\infty} \left( \sum_{k=j}^{\infty} k^{-2} \right) \mathbb{E}[X_1^2; j-1 < X_1 \leq j] \\
&\leq C \sum_{j=1}^{\infty} j^{-1} \mathbb{E}[X_1^2; j-1 < X_1 \leq j] \\
&\leq C \sum_{j=1}^{\infty} \mathbb{E}[|X_1|; j-1 < X_1 \leq j] \\
&= \mathbb{E}[|X_1|] < \infty.
\end{aligned}$$

Therefore, by Kolmogorov's Theorem (Theorem 56),

$$\lim_{N \rightarrow \infty} N^{-1} \sum_{k=1}^N (Y_k - \mathbb{E}[Y_k]) = 0$$

with probability 1. If we apply (3.2) we see that

$$\lim_{N \rightarrow \infty} N^{-1} \sum_{k=1}^N Y_k = 0.$$

This combined with (3.1) proves the first assertion of the theorem. The second assertion follows from the first assertion and Kolmogorov's Zero-One Law. **QED**

Kolmogorov's Strong Law of Large Numbers has the following converse.

**Theorem 58** *Let  $\{X_n\}_{n=1}^{\infty}$  be a sequence of independent identically distributed random variables with  $\mathbb{E}[|X_n|] = \infty$ . Put  $S_N = X_1 + \cdots + X_N$ . Then*

$$\Pr \left( \omega : \lim_{N \rightarrow \infty} N^{-1} S_N(\omega) \text{ converges} \right) = 0.$$

**Proof:** By the Borel-Cantelli Lemma,  $\Pr(\{\omega : |X_k(\omega)| \geq k \text{ i.o.}\}) = 1$ . Therefore the sequence  $N^{-1} S_N$  changes by at least 1 infinitely often with probability 1, so it cannot converge. **QED**

The hypothesis of mutual independence is not needed for the conclusion of the Strong Law of Large Numbers to remain valid. This was demonstrated by Etemadi in 1981:

**Theorem 59 (Etemadi's Strong Law of Large Numbers)** *Let  $\{X_n\}_{n=1}^{\infty}$  be pairwise independent, identically distributed random variables with  $\mathbb{E}[|X_1|] < \infty$ . Then, with probability 1,*

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n X_k = \mathbb{E}[X_1].$$

**Proof:** Without loss of generality let us assume that the  $X_k$  are non-negative. Let  $Y_k = X_k I_{\{X_k \leq k\}}$ . Put

$$T_n = \sum_{k=1}^n Y_k.$$

Choose  $a \in (1, 2)$  and let  $k_n$  denote the greatest integer in  $a^n$ .

$$\begin{aligned} \sum_{n=1}^N \text{Var}[T_{k_n}/k_n] &= \sum_{n=1}^N \frac{1}{k_n^2} \sum_{j=1}^{k_n} \text{Var}[Y_j] \\ &\leq C_a \sum_{j=1}^{k_N} \frac{1}{j^2} \text{Var}[Y_j] \\ &\leq C_a \sum_{j=1}^{k_N} \frac{1}{j^2} \text{E}[Y_j^2] \\ &= C_a \sum_{j=1}^{k_N} \frac{1}{j^2} \sum_{i=0}^{j-1} \text{E}[Y_j^2; i < Y_j \leq i+1] \\ &= C_a \sum_{j=1}^{k_N} \frac{1}{j^2} \sum_{i=0}^{j-1} \text{E}[X_1^2; i < X_1 \leq i+1] \\ &\leq C'_a \sum_{i=0}^{k_N} \frac{1}{i+1} \text{E}[X_1^2; i < X_1 \leq i+1] \\ &\leq C'_a \sum_{i=0}^{k_N} \text{E}[X_1; i < X_1 \leq i+1] \\ &\leq C'_a \text{E}[X_1]. \end{aligned}$$

Therefore it follows, in the usual manner, from Chebychev's inequality and the Borel-Cantelli Lemma that

$$\lim_{n \rightarrow \infty} k_n^{-1}(T_{k_n} - \text{E}[T_{k_n}]) = 0$$

with probability 1. As in the proof of Kolmogorov's Strong Law of Large Numbers we then see that

$$\lim_{n \rightarrow \infty} k_n^{-1} \sum_{j=1}^{k_n} X_k = \text{E}[X_1]$$

with probability 1. Finally, since the  $X_k$  are non-negative we have

$$\frac{1}{a} \text{E}[X_1] \leq \liminf_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n X_k \leq \limsup_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n X_k \leq a \text{E}[X_1]$$

with probability 1 for each rational number  $a \in (1, 2)$ . The result now follows from the Pinching Theorem.

**QED**

## 3.2 A problem of records

An urn contains the integers from 1 to  $n$ , which are sampled successively without replacement. To model this probabilistically, we must define  $(\Omega, \mathcal{F}, \Pr)$ . Let  $\Omega$  be  $S_n$  the set of all permutations of  $\{1, \dots, n\}$ . We let  $\mathcal{F} = 2^\Omega$ , and let  $\Pr$  be the uniform probability measure on  $\mathcal{F}$ .

Now, let  $A_k$  be the event of a “record at time  $k$ ”, that is that the  $k^{\text{th}}$  integer selected exceeds all of its predecessors. Let  $I_k$  be the indicator of  $A_k$ , and let  $S_m = I_1 + \dots + I_m$ , for  $0 \leq m \leq n$ . ( $S_0 = 0$ .)

By induction on  $k$  we can prove that  $\Pr(A_k) \equiv \mathbb{E}[I_k] = 1/k$ . We can also prove that the events  $A_k$  are mutually independent. So for any  $n$ , the random variable  $\{I_k\}_{k=1}^n$  are mutually independent. We wish to consider the behavior of  $S_n/\mathbb{E}[S_n]$  as  $n \rightarrow \infty$ . First observe that

$$\begin{aligned} \lim_{n \rightarrow \infty} \Pr \left( \left\{ \omega : \left| \frac{S_n(\omega)}{\mathbb{E}[S_n]} - 1 \right| \geq \epsilon > 0 \right\} \right) &= \lim_{n \rightarrow \infty} \Pr (\{ \omega : |S_n(\omega) - \mathbb{E}[S_n]| \geq \epsilon \mathbb{E}[S_n] > 0 \}) \\ &\leq \lim_{n \rightarrow \infty} \frac{\text{Var}[S_n]}{\epsilon^2 (\mathbb{E}[S_n])^2} \\ &= \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n \text{Var}[I_k]}{\epsilon^2 (\mathbb{E}[S_n])^2} \\ &\leq \lim_{n \rightarrow \infty} \frac{\sum_{k=1}^n \mathbb{E}[I_k]}{\epsilon^2 (\mathbb{E}[S_n])^2} \\ &= \lim_{n \rightarrow \infty} \frac{1}{\epsilon^2 \mathbb{E}[S_n]} \\ &\leq \lim_{n \rightarrow \infty} \frac{C}{\epsilon^2 \log(n)} \\ &= 0. \end{aligned}$$

This is a sort of Weak Law of Large Numbers. What about a Strong Law of Large Numbers in this example? It does make sense to take an infinite sequence of independent and identically distributed random variables with a continuous distribution function  $F$ ,  $\{X_k\}_{k=1}^\infty$ , on  $(\Omega, \mathcal{F}, \Pr)$ , and define

$$A_k = \{ \omega : X_k(\omega) > \max\{X_1(\omega), \dots, X_{k-1}(\omega)\} \},$$

where  $A_1 = \Omega$ . Let  $I_k$  be the indicator of  $A_k$ . It can be shown that the  $I_k$  are mutually independent with  $\mathbb{E}[I_k] = 1/k$ .

Alternatively, we could just assume that we have an infinite sequence of mutually independent indicator functions  $\{I_k\}_{k=1}^\infty$  with  $\mathbb{E}[I_k] = 1/k$ . Either way, put  $S_n = I_1 + \dots + I_n$ , and we would like to show that

$$\Pr \left( \left\{ \omega : \lim_{n \rightarrow \infty} \frac{S_n(\omega)}{\log(n)} = 1 \right\} \right) = 1.$$

This will follow from Kronecker’s Lemma if we can show that

$$\Pr \left( \left\{ \omega : \sum_{k=2}^{\infty} \frac{I_k(\omega) - k^{-1}}{\log(k)} \text{ is convergent} \right\} \right) = 1.$$

In turn, this follows from the proof of Kolmogorov’s Theorem (Theorem 56), once we have shown that

$$\sum_{k=2}^{\infty} \text{Var} \left[ \frac{I_k}{\log(k)} \right] < \infty.$$

This is easy, since

$$\begin{aligned} \sum_{k=2}^{\infty} \text{Var} \left[ \frac{I_k}{\log(k)} \right] &\leq \sum_{k=2}^{\infty} \frac{\text{E}[I_k]}{(\log(k))^2} \\ &= \sum_{k=2}^{\infty} \frac{1}{k(\log(k))^2} \\ &< \infty. \end{aligned}$$

We can generalize this further. Suppose that  $I_k$  are mutually independent indicator functions with  $\text{E}[I_k] = p_k > 0$ . Put  $S_n = I_1 + \cdots + I_n$ , and put  $P_k = p_1 + \cdots + p_k$ . Then

$$\Pr \left( \left\{ \omega : \frac{S_n(\omega)}{\text{E}[S_n]} = 1 \right\} \right) = 1$$

since for every integer  $N > 1$

$$\begin{aligned} \sum_{k=1}^N \text{Var} \left[ \frac{I_k}{\text{E}[S_k]} \right] &\leq \sum_{k=1}^N \frac{p_k}{P_k^2} \\ &< \frac{1}{P_1} + \sum_{k=2}^N \frac{p_k}{P_{k-1}P_k} \\ &= \frac{1}{P_1} + \sum_{k=2}^N \left( \frac{1}{P_{k-1}} - \frac{1}{P_k} \right) \\ &= \frac{2}{P_1} - \frac{1}{P_N} \\ &< \frac{2}{P_1} \end{aligned}$$

Note that this encompasses both our result on records, which is  $p_k = 1/k$  and our strong law of large numbers for events in the iid case, where  $p_k = p > 0$ .

Suppose that  $x := (x_1, x_2, \dots)$  is a sequence of real numbers. We say that a record value for  $x$  occurs at time  $t$  if  $x_t > x_j$  for all positive integers  $j$  which are less than  $t$ . To count the number of records which in occur in  $x$  by time  $t$ ,  $N_t(x)$ , define  $I_t(x) = 1$  if a record is set at time  $t$  and 0 otherwise, so that we have

$$N_t(x) = \sum_{k=1}^t I_k(x)$$

for each positive integer  $t$ .

Also, if the sequence  $x$  has the property that for each positive integer  $m$  there exists an integer  $n > m$  so that  $x_n > x_m$  then we let  $L_n(x)$  denote the time of the  $n^{\text{th}}$  record. That is,  $L_1(x) = 1$ , and  $L_n(x) = \inf\{m : x_m > x_{L_{n-1}(x)}\}$  for  $n > 1$ . Observe that  $N_{L_n(x)}(x) = n$  and  $L_{N_n(x)}(x) \leq n$  for each positive integer  $n$ .

When the deterministic sequence  $x$  is replaced by an iid sequence  $X$  of real valued random variables,  $L_n(X)$  and  $R_n(X)$  are sequences of random variables whose limiting behavior is of interest. The results for iid sequences with a continuous distribution are well-known. The following theorem is straightforward to prove since the  $I_k(X)$  turn out to be independent. See, for example, page 45 of Durrett[1991].

**Theorem 60** Suppose that  $X := (X_1, X_2, \dots)$  is a sequence of iid random variables whose common distribution function is continuous. Then

$$\lim_{n \rightarrow \infty} \frac{N_n(X)}{\log(n)} = 1 \text{ a.s.} \quad (3.3)$$

and

$$\lim_{n \rightarrow \infty} \frac{\log(n)}{\mathbb{E}[N_n(X)]} = 1. \quad (3.4)$$

A standard device gives the following corollary:

**Corollary 61** Under the hypotheses of the preceding theorem,

$$\lim_{n \rightarrow \infty} \frac{L_n(X)}{n} = 1 \text{ a.s.}$$

**Proof:** Let  $A$  be the set of all  $\omega$  such that

$$\lim_{n \rightarrow \infty} L_n(X(\omega)) = +\infty$$

and

$$\lim_{n \rightarrow \infty} \frac{N_n(X(\omega))}{\log(n)} = 1.$$

$\Pr(A) = 1$ . Observe that

$$\frac{\log(L_n(X(\omega)))}{n} = \frac{\log(L_n(X(\omega)))}{N_{L_n(X(\omega))}(X(\omega))}$$

so for each  $\omega \in A$ ,

$$\frac{\log(L_n(X(\omega)))}{n}$$

is a subsequence of a sequence which converges to 1. **QED**

When the sequence is a sequence of iid integer valued random variables, it is more difficult to get good results since the  $I_k(X)$  are no longer independent. The fundamental paper in this case is Vervaat[1973]. The conclusions of the following theorem may be reached from Theorem 4.1 of that paper, but at the price of more complicated arguments, and, as we shall see, stronger assumptions on the distribution of our iid sequence.

**Theorem 62** Suppose that  $Z := (Z_1, Z_2, \dots)$  is an iid sequence of positive integer valued random variables. Put  $f_d = \Pr(Z_n = d)$  and  $F_d = \Pr(Z_n \leq d)$ . If  $f_d > 0$  for all positive integers  $d$  and

$$\sum_{k=2}^{\infty} \frac{1}{k \log(k)} \sum_{d=1}^{\infty} (F_d^k - F_{d-1}^k - kF_{d-1}^{k-1}f_d) < \infty \quad (3.5)$$

then

$$\lim_{n \rightarrow \infty} \frac{N_n(Z)}{\log(n)} = 1 \text{ a.s.}, \quad (3.6)$$

$$\lim_{n \rightarrow \infty} \frac{L_n(X)}{n} = 1 \text{ a.s.} \quad (3.7)$$

and

$$\lim_{n \rightarrow \infty} \frac{\log(n)}{\mathbb{E}[N_n(Z)]} = 1. \quad (3.8)$$

In particular, (3.5) holds if

$$\sum_{d=1}^{\infty} \left( \frac{f_d}{1 - F_d} \right)^2 \frac{-1}{\log(1 - F_d)} < \infty. \quad (3.9)$$

**Proof:** Let  $U := (U_1, U_2, \dots)$  be an iid sequence of continuous uniform random variables on  $(0, 1)$ , and assume that  $U$  and  $Z$  are independent. If we define the sequence  $X$  by  $X_n = Z_n + U_n$  then  $X$  satisfies the hypotheses of Theorem 60. What is more, for each integer  $n > 1$ ,

$$N_n(X) = N_n(Z) + \sum_{k=1}^n (I_k(X) - I_k(Z)) \quad (3.10)$$

and  $I_k(X) - I_k(Z) \geq 0$ . What is more,

$$\begin{aligned} \mathbb{E}[I_k(X) - I_k(Z)] &= \frac{1}{k} - \sum_{d=1}^{\infty} F_{d-1}^{k-1} f_d \\ &= \frac{1}{k} \sum_{d=1}^{\infty} (F_d^k - F_{d-1}^k - k F_{d-1} f_d) \end{aligned} \quad (3.11)$$

$$= \frac{k-1}{2} \sum_{d=1}^{\infty} (F_{d-1} + u_{k,d} f_d)^{k-2} f_d^2 \quad (3.12)$$

where  $0 < u_{k,d} < 1$  for each  $d$  and  $k$ .

We see from (3.11) that (3.5) and

$$\sum_{k=2}^{\infty} \frac{\mathbb{E}[I_k(X) - I_k(Z)]}{\log(k)} < \infty \quad (3.13)$$

are equivalent. Therefore (3.8) follows from (3.4), (3.10), (3.13) and Kronecker's lemma.

To prove (3.6) observe that since each  $I_k(X) - I_k(Z)$  is a non-negative random variable, (3.13) implies that

$$\sum_{k=2}^{\infty} \frac{I_k(X) - I_k(Z)}{\log(k)} < \infty \text{ a.s.},$$

so Kronecker's lemma gives

$$\lim_{n \rightarrow \infty} \frac{N_n(X) - N_n(Z)}{\log(n)} = 0 \text{ a.s.}$$

This combined with (3.3) and (3.10) shows that (3.5) implies (3.6). Finally, (3.7) follows from the same argument used to establish the Corollary to Theorem 60.

My colleague, Hans Volkmer, has pointed out that if for  $0 < a < x < 1$  we define

$$H(x) = \sum_{k=2}^{\infty} \frac{k-1}{\log(k)} x^{k-2} \quad (3.14)$$

then it follows from Exercise 6 on page 242 of Titchmarsh[1939] and partial summation that there are positive constants  $A < B$  such that

$$A \leq -H(x)(1-x)^2 \log(1-x) \leq B$$

for  $a < x < 1$ . This observation along with (3.12) shows that (3.9) implies (3.5). **QED**

If we put  $r_d = f_d/(1 - F_{d-1})$  for  $d \geq 1$  then

$$\sum_{d=1}^{\infty} r_d^a < \infty \tag{3.15}$$

for some  $a > 1$  implies (3.9), and, therefore (3.5). To see why, observe that

$$(1 - F_d) = \prod_{k=1}^d (1 - r_k).$$

Put  $r = \max\{r_1, r_2, \dots\}$ . Since  $0 < r_k < 1$  for all  $k$ , (3.15) gives  $0 < r < 1$ . Put  $R_d = r_1 + \dots + r_d$ . Since  $F_d$  converges to 1 we must have  $R_d$  divergent. Then

$$\sum_{d=1}^{\infty} \left( \frac{f_d}{1 - F_d} \right)^2 \frac{-1}{\log(1 - F_d)} \leq \frac{1}{(1 - r)^2} \sum_{d=1}^{\infty} \frac{r_d^2}{R_d}.$$

We immediately see that if  $1 < a \leq 2$  we are done. If  $a > 2$ , apply Hölder's inequality with  $p = a - 1$ ,  $q = p/(p - 1)$  and

$$\frac{r_d^2}{R_d} = r_d^{2-(1/q)} \frac{r_d^{1/q}}{R_d}$$

to get

$$\sum_{d=1}^{\infty} \frac{r_d^2}{R_d} \leq \left( \sum_{d=1}^{\infty} r_d^a \right)^{1/(a-1)} \left( \sum_{d=1}^{\infty} \frac{r_d}{R_d^q} \right)^{1/q} < \infty$$

by the Theorem of Abel and Dini (page 290 of Knopp[1928]). **QED**

To deduce the conclusions of Theorem 62 from Theorem 4.1 of Vervaat(1973), we must replace (3.5) with the condition

$$\lim_{n \rightarrow \infty} \frac{R_n^{(2)}}{f(R_n)} = 0, \tag{3.16}$$

where  $f(x) = \sqrt{2x \log \log(x)}$  and  $R_n^{(2)} = r_1^2 + \dots + r_n^2$ .

Note that (3.16) implies

$$\sum_{d=1}^{\infty} \frac{r_d^2}{R_d} < \infty, \tag{3.17}$$

and, therefore, (3.9), if  $r < 1$ .

To see why, we consider two cases. In the first, assume that (3.15) holds with  $a = 2$ . Then both (3.16) and (3.17) hold. Next consider the case where (3.15) fails for  $a = 2$ . Then

$$\sum_{d=1}^{\infty} \frac{r_d^2}{R_d} = \sum_{d=1}^{\infty} \frac{r_d^2}{\left( R_d^{(2)} \right)^{3/2}} \left( \frac{R_d^{(2)}}{R_d^{2/3}} \right)^{3/2} < \infty$$

since we have

$$\sum_{d=1}^{\infty} \frac{r_d^2}{\left(R_d^{(2)}\right)^{3/2}} < \infty$$

by the Theorem of Abel and Dini and

$$\frac{R_d^{(2)}}{R_d^{2/3}}$$

is bounded above since (3.16) says

$$\frac{R_d^{(2)}}{R_d^{2/3}} \frac{R_d^{1/6}}{\sqrt{\log(\log(R_d))}}$$

converges to 0 while  $R_d \rightarrow \infty$  as  $d \rightarrow \infty$ .

We conclude with an example. Suppose that  $r_d = 1/(d+1)^p$  where  $0 < p < 1/3$ . Then (3.16) fails while (3.17) holds.

### Acknowledgments

I should like to thank Barry Arnold and H. N. Nagaraja for pointing out the relation between this work and that of W. Vervaat. I should like to thank my colleague Hans Volkmer for his help with the analysis of the function  $H$  in (3.14).

### Bibliography

- Durrett, Richard. (1991) *Probability: Theory and Examples*. Wadsworth & Brooks/Cole. Pages 45-46.  
 Knopp, Konrad. (1928) *Theory and Application of Infinite Series*. Blackie and Son Limited. Pages 290 - 291.  
 Titchmarsh, E. C. (1939) *The Theory of Functions*. Oxford University Press. Page 242.  
 Vervaat, Wim (1973) *Limit theorems for records from discrete distributions*. Stochastic Processes Appl. 1, y317-334.

### 3.3 Moments in higher dimensions

Moments are useful for vector-valued random variables as well. Let  $X = (X_1, \dots, X_n)^t$ , and suppose that  $E[X_i] = 0$  and  $E[X_i^2] < \infty$  for  $i = 1$  to  $n$ . We want to make the following

**Definition 63 (Covariance)** For any  $a \in R^n$  define  $\Sigma(a)$  by

$$\Sigma(a) = \sum_{j=1}^n \sum_{k=1}^n a_k a_j E[X_j X_k] = E[\langle a | X \rangle^2] = E[a^t X X^t a],$$

where  $\langle * | * \rangle$  denotes the standard inner product on  $R^n$ , and  $^t$  denotes the transpose of a matrix.  $\Sigma(a)$  is a quadratic form. If we define  $\sigma_{j,k} = E[X_j X_k]$ , called the **covariance** of  $X_j$  and  $X_k$ , and we let  $\Sigma$  denote the (symmetric) matrix of the  $\sigma_{j,k}$  then

$$\Sigma(a) = \langle a | \Sigma a \rangle .$$

$\Sigma$  is called the **covariance matrix** of  $X$ . If the components of  $X$  do not have 0 mean, then the covariance matrix of  $X$  is that of the vector whose  $i^{\text{th}}$  component is  $X_i - E[X_i]$ .

**Theorem 64** If  $X$  and  $Y$  are independent  $R^n$  valued random variables with covariance matrices  $\Sigma_X$  and  $\Sigma_Y$  respectively then  $X + Y$  has covariance matrix  $\Sigma_X + \Sigma_Y$ .

**Proof:** : Compute it out. **QED**

**Theorem 65** A covariance quadratic form is always positive semidefinite. It is positive definite if and only if the probability measure of the random variable  $X$  on which it is based is not concentrated on a subspace of  $R^n$  of dimension less than  $n$ .

**Proof:** It is clear from the definition that  $\Sigma(a) \geq 0$  for all  $a \in R^n$ .

Suppose that  $\Sigma(a)$  is not positive definite. Then for some  $a \neq \vec{0}$  we have  $\Sigma(a) = 0$ . Hence with probability 1,  $\langle X|a \rangle = 0$ , that is  $\Pr(\langle X|a \rangle = 0) = 1$ , so the probability measure of  $X$  is concentrated on the orthogonal complement of the span of  $a$ , which is a subspace of dimension  $n - 1$ .

Conversely, if the distribution of  $X$  is concentrated on a proper subspace of  $R^n$  then this space is contained in the orthogonal complement of the span of some nonzero vector  $a$ , and then  $\langle X|a \rangle = 0$  with probability 1, giving  $\Sigma(a) = E[\langle X|a \rangle^2] = 0$ , as claimed. **QED**

As an example, let  $X = (X_1, X_2)$  and suppose that  $X$  is uniformly distributed on the unit circle, so that  $X_1^2 + X_2^2 = 1$ . By symmetry,  $E[X] = (0, 0)^t$ . Let  $\vec{a} = (a_1, a_2)^t$ . What is  $\Sigma(\vec{a})$ ? We can calculate directly:

$$\begin{aligned}\Sigma(\vec{a}) &= E[\langle \vec{a}|X \rangle^2] \\ &= E[(a_1 X_1 + a_2 X_2)^2] \\ &= a_1^2 E[X_1^2] + 2a_1 a_2 E[X_1 X_2] + a_2^2 E[X_2^2]\end{aligned}$$

We have  $X_1 = \cos(\Theta)$  and  $X_2 = \sin(\Theta)$  where  $\Theta$  is uniformly distributed on  $[0, 2\pi]$ , so

$$\begin{aligned}E[X_1 X_2] &= \int_0^{2\pi} \cos(\theta) \sin(\theta) \frac{1}{2\pi} d\theta \\ &= 0 \\ E[X_1^2] &= \int_0^{2\pi} \cos^2(\theta) \frac{1}{2\pi} d\theta \\ &= \frac{1}{2} \\ E[X_2^2] &= \int_0^{2\pi} \sin^2(\theta) \frac{1}{2\pi} d\theta \\ &= \frac{1}{2}\end{aligned}$$

so

$$\Sigma = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}.$$

### 3.3.1 Lord Raleigh's Random Walk

Starting at some point in the plane, select a point on the unit circle, centered at that point, the choice being uniformly distributed on that unit circle. Repeat this choice independently.

Let  $V^{(n)}$  be the position at time  $n$ , starting from  $(0, 0)$ . Then we have  $V^{(n)} = (v_1^{(n)}, v_2^{(n)})$ , and

$$V^{(n)} = \sum_{j=1}^n X^{(j)},$$

where the  $X^{(j)}$  are independent and identically distributed on the unit circle. We want to know something about the length,  $\|V^{(n)}\|$ , of  $V^{(n)}$ . The first thing to try is  $E[\|V^{(n)}\|]$ , but this is hard to get a hold of. However,  $E[\|V^{(n)}\|^2]$  is very easy to compute, as  $V^{(n)}$  has covariance matrix equal to

$$\Sigma = \begin{bmatrix} n/2 & 0 \\ 0 & n/2 \end{bmatrix}$$

, and so  $E[\|V^{(n)}\|^2] = n$ .

Something to think about: Let

$$p_n = \Pr(\{\omega : \|V^{(n)}(\omega)\| \leq 1\}).$$

What can be said about the sequence  $p_n$ ?

### 3.3.2 Product Measure and Fubini's Theorem

**Definition 66 (Product Probability Space)** For each  $k \in \{1, 2, \dots, n\}$  let  $(\Omega_k, \mathcal{F}_k, \Pr_k)$  be a probability space. Then the product probability space,

$$(\Omega, \mathcal{F}, \Pr) = \prod_{k=1}^n (\Omega_k, \mathcal{F}_k, \Pr_k),$$

is defined as follows:

1.  $\Omega$  is the Cartesian product of the  $\Omega_k$ , so that  $\omega \in \Omega$  is an  $n$ -tuple:  $\omega = (\omega_1, \dots, \omega_n)$ , where  $\omega_k \in \Omega_k$  for each  $k$ ;
2.  $\mathcal{F}$  is the smallest sigma algebra of subsets of  $\Omega$  which contains all cylinder sets of the form  $\{\omega \in \Omega : \omega_j \in A_j \in \mathcal{F}_j, j \in \{1, \dots, n\}\}$ . The set of all cylinder sets is a semi-algebra, and, therefore, there is a unique probability measure  $\Pr$  on  $\mathcal{F}$  such that

$$\Pr(\{\omega \in \Omega : \omega_j \in A_j \in \mathcal{F}_j, j \in \{1, \dots, n\}\}) = \prod_{k=1}^n \Pr_k(A_k).$$

For more details on this construction, see *Real and Complex Analysis* by Walter Rudin.

**Two facts:** The same construction goes through for countable and uncountable products. Here the cylinder sets used must be of the form

$$\prod_{\alpha \in I} A_\alpha,$$

where only finitely many of the  $A_\alpha$  may differ from their respective  $\Omega_\alpha$ .

This construction shows that there are indeed probability spaces having infinite sequences of independent and identically distributed random variables. The simplest way to get such a space is to take a distribution function,  $F$ , on the real line, and form the countable Cartesian product of  $((-\infty, \infty), \mathcal{B}, \mu_F)$ .

**Theorem 67 (Fubini's Theorem)** Suppose that  $(\Omega, \mathcal{F}, \Pr)$  is the product probability space

$$(\Omega_1, \mathcal{F}_1, \Pr_1) \times (\Omega_2, \mathcal{F}_2, \Pr_2).$$

Let  $f(\omega)$  be an integrable function on  $(\Omega, \mathcal{F}, \Pr)$ . Then

1. For each fixed  $\omega_1 \in \Omega_1$  and  $\omega_2 \in \Omega_2$ , the functions  $f(\omega_1, \cdot)$  and  $f(\cdot, \omega_2)$  are  $\mathcal{F}_2$  and  $\mathcal{F}_1$  measurable, respectively;
2. The functions

$$\Psi_{3-j}(\omega_{3-j}) \equiv \int_{\Omega_j} f(\omega_1, \omega_2) d\Pr_j(\omega_j)$$

are  $\mathcal{F}_{3-j}$  measurable for  $j = 1$  and  $j = 2$ ;

- 3.

$$\int_{\Omega} f d\Pr = \int_{\Omega_{3-j}} \left( \int_{\Omega_j} f(\omega_1, \omega_2) d\Pr_j(\omega_j) \right) d\Pr_{3-j}(\omega_{3-j})$$

This concept of iterated integration will generalize by the notion of conditional expectation.

For a proof of Fubini's Theorem, see *Real and Complex Analysis* by Walter Rudin or *Real Analysis and Probability* by Robert Ash. The spaces need not be probability spaces.

**Example:** Suppose that a fish lays  $N$  eggs, where  $N$  is a non-negative integer valued random variable with  $p_k = \Pr(N = k)$ . Suppose that each egg survives with probability  $p$  and is destroyed with probability  $1 - p$ . We assume that  $p$  is independent of  $N$  and is the same for each egg. Let  $X$  be the number of eggs which survive. Find the probability distribution of  $X$ , or  $E[X]$ ,  $\text{Var}[X]$ , and so on.

**Solution:** Let us define the simplest probability space on which this problem makes sense. If  $N$  is not assumed to be bounded then we need an infinite product space.

$$(\Omega, \mathcal{F}, \Pr) = \left[ \prod_{j=1}^{\infty} (\{0, 1\}_j, 2^{\{0,1\}_j}, p_j(\{1\}) = p) \right] \times (W, 2^W, \{p_k\}_{k=0}^{\infty})$$

where  $W = \{0, 1, \dots\}$ . Each  $\omega \in \Omega$  can be viewed as an ordered pair  $(x, k)$  where  $x$  is an infinite sequence of 0's and 1's. Define  $\nu(\omega) = \nu((x, k)) = k$ , so that  $\Pr(\{\omega : \nu(\omega) = k\}) = p_k$ . If  $x_i$  is the  $i^{\text{th}}$  coordinate of  $x$  then

$$X(\omega) = X((x, k)) = \sum_{i=1}^{\nu((x,k))} x_i = \sum_{i=1}^k x_i.$$

To calculate  $E[X]$  we use Fubini's Theorem with  $W$  as one factor:

$$\begin{aligned} E[X] &= \int_W \left( \int_{\prod_{j=1}^{\infty} \{0,1\}_j} X \right) \\ &= E[p\nu] \\ &= pE[\nu]. \end{aligned}$$

The other moments of  $X$  can be computed in a similar manner.

## Chapter 4

# Convergence in Distribution

These will all be problems in analysis in the real numbers.

**Definition 68 (Convergence in distribution)** Let  $\{F_n\}_{n=1}^\infty$  be a sequence of distribution functions on the real line, and let  $F$  be another distribution function on the real line. We say that  $\{F_n\}_{n=1}^\infty$  converges to  $F$  **in distribution** if and only if  $F_n(x) \rightarrow F(x)$  as  $n \rightarrow \infty$  at every point  $x$  of continuity of  $F$ .

If  $F_n$  is the distribution function of the real-valued random variable  $X_n$  and  $F$  is the distribution function of the real-valued random variable  $X$  then we say that  $X_n$  converges to  $X$  **in distribution** if the  $F_n$  converge to  $F$  in distribution.

Note that this is very, very weak type of convergence.

**Example:** Let  $X$  be  $\mathcal{N}(0, 1)$  on  $(\Omega, \mathcal{F}, \text{Pr})$ . Let  $X_k = (-1)^k X$  and let  $F_k$  be the distribution function of  $X_k$ . Then all the  $F_k$  are the same, but the sequence of random variables  $\{X_k\}_{k=1}^\infty$  only converges on a set of probability 0.

Note, however, that if  $\{X_k\}_{k=1}^\infty$  only converges on a set of probability 1, then  $\{X_k\}_{k=1}^\infty$  converges in distribution as well.

Also, a sequence of random variables converges with probability 1 to 0 if and only if it converges in distribution to 0.

**Example:** Select a point at random from  $\{1/n, 2/n, \dots, n/n\}$ , and call the resulting distribution function  $F_n$ . Let  $F$  denote the distribution function of the continuous uniform distribution on  $[0, 1]$ . Then for each  $x \in (-\infty, \infty)$ ,  $F_n(x) \rightarrow F(x)$ .

The main source of examples is from the theory of sums of independent random variables.

**Definition 69 (Convolution:)** If  $F$  and  $G$  are probability distribution functions we define the **convolution** of  $F$  and  $G$ , denoted  $F * G$  by

$$F * G(x) = \int_{\mathbb{R}} F(x-t) dG(t).$$

If  $F$  and  $G$  have densities  $f$  and  $g$  respectively, then  $F * G$  has density

$$f * g(x) = \int_{\mathbb{R}} f(x-t)g(t) dt.$$

**Theorem 70** Let  $X$  and  $Y$  be independent random variables on  $(\Omega, \mathcal{F}, \Pr)$  with distribution functions  $F$  and  $G$  respectively. Then  $F * G$  is the distribution function of  $X + Y$ , and

$$F * G(t) = \mathbb{E}[F(t - Y)]$$

**Proof:**

$$\begin{aligned} \Pr(\{\omega : X(\omega) + Y(\omega) \leq t\}) &= \mathbb{E}[I_{\{\omega : X(\omega) + Y(\omega) \leq t\}}] \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} I_{\{(x,y) : x+y \leq t\}} dF(x) dG(y) \\ &= \int_{\mathbb{R}} \left( \int_{-\infty}^{t-y} 1 dF(x) \right) dG(y) \\ &= \int_{\mathbb{R}} F(t - y) dG(y) \\ &= \mathbb{E}[F(t - Y)] \end{aligned}$$

**Corollary 71**  $F * G = G * F$

**Theorem 72** Let  $X$  and  $Y$  be independent random variables on  $(\Omega, \mathcal{F}, \Pr)$  with distribution functions  $F$  and  $G$  respectively, and that  $G$  has density  $g$ . Then  $F * G$  has density

$$\int_{\mathbb{R}} g(v - t) dF(t).$$

**Proof:** From Theorem 70 and its Corollary,

$$\begin{aligned} \Pr(X + Y \leq x) &= \int_{\mathbb{R}} G(x - t) dF(t) \\ &= \int_{\mathbb{R}} \left( \int_{-\infty}^{x-t} g(u) du \right) dF(t) \\ &= \int_{\mathbb{R}} \left( \int_{-\infty}^x g(v - t) dv \right) dF(t) \\ &= \left( \int_{-\infty}^x \left( \int_{\mathbb{R}} g(v - t) dF(t) \right) dv \right) \end{aligned}$$

which proves the theorem. Note that the last step follows from Fubini's Theorem. **QED**

Now that we have defined the convolution of two distributions, it is easy to see that the convolution is both associative and commutative, by its relation to the sum of independent random variables. It is easy to see that we may generalize to the sum of  $n$  independent random variables and the convolution of  $n$  distribution functions.

**Theorem 73** If  $\{X_j\}_{j=1}^N$  are  $N$  independent random variables with corresponding distribution functions  $F_j$ , then  $X_1 + \cdots + X_N$  has as its distribution function  $F_1 * \cdots * F_N$ .

Now, suppose that  $\{X_n\}_{n=1}^{\infty}$  is a sequence of independent and identically distributed random variables on  $(\Omega, \mathcal{F}, \Pr)$  with  $\mathbb{E}[X_n] = 0$  and  $\text{Var}[X_n] = \sigma^2 > 0$ . It is natural to study the random variables

$$Y_n = \frac{\sum_{k=1}^n X_k}{\sqrt{n\sigma^2}}$$

since  $E[Y_n] = 0$  and  $\text{Var}[Y_n] = 1$ .

We shall show that the sequence of distribution functions of such  $Y_n$  must have a subsequence which converges in distribution. The content of the **Central Limit Theorem** is that not only does a subsequence converge, but in fact the full sequence converges in distribution to the  $\mathcal{N}(0, 1)$  distribution function,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{u^2}{2}\right) du.$$

Note that if  $G_n$  is the distribution function of  $Y_n$  and  $F^{(n)}$  is the distribution function of  $X_1 + \cdots + X_n$  then

$$\begin{aligned} G_n(x) &= \Pr\left(\frac{\sum_{k=1}^n X_k}{\sqrt{n\sigma^2}} \leq x\right) \\ &= F^{(n)}(x\sqrt{n\sigma^2}). \end{aligned}$$

Hence the Central Limit Theorem may be stated as follows:

**Theorem 74 (Central Limit Theorem):** *Let  $F$  be any distribution function with*

$$\begin{aligned} \int_{\mathbb{R}} x dF(x) &= 0 \\ \int_{\mathbb{R}} x^2 dF(x) &= \sigma^2 \in (0, \infty). \end{aligned}$$

*Let  $F^{(n)}$  be the  $n$ -fold convolution of  $F$  with itself. Then for each  $x \in (-\infty, \infty)$ ,  $F^{(n)}(x\sqrt{n\sigma^2}) \rightarrow \Phi(x)$  as  $n \rightarrow \infty$ .*

There are, however, many other instances of convergence in distribution to distribution other than  $\mathcal{N}(0, 1)$ . One such example is the distribution of the maximum of positive iid random variables.

Suppose that  $X_1, X_2, \dots$  are iid positive random variables with distribution function  $F(x)$  and put  $G(x) = 1 - F(x)$ . Put  $M_n = \max\{X_1, \dots, X_n\}$ . We should like to find a (deterministic) sequence  $b_n$  so that  $M_n/b_n$  converges in distribution. Observe that

$$\Pr(M_n/b_n \leq x) = (F(b_n x))^n = (1 - G(b_n x))^n.$$

From elementary calculus we see that we need  $nG(b_n x)$  to converge to  $g(x)$  for some  $g(x)$  so that we will have

$$\lim_{n \rightarrow \infty} \Pr(M_n/b_n \leq x) = \exp(-g(x)).$$

For example, if  $G(x) = c/x^p$  for  $x > 1$ , then  $b_n = n^{1/p}$  gives  $g(x) = cx^{-p}$  and

$$\Pr(M_n/b_n \leq x) = \exp(-cx^{-p}).$$

Another example is the so-called **Arcsine Law**. (The following is a condensed version of material found in Chapter 3 of Volume I of W. Feller's book.)

Let  $\{X_n\}_{n=1}^{\infty}$  be an independent and identically distributed sequence of random variables on  $(\Omega, \mathcal{F}, \Pr)$  and let  $S_n$  be their  $n^{\text{th}}$  partial sum. Let  $\xi$  be the indicator function for  $(0, \infty)$  and define the random variable  $N_n$  to be

$$\sum_{k=1}^n \xi(S_k).$$

$N_n$  is simply the number of times that  $S_n$  is positive in the first  $n$  trials. Thus a reasonable normalization of  $N_n$  is to divide  $N_n$  by  $n$  to obtain a random variable which is always between 0 and 1. Thus the sequence  $\text{Var}[N_n/n]$  must be bounded. Let us calculate  $E[N_n/n]$ . First,

$$\begin{aligned} E[N_n] &= E\left[\sum_{k=1}^n \xi(S_k)\right] \\ &= \sum_{k=1}^n \Pr(\{\omega : S_k(\omega) > 0\}). \end{aligned}$$

Let  $a_k = \Pr(\{\omega : S_k(\omega) > 0\})$ . If we can show that  $a_k \rightarrow \alpha$  then  $E[N_n/n] \rightarrow \alpha$  as well.

On the other hand,

$$\begin{aligned} E[N_n^2] &= E\left[\sum_{k=1}^n \sum_{j=1}^n \xi(S_k)\xi(S_j)\right] \\ &= \sum_{k=1}^n \sum_{j=1}^n [\xi(S_k)\xi(S_j)] \\ &= \sum_{k=1}^n a_k + 2 \sum_{k=2}^n \sum_{j=1}^{k-1} \Pr(\{\omega : S_k(\omega) > 0, S_j(\omega) > 0\}) \\ &= \sum_{k=1}^n a_k + 2 \sum_{k=1}^{n-1} \sum_{j=1}^{n-k} \Pr(\{\omega : S_k(\omega) > 0, S_{k+j}(\omega) > 0\}) \\ &= \sum_{k=1}^n a_k + \sum_{k=1}^{n-1} \sum_{j=1}^{n-k} \Pr(\{\omega : S_k(\omega) > 0, S_{k+j}(\omega) > 0\}) \\ &\quad + \sum_{k=1}^{n-1} \sum_{j=1}^{n-k} \Pr(\{\omega : S_j(\omega) > 0, S_{k+j}(\omega) > 0\}) \end{aligned}$$

From the homework we know that if  $X$  and  $Y$  are independent random variables, then

$$\begin{aligned} &\Pr(\{\omega : X(\omega) > 0, X(\omega) + Y(\omega) > 0\}) + \Pr(\{\omega : Y(\omega) > 0, X(\omega) + Y(\omega) > 0\}) \\ &= \\ &\Pr(\{\omega : X(\omega) > 0\}) \Pr(\{Y(\omega) > 0\}) + \Pr(\{\omega : X(\omega) + Y(\omega) > 0\}) \end{aligned}$$

(Proof uses indicator functions and properties of expectation.)

This gives us

$$E\left[\left(\frac{N_n}{n}\right)^2\right] = \frac{1}{n^2} \left( \sum_{k=1}^n a_k + \sum_{k=1}^{n-1} \sum_{l=1}^{n-k} (a_k a_l + a_{k+l}) \right).$$

Now, suppose that

$$\lim_{k \rightarrow \infty} a_k = \alpha.$$

Then we see that

$$\lim_{n \rightarrow \infty} \mathbb{E} \left[ \left( \frac{N_n}{n} \right)^2 \right] = 0 + \frac{\alpha^2}{2} + \frac{\alpha}{2}.$$

Therefore

$$\lim_{n \rightarrow \infty} \text{Var} \left( \frac{N_n}{n} \right) = \frac{\alpha - \alpha^2}{2}$$

and this limiting variance is positive if and only if  $\alpha \in (0, 1)$ . This, in turn, leads us to suspect that for all  $x$ ,

$$\Pr \left( \left\{ \omega : \frac{N_n(\omega)}{n} \leq x \right\} \right)$$

converges to  $F(x)$ . In fact, the arcsine law says that if

$$\lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n a_k = \frac{1}{2}$$

then

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{2}{\pi} \arcsin(\sqrt{x}) & \text{if } x \in [0, 1] \\ 1 & \text{if } x \geq 1. \end{cases}$$

## 4.1 Convergence in Distribution in One Dimension

**Theorem 75** Let  $\{F_n\}_{n=1}^{\infty}$  be a sequence of distribution functions and let  $F$  be another distribution function. Then  $F_n(x)$  converges to  $F(x)$  for all  $x \in D$ , a dense subset of  $(-\infty, \infty)$  if and only if  $F_n(x)$  converges to  $F(x)$  at every  $x$  which is a continuity point of  $F$ .

**Proof:** Suppose that  $F_n(x)$  converges to  $F$  at every point of continuity of  $F$ .  $F$  has at most a countable number of points of discontinuity, so  $F_n(x)$  converges to  $F(x)$  for all  $x$  in the real numbers.

Now, suppose that the convergence set  $D$  of  $F_n(x)$  is dense in the real numbers. Let  $x$  be a point of continuity of  $F$ . Choose  $\epsilon > 0$ . Then there exist  $\underline{x} \leq x \leq \bar{x} \in D$  such that

$$\begin{aligned} F(\bar{x}) - F(x) &< \epsilon \\ F(x^-) - F(\underline{x}) &< \epsilon \end{aligned}$$

Thus we have

$$\begin{aligned} \limsup_{n \rightarrow \infty} F_n(x) &\leq \limsup_{n \rightarrow \infty} F_n(\bar{x}) \\ &= F(\bar{x}) \\ &\leq F(x) + \epsilon \end{aligned}$$

and

$$\begin{aligned} \liminf_{n \rightarrow \infty} F_n(x) &\geq \liminf_{n \rightarrow \infty} F_n(\underline{x}) \\ &= F(\underline{x}) \\ &\geq F(x) - \epsilon. \end{aligned}$$

Since  $\epsilon$  was arbitrary,

$$\liminf_{n \rightarrow \infty} F_n(x) = F(x) = \limsup_{n \rightarrow \infty} F_n(x)$$

so  $F_n(x)$  converges to  $F(x)$ , as claimed. **QED**

**Theorem 76** Let  $\{F_n\}_{n=1}^{\infty}$  be a sequence of distribution functions and let  $F$  be another distribution function. Then  $F_n$  converges to  $F$  in distribution if and only if

$$\lim_{n \rightarrow \infty} \int_{(-\infty, \infty)} f(x) dF_n(x) = \int_{(-\infty, \infty)} f(x) dF(x) \quad (4.1)$$

for every bounded continuous function  $f : (-\infty, \infty) \rightarrow (-\infty, \infty)$ .

**Proof:** Suppose that  $F_n$  converges to  $F$  in distribution. Let  $D$  be the set where all the  $F_n$  and  $F$  are continuous. We know that  $D$  is dense in the real numbers because it is the complement of a countable set of real numbers. Let  $f$  be a given bounded continuous function, and let  $M = \sup_x |f(x)|$ . Choose  $\epsilon > 0$ . Then there exist  $a < b \in D$  so that

$$\int_{[a,b]^c} dF(x) < \epsilon,$$

and there is a positive integer  $N(\epsilon)$  so that if  $n \geq N(\epsilon)$  then

$$\int_{[a,b]^c} dF_n(x) < 2\epsilon.$$

Thus, for all  $n \geq N(\epsilon)$ ,

$$\left| \int_{(-\infty, \infty)} f(x) dF_n(x) - \int_{(-\infty, \infty)} f(x) dF(x) \right| \leq \left| \int_{[a,b]} f(x) dF_n(x) - \int_{[a,b]} f(x) dF(x) \right| + 3M\epsilon.$$

Now, since  $f$  is continuous, it is uniformly continuous on  $[a, b]$ . Therefore we can find a step function,  $\phi_\epsilon$  such that for all  $x \in [a, b]$ ,

$$|\phi_\epsilon(x) - f(x)| \leq \epsilon,$$

and the points of discontinuity of  $\phi_\epsilon$  are all elements of  $D$ . Let  $\{x_j\}_{j=1}^l$  be these points of discontinuity, and let  $\bar{x}_j$  denote the midpoint of  $[x_j, x_{j+1}]$ . Note that

$$\int_{x_j}^{x_{j+1}} \phi_\epsilon(x) dF_n(x) = \phi_\epsilon(\bar{x}_j)(F_n(x_{j+1}) - F_n(x_j))$$

Applying this to the preceding inequality and using the triangle inequality we obtain

$$\begin{aligned} \left| \int_{(-\infty, \infty)} f(x) dF_n(x) - \int_{(-\infty, \infty)} f(x) dF(x) \right| &\leq 3M\epsilon + \left| \int_{[a,b]} f(x) dF_n(x) - \int_{[a,b]} \phi_\epsilon(x) dF_n(x) \right| \\ &\quad + \left| \int_{[a,b]} \phi_\epsilon(x) dF_n(x) - \int_{[a,b]} \phi_\epsilon(x) dF(x) \right| \\ &\quad + \left| \int_{[a,b]} \phi_\epsilon(x) dF(x) - \int_{[a,b]} f(x) dF(x) \right| \\ &\leq 3M\epsilon + \epsilon + o(1) + \epsilon. \end{aligned}$$

Since  $\epsilon$  was arbitrary, (4.1) holds.

Now, the converse. We will not need to assume that (4.1) holds for all bounded continuous functions. We only need non-increasing continuous functions  $f$  which satisfy  $f(x) = 1$  for  $x \leq a$  and  $f(x) = 0$  for  $x \geq b$  for some  $a < b$ . If all such functions satisfy (4.1) then  $F_n$  converges to  $F$  in distribution.

So, suppose that  $x$  is a point of continuity of  $F$ . Without loss of generality, we may assume  $x = 0$ , since the translate of a distribution function is a distribution function. For each  $\delta > 0$ , take  $f_\delta$  to be of the type discussed in the preceding paragraph, where  $a = 0$  and  $b = \delta$ . Then we have

$$F_n(0) = \int_{(-\infty, 0]} f_\delta(x) dF_n(x) \leq \int_{(-\infty, \infty)} f_\delta(x) dF_n(x)$$

so

$$\begin{aligned} \limsup_{n \rightarrow \infty} F_n(0) &\leq \limsup_{n \rightarrow \infty} \int_{(-\infty, \infty)} f_\delta(x) dF_n(x) \\ &= \int_{(-\infty, \infty)} f_\delta(x) dF(x) \\ &\leq \int_{(-\infty, \delta]} 1 dF(x) \\ &= F(\delta) \end{aligned}$$

Since  $\delta$  was arbitrary, and distribution functions are continuous from the right, we have

$$\limsup_{n \rightarrow \infty} F_n(0) \leq F(0).$$

For the inequality for the liminf, let  $g_\delta$  be in the same class of functions as  $f_\delta$ , but with  $a = -\delta$  and  $b = 0$ . Then

$$\begin{aligned} \liminf_{n \rightarrow \infty} F_n(0) &\geq \limsup_{n \rightarrow \infty} \int_{(-\infty, \infty)} g_\delta(x) dF_n(x) \\ &= \int_{(-\infty, \infty)} g_\delta(x) dF(x) \\ &\geq F(-\delta). \end{aligned}$$

Since  $F$  is continuous at 0 we have

$$\liminf_{n \rightarrow \infty} F_n(0) \geq F(0)$$

as well. **QED**

**Corollary 77** *If  $F_n$  converges to  $F$  in distribution and  $f$  is a bounded function with a finite number of discontinuities then (4.1) still holds.*

**Proof:** Examination of the proof of Theorem 76 shows that such a function  $f$  may still be uniformly approximated by step functions in the manner required to prove (4.1). **QED**

## 4.2 Compactness of Distributions and the Helly Selection Principle

The subset of distribution functions is not a compact subset (of the set of functions from the real numbers to the real numbers) as it fails to be sequentially compact (see Munkre's *Topology*). To see this, let  $F_n$  be the distribution function which is 0 for  $x < n$  and 1 for  $x \geq n$ . The sequence  $F_n(x)$  converges to 0 for every  $x$ , so the limiting function is not a distribution function. Since the sequence converges to the function which is constantly 0, so does every subsequence, so there is no subsequence converging to a distribution function. The problem is that the domains of these distribution functions, that is the real numbers, is not compact. Helly's theorem gives a partial solution to this inconvenient problem.

**Definition 78 (Subdistribution Function)**  $F(x)$  is a **subdistribution function** if and only if

1.  $F$  is monotone;
2.  $F$  is right continuous;
3.  $0 \leq F(-\infty) \leq F(\infty) \leq 1$ .

**Theorem 79 (Helly's Theorem)** Let  $\{F_n\}_{n=1}^{\infty}$  be a sequence of distribution functions on the real line. Then there is a subdistribution function  $F$  and subsequence  $\{F_{n_k}\}_{k=1}^{\infty}$  of  $\{F_n\}_{n=1}^{\infty}$  which converges to  $F$  at every point of continuity of  $F$ .

**Proof:** Let  $D \equiv \{x_j\}_{j=1}^{\infty}$  be a countable dense subset of the real numbers. Since  $\{F_n(x_1)\}_{n=1}^{\infty}$  is a bounded sequence of real numbers it contains a convergent subsequence. Hence there is a subsequence  $\{F_{1,n}\}_{n=1}^{\infty}$  of  $\{F_n\}_{n=1}^{\infty}$  such that

$$\lim_{n \rightarrow \infty} F_{1,n}(x_1) = y_1 \in [0, 1].$$

In turn,  $\{F_{1,n}(x_2)\}_{n=1}^{\infty}$  is a bounded set of real numbers, so it too has a convergent subsequence. Hence there is a subsequence  $\{F_{2,n}\}_{n=1}^{\infty}$  of  $\{F_{1,n}\}_{n=1}^{\infty}$ , and hence of  $\{F_n\}_{n=1}^{\infty}$ , with the property that

$$\lim_{n \rightarrow \infty} F_{2,n}(x_1) = y_j \in [0, 1]$$

for  $j \in \{1, 2\}$ . Continuing inductively, for each positive integer  $m$  we can find a subsequence  $\{F_{m,n}\}_{n=1}^{\infty}$  of  $\{F_n\}_{n=1}^{\infty}$  with the property that

$$\lim_{n \rightarrow \infty} F_{m,n}(x_1) = y_j \in [0, 1]$$

for  $j \in \{1, \dots, m\}$ . Now define  $F_D : D \rightarrow [0, 1]$  by  $F_D(x_j) = y_j$ , and let  $F_{n_k} = F_{k,k}$ , the so-called diagonal sequence. Then for any  $x \in D$ ,  $F_{n_k}(x)$  converges to  $F_D(x)$  as  $k \rightarrow \infty$ . Since each  $F_{k,k}$  is non-decreasing,  $F_D$  is non-decreasing. Now define  $F : (-\infty, \infty) \rightarrow [0, 1]$  by putting  $F(x) = \inf\{F_D(y) : y > x\}$ .  $F$  is clearly non-decreasing. We need only prove that  $F$  is right continuous. Let  $\{z_n\}_{n=1}^{\infty}$  strictly decrease to  $x$ . Then

$$b \equiv \lim_{n \rightarrow \infty} F(z_n) \geq F(x). \tag{4.2}$$

Suppose that  $b > F(x)$ . Let  $y_0 \in D$ ,  $y_0 > x$  satisfy  $F_D(y_0) < b$ . For large enough values of  $n$ ,  $x < z_n < y_0$ , giving  $F(z_n) \leq F(y_0) < b$ , contradicting (4.2). Therefore,  $F$  is right continuous, and, therefore, a subdistribution function.

It remains to show that  $\{F_n\}_{n=1}^\infty$  converges to  $F$  at the continuity points of  $F$ . Let  $x$  be a continuity point of  $F$ , and suppose that  $y \in D$  and  $y > x$ . Then

$$\limsup_{k \rightarrow \infty} F_{n_k}(x) \leq \limsup_{k \rightarrow \infty} F_{n_k}(y) = F_D(y).$$

Now taking the infimum over all such  $y \in D$  we see from the definition of  $F$  that

$$\limsup_{k \rightarrow \infty} F_{n_k}(x) \leq F(x). \quad (4.3)$$

Also, if  $x' < y < x$  and  $y \in D$  we have

$$\begin{aligned} F(x') &\leq F_D(y) \\ &= \lim_{k \rightarrow \infty} F_{n_k}(y) \\ &= \liminf_{k \rightarrow \infty} F_{n_k}(y) \\ &\leq \liminf_{k \rightarrow \infty} F_{n_k}(x). \end{aligned}$$

Letting  $x'$  converge upward to  $x$  we see that

$$F(x) = F(x^-) \leq \liminf_{k \rightarrow \infty} F_{n_k}(x). \quad (4.4)$$

Combining (4.3) and (4.4) shows that

$$F(x) = \lim_{k \rightarrow \infty} F_{n_k}(x),$$

as desired. **QED**

This same result can be obtained by using more abstract theorems from functional analysis. See Rudin, *Real and Complex Analysis*.

Let  $[-\infty, \infty]$  be the two point compactification of the real line. Let  $C^0[-\infty, \infty]$  denote the continuous functions on  $(-\infty, \infty)$  that have finite limits at  $\pm\infty$ . The Riesz Representation Theorem says that every bounded, non-negative linear functional  $\Lambda$  on  $C^0[-\infty, \infty]$  is of the form

$$\Lambda(f) = c_0 \int_{(-\infty, \infty)} f(x) dF(x) + c_1 f(-\infty) + c_2 f(+\infty),$$

where  $F$  is a probability distribution on  $(-\infty, \infty)$  and the constants  $c_i$  are non-negative.

The Banach-Alaoglu Theorem tells us that the space of all such linear functionals is weak-star compact, that is, that every sequence of such linear functionals has a subsequence which is convergent to such a linear functional at every element of  $C^0[-\infty, \infty]$ . We now associate to each  $F_n$  the linear functional  $L_n$  by the rule

$$L_n(f) = \int_{(-\infty, \infty)} f(x) dF_n(x)$$

and let the subsequential limit of the  $F_n$  be the the subdistribution function  $c_0 F$  corresponding to the limiting linear functional  $L$ , and apply Theorem 76.

## Chapter 5

# The Central Limit Theorem and Characteristic Functions

The familiar form of the Central Limit Theorem is that if  $\{X_n\}_{n=1}^{\infty}$  are independent and identically distributed random variables with  $E[X_n] = \mu$  and  $\text{Var}[X_n] = \sigma^2 > 0$  then

$$\lim_{n \rightarrow \infty} \Pr \left( \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq x \right) = \Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-u^2/2) du.$$

The plan from here is as follows:

1. Prove the Central Limit Theorem directly via contraction operators, as done by H. F. Trotter, *An elementary proof of the Central Limit Theorem*, Archiv der Mathematik Vol 10, 1959.
2. Develop the Theory of Fourier Transforms of Probability Distributions.
3. Use Fourier Transforms to give an analytic proof of the Central Limit Theorem.

### 5.1 Preliminaries

1. Let  $C_b$  be the class of bounded, uniformly continuous functions on the real line.
2. Let  $X$  be a real valued random variable and let  $F$  be its distribution function. We shall associate to  $F$  an operator  $T : C_b \rightarrow C_b$  by the rule:

$$T(f)(x) = \int_{-\infty, \infty} f(x+y) dF(y) \equiv E[f(x+X)].$$

Note the following:

- (a) The range of  $T_b$  really is a subset of  $C_b$ . Furthermore, if we put  $\|f\| \equiv \sup\{|f(x)| : -\infty < x < \infty\}$ , the  $T$  is a bounded linear operator on  $C_b$ .
- (b)  $T$  is a weak contraction, that is,  $\|T(f)\| \leq \|f\|$ .

Now, what about compositions? Suppose that  $X$  and  $Y$  are independent random variables with distribution functions  $F$  and  $G$  respectively. Let  $T$  be associated with  $F$  and  $U$  be associated with  $G$  in the manner described above.

**Proposition 80**

$$\begin{aligned} T \circ U &= U \circ T \\ (T \circ U)(f)(x) &= \int_{(-\infty, \infty)} f(x+y) d(F * G)(y) \\ &= E[f(x + X + Y)] \end{aligned}$$

**Proof:** Let  $U(f) = g$ . Then

$$\begin{aligned} (T \circ U)(f)(x) &= T(g)(x) \\ &= E[g(x + X)] \\ &= E[E[f(x + Y + X)]] \\ &= \int_{(-\infty, \infty)} f(x+y) d(F * G)(y) \\ &= E[f(x + X + Y)] \end{aligned}$$

which gives us all the claims in the proposition. **QED**

Now, by induction, if  $\{X_k\}_{k=1}^N$  are independent, and  $T_k$  is associated to  $X_k$ , we have

$$(T_1 \circ \dots \circ T_N)(f)(x) = E[f(X_1 + \dots + X_N + x)].$$

If these  $X_k$  are independent and identically distributed we will write  $T^N$  in place of  $T_1 \circ \dots \circ T_N$ . Note that

$$T^N(f)(0) = E[f(X_1 + \dots + X_N)] = \int_{(-\infty, \infty)} f(y) dF^{(N)}(y).$$

We shall want to rescale at some point by dividing through by  $\sqrt{N}$  in a convenient place.

Now, let  $\{X_n\}_{n=1}^\infty$  be a given sequence of independent and identically distributed random variables on  $(\Omega, \mathcal{F}, \Pr)$  with  $E[X_n] = 0$  and  $\text{Var}[X_n] = 1$ . Let  $F$  be the common distribution function of these random variables. Let  $T_n$  be the operator corresponding to  $X_1/\sqrt{n}$ , or, equivalently, to  $F(\sqrt{n}x)$ . Then according to Proposition 80,  $T_n^n$  corresponds to

$$n^{-1/2}(X_1 + \dots + X_n).$$

For this reason, we study  $T_n^n$ . Finally, let  $S$  be the operator corresponding to the standard normal distribution. We are interested in the behavior of  $T_n^n - S$ .

**Some facts about the standard normal distribution.**

A random variable  $X$  is said to have the standard normal distribution if for every  $x$ ,

$$\Pr(\{\omega : X(\omega) \leq x\}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-u^2/2) du$$

**Definition 81 (Self-similar of index 2)** If  $\{X_k\}_{k=1}^{\infty}$  is a sequence of independent and identically distributed random variables with mean 0, and distribution function  $F$ , and if

$$n^{-1/2}(X_1 + \cdots + X_n)$$

has the same distribution as  $X_1$  for every  $n$ , then we call that distribution  $F$  self-similar of index 2.

The only property of the standard normal distribution that we will use is that it is self-similar of index 2. One can prove this directly via calculus. The Central Limit Theorem shows that it is the only such distribution with a finite second moment.

## 5.2 Trotter's Proof of the Central Limit Theorem

Let  $S_n$  correspond to  $Y/\sqrt{n}$  where  $Y$  has the standard normal distribution. If  $\{Y_n\}_{n=1}^{\infty}$  are independent and identically distributed with the standard normal distribution then  $S_n^n$  corresponds to

$$n^{-1/2}(Y_1 + \cdots + Y_n),$$

which as the standard normal distribution. So we can study  $S_n^n - T_n^n$  which is more natural.

The Central Limit Theorem will be proved once we show that for all functions  $f : (-\infty, \infty) \rightarrow (-\infty, \infty)$  with uniformly continuous first and second derivatives that

$$\lim_{n \rightarrow \infty} \|T_n^n(f) - S_n^n(f)\| = 0.$$

This is a consequence of the proof of Theorem 76.

To finish the proof we will need to prove a few lemmas.

**Lemma 82** Suppose that  $A$  and  $B$  are contraction operators on a normed vector space  $X$  and  $A$  and  $B$  commute. Then if  $x \in X$  we have

$$\|A^n(x) - B^n(x)\| \leq n\|A(x) - B(x)\|.$$

**Proof:** We simply factor:

$$\begin{aligned} A^n(x) - B^n(x) &= (A^n - B^n)(x) \\ &= \left( \sum_{k=0}^{n-1} A^{n-1-k} B^k \right) (A - B)(x) \end{aligned}$$

so

$$\begin{aligned} \|A^n(x) - B^n(x)\| &\leq \left( \sum_{k=0}^{n-1} \|A^{n-1-k} B^k\| \right) \|(A - B)(x)\| \\ &\leq n\|A(x) - B(x)\|, \end{aligned}$$

since  $A$  and  $B$  are contraction operators.

**QED**

**Lemma 83** *If  $f \in C^2(\mathbf{R})$  and  $f$  has uniformly continuous first and second derivatives then*

$$\lim_{n \rightarrow \infty} n \|T_n(f) - S_n(f)\| = 0.$$

**Proof:** Since  $f$  is assumed twice differentiable we may write (via Taylor's Theorem)

$$f(x+y) = f(x) + f'(x)y + \frac{1}{2}y^2 f''(x) + \frac{1}{2}y^2(f''(x+\theta y) - f''(x))$$

where  $\theta \in [0, 1]$  and  $\theta$  depends on  $x$  and  $y$ . We then have

$$\begin{aligned} T_n(f)(x) &= E[f(x + n^{-1/2}X)] \\ &= E[f(x) + f'(x)n^{-1/2}X + \frac{1}{2n}X^2 f''(x) + \frac{1}{2n}X^2(f''(x + \theta n^{-1/2}X) - f''(x))] \\ &= f(x) + f'(x)n^{-1/2}E[X] + \frac{1}{2n}E[X^2]f''(x) + \frac{1}{2n}E[X^2(f''(x + \theta n^{-1/2}X) - f''(x))] \\ &= f(x) + \frac{1}{2n}f''(x) + \frac{1}{2n}E[X^2(f''(x + \theta n^{-1/2}X) - f''(x))] \end{aligned}$$

since  $E[X] = 0$  and  $E[X^2] = \text{Var}[X] = 1$ . Therefore

$$n \left\| T_n(f) - f - \frac{1}{2n}f'' \right\| \leq \left\| E\left[\frac{1}{2}X^2|(f''(x + \theta n^{-1/2}X)) - f''(x)|\right] \right\|$$

We shall show that this larger quantity converges to 0 as  $n \rightarrow \infty$ .

Let  $M > 0$  be given. Then

$$\begin{aligned} R_n(x) &\equiv E\left[\frac{1}{2}X^2|(f''(x + \theta n^{-1/2}X)) - f''(x)|\right] \\ &= E\left[\frac{1}{2}X^2|(f''(x + \theta n^{-1/2}X)) - f''(x)|; |X| \leq M\right] \\ &\quad + E\left[\frac{1}{2}X^2|(f''(x + \theta n^{-1/2}X)) - f''(x)|; |X| > M\right] \end{aligned}$$

Let  $A = \|f''\|$  and let

$$\epsilon(t) = \sup\{|f''(x+y) - f''(x)| : x \in \mathbf{R}, |y| < t\}$$

Since the second derivative of  $f$  is assumed to be uniformly continuous,  $\epsilon(t) \rightarrow 0$  as  $t \rightarrow 0$ . (This should remind you of the proof of the Weierstrass Approximation Theorem, Theorem 18.) Therefore

$$R_n(x) \leq \frac{1}{2}\epsilon(n^{-1/2}M) + AE[X^2; |X| \geq M].$$

As this estimate of  $R_n(x)$  is independent of  $x$ , we have

$$\|R_n\| \leq \frac{1}{2}\epsilon(n^{-1/2}M) + AE[X^2; |X| \geq M],$$

and therefore

$$\limsup_{n \rightarrow \infty} \|R_n\| \leq AE[X^2; |X| \geq M],$$

for any  $M$ . However,

$$\lim_{M \rightarrow \infty} \mathbb{E}[X^2 : |X| > M] = 0.$$

Therefore

$$\limsup_{n \rightarrow \infty} \|R_n\| = 0.$$

Since a random variable with the standard normal distribution has mean 0 and variance 1, the preceding holds when  $T_n$  is replaced with  $S_n$ . Therefore

$$\begin{aligned} \lim_{n \rightarrow \infty} n \|T_n(f) - S_n(f)\| &\leq \lim_{n \rightarrow \infty} n \|T_n(f) - f - \frac{1}{2n} f''\| + \lim_{n \rightarrow \infty} n \|S_n(f) - f - \frac{1}{2n} f''\| \\ &= 0, \end{aligned}$$

as desired. QED

Now we can prove

**Theorem 84 (Central Limit Theorem)** *Let  $F$  be a distribution function with*

$$\begin{aligned} \int_{\mathbf{R}} x \, dF(x) &= 0, \\ \int_{\mathbf{R}} x^2 \, dF(x) &= 1, \end{aligned}$$

and let  $F^{(n)}$  be the  $n$ -fold convolution of  $F$  with itself. Then  $F^{(n)}(\sqrt{n}x) \rightarrow \Phi(x)$  for every real number  $x$ .

**Proof:** Let  $T_n$  and  $S_n$  be as above. Then from Lemmas 82 and 83 we see that

$$\lim_{n \rightarrow \infty} \|T_n^n(f) - S_n^n(f)\| = 0.$$

for all  $f \in C^2(\mathbf{R})$  which are uniformly continuous and which have uniformly continuous second derivatives. Thus for all such  $f$ , letting  $F_n(x) = F^{(n)}(\sqrt{n}x)$ ,

$$\int_{\mathbf{R}} f(x) \, dF_n(x) \rightarrow \int_{\mathbf{R}} f(x) \, d\Phi(x).$$

The Central Limit Theorem then follows from the proof of Theorem 76. QED

### 5.3 Extensions of the Central Limit Theorem and Related Questions

1. The extension to vector-valued random variables is straightforward, as the preceding proof easily generalizes to higher dimensions.
2. What can be said in the case where the random variables are independent, but not identically distributed? This very interesting question has been addressed successfully by Lindeberg (1922) and Feller (1935).

3. The case of non-independent random variables is very deep and very hard. As one can easily see, the proof above will not work at all.
4. From the practical standpoint, how quickly does the sequence of standardized distribution converge to the standard normal?

The answer to the second question is given by the Lindeberg-Feller Theorem. (See, for example, W. Feller, *An Introduction to Probability Theory and Its Applications, Volume II.*) In 1922 Lindeberg gave sufficient conditions for the conclusion of the Central Limit Theorem to hold. In 1935 Feller showed that these conditions were necessary.

**Definition 85 (The Lindeberg Condition)** Let  $\{X_n\}_{n=1}^\infty$  be a sequence of independent random variables on  $(\Omega, \mathcal{F}, \Pr)$ , such that  $E[X_n] = 0$  for all  $n$ . Let  $\sigma_n^2 = E[X_n^2]$ . Let

$$c_n = \sqrt{\sum_{j=1}^n \sigma_j^2}.$$

Then random variables  $\{X_n\}_{n=1}^\infty$  are said to satisfy the **Lindeberg Condition** if and only if for every  $\delta > 0$ ,

$$\lim_{n \rightarrow \infty} c_n^{-2} \sum_{j=1}^n E[X_j^2; |X_j| > \delta c_n] = 0.$$

**Theorem 86 (Lindeberg-Feller Theorem)** Let  $\{X_n\}_{n=1}^\infty$  be a sequence of independent random variables on  $(\Omega, \mathcal{F}, \Pr)$  with  $E[X_n] = 0$  for all  $n$ .  $\{X_n\}_{n=1}^\infty$  satisfy the Lindeberg Condition if and only if

$$\lim_{n \rightarrow \infty} \Pr\left(\left\{\omega : c_n^{-1} \sum_{j=1}^n X_j(\omega) \leq x\right\}\right) = \Phi(x)$$

for every real number  $x$ .

**Proof:** See the reference cited above, or Chung, *A Course in Probability Theory*, or Ash, *Real Analysis and Probability*.

Note that if the  $\{X_n\}_{n=1}^\infty$  are independent and identically distributed with first moment equal to 0 and second moment finite and positive, the Lindeberg Condition is trivially satisfied, for then  $c_n$  is proportional to  $\sqrt{n}$ , and we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n E[X_j^2; |X_j| > \delta c_n] = E[X_1^2; |X_1| > \delta c \sqrt{n}] = 0.$$

Note also that finite first and second moments are important. If each  $X_n$  has the Cauchy density, which is self-similar with index 1, that is

$$\frac{1}{n} \sum_{k=1}^n X_k$$

has the same distribution as  $X_1$  for every  $n$ , then normalizing the sum by  $\sqrt{n}$  will not lead to a convergent sequence of distribution functions.

## 5.4 Characteristic Functions

**Reference:** Any book on probability, but especially Chapter 15 of Feller's *An Introduction to Probability Theory and Its Applications, Volume II.*)

**Definition 87 (Characteristic Function)** Let  $X$  be a random variable on  $(\Omega, \mathcal{F}, \Pr)$  with distribution function  $F$ . Then the **characteristic function** of  $X$ , denoted by  $\phi_X$  or by  $\hat{F}$ , is a function from  $\mathbf{R}$  to  $\mathbf{C}$  defined by

$$\phi_X(\lambda) = \mathbb{E}[\exp(i\lambda X)] = \int_{\mathbf{R}} \exp(it\lambda) dF(t).$$

If  $X$  is a vector-valued random variable, the  $\phi_X$  is a map from  $\mathbf{R}^n$  to  $\mathbf{C}$  defined by

$$\phi_X(\vec{\lambda}) = \mathbb{E}[\exp(\vec{\lambda} \cdot X)].$$

**Note:** The characteristic function is simply the Fourier transform of the measure  $\mu_F$  induced on the real line by  $X$ . It is well-defined because  $\exp(i\lambda X)$  is a bounded random variable, or, if you like, because  $\mu_F$  is a probability measure on  $\mathbf{R}$ .

## 5.5 Properties of Characteristic Functions

1. If  $X$  has characteristic function  $\phi$  then  $aX + b$  has characteristic function  $\exp(ib\lambda)\phi(a\lambda)$ , since

$$\mathbb{E}[\exp(i\lambda(aX + b))] = \mathbb{E}[\exp(ib\lambda)\exp(ia\lambda X)] = \exp(ib\lambda)\phi(a\lambda).$$

2. Let  $\{X_k\}_{k=1}^N$  be independent random variables on  $(\Omega, \mathcal{F}, \Pr)$ , and let  $\phi_j$  be the characteristic function of  $X_j$ , and let  $\phi$  be the characteristic function of  $X_1 + \cdots + X_N$ . Then

$$\phi(\lambda) = \prod_{k=1}^N \phi_k(\lambda),$$

since

$$\mathbb{E}[\exp(i\lambda \sum_{k=1}^N X_k)] = \mathbb{E}[\prod_{j=1}^N \exp(i\lambda X_j)] = \prod_{j=1}^N \mathbb{E}[\exp(i\lambda X_j)],$$

where the last equality follows from the independence of the  $X_j$ .

3. Let  $\phi$  be the characteristic function of  $X$ . Then  $\phi$  is uniformly continuous on  $\mathbf{R}$ , since

$$|\mathbb{E}[\exp(i(x+y)X)] - \mathbb{E}[\exp(ixX)]| \leq \mathbb{E}[|\exp(iyX) - 1|]$$

which converges to 0 as  $y \rightarrow 0$  by the dominated convergence theorem.

4. Let  $\phi$  be the characteristic function of  $X$ . Then  $\phi$  is real valued if and only if  $X$  and  $-X$  have the same distribution.

Suppose that  $X$  and  $-X$  have the same distribution. Then

$$\phi(\lambda) = \mathbb{E}[\exp(i\lambda(-X))] = \mathbb{E}[\exp(-i\lambda X)] = \overline{\mathbb{E}[\exp(i\lambda X)]} = \overline{\phi(\lambda)}$$

so  $\phi(\lambda)$  is real.

To prove the converse, we need the theorem (proven below) that asserts that two random variables with the characteristic function are identically distributed. Then we just note that the proof of the first half of the proposition shows that the characteristic function of  $-X$  is the complex conjugate of the characteristic function  $X$ . If the characteristic function of  $X$  is real, then  $X$  and  $-X$  have the same characteristic function.

5. Suppose that  $X_1$  and  $X_2$  are random variables on  $(\Omega, \mathcal{F}, \Pr)$  and that

$$E[\exp(itX_1)]E[\exp(isX_2)] = E[\exp(i(tX_1 + sX_2))]$$

for all real ordered pairs  $(t, s)$ . Then  $X_1$  and  $X_2$  are independent.

The proof of this statement also requires the Uniqueness Theorem. The hypothesis states that the characteristic function of the vector random variable  $(X_1, X_2)$  is the same as the characteristic function of an independent pair of random variables whose marginal distributions are those of  $X_1$  and  $X_2$ . Hence the joint distribution of  $(X_1, X_2)$  is the same as that of a pair of independent random variables, so  $X_1$  and  $X_2$  are independent.

**Note:** It is not the case that if  $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$  for all real numbers  $t$  that  $X$  and  $Y$  are independent. As a counter-example, suppose that  $X$  has the Cauchy distribution, that is

$$\Pr(X \leq x) = \frac{1}{\pi} \int_0^x \frac{1}{1+u^2} du.$$

Then

$$\phi_X(t) = \exp(-|t|).$$

(This can be shown either by using the calculus of residues, or by observing that the Cauchy distribution is self-similar of index 1, shown by integration by partial fractions. ESK)

Then  $\phi_{X+X}(t) = \phi(2X)(t) = \exp(-2|t|)$  and if  $X$  and  $Y$  are independent and each has the Cauchy distribution, then  $\phi_{X+Y} = \exp(-2|t|)$  as well.

**Theorem 88 (Uniqueness of Characteristic Functions)** *Let  $X$  and  $Y$  be random variables on  $(\Omega, \mathcal{F}, \Pr)$  with distribution functions  $F$  and  $G$  respectively. If the characteristic function of  $X$  is the same as the characteristic function of  $Y$  then  $F = G$ .*

**Proof:** We shall give the proof for real-valued random variables. The extension to vector-valued random variables is straight-forward.

First,

$$\exp(-itu)\phi_X(u) = \int_{\mathbf{R}} \exp(iu(s-t))dF(s).$$

Next, we integrate both sides of this expression with respect to  $dG$ , treating  $u$  as the variable:

$$\begin{aligned} \int_{\mathbf{R}} \exp(-itu)\phi_X(u)dG(u) &= \int_{\mathbf{R}} \left( \int_{\mathbf{R}} \exp(iu(s-t))dF(s) \right) dG(u) \\ &= \int_{\mathbf{R}} \left( \int_{\mathbf{R}} \exp(iu(s-t))dG(u) \right) dF(s) \quad (\text{Fubini's theorem}) \\ &= \int_{\mathbf{R}} \phi_Y(s-t) dF(s) \end{aligned}$$

giving us the crucial formula:

$$\int_{\mathbf{R}} \exp(-itu) \phi_X(u) dG(u) = \int_{\mathbf{R}} \phi_Y(s-t) dF(s). \quad (5.1)$$

Now, suppose that  $N$  is independent of  $X$  and  $N$  has the standard normal distribution. Let  $G$  be the distribution function of  $N/a$  for any positive real number  $a$ . (The important property of this family of normal distributions is that their characteristic functions can be scaled to be densities.) Then (5.1) says

$$\frac{a}{\sqrt{2\pi}} \int_{\mathbf{R}} \exp(-itu) \phi_X(u) \exp\left(-\frac{a^2 u^2}{2}\right) du = \int_{\mathbf{R}} \exp\left(-\frac{(s-t)^2}{2a^2}\right) dF(s). \quad (5.2)$$

If we multiply both sides of (5.2) by  $(2a^2\pi)^{-1/2}$  we get

$$\frac{1}{2\pi} \int_{\mathbf{R}} \exp(-itu) \phi_X(u) \exp\left(-\frac{a^2 u^2}{2}\right) du = \frac{1}{\sqrt{2a^2\pi}} \int_{\mathbf{R}} \exp\left(-\frac{(s-t)^2}{2a^2}\right) dF(s). \quad (5.3)$$

According to Theorem 72, the expression on the right-hand side of (5.3) is the density of  $X + aN$ .

So now we need a lemma about families of random variables related by a change in scale.

**Lemma 89** *Let  $N$  be any random variable, and let  $Y_a = -aN$  for each positive real number  $a$ . If  $f$  is any bounded, measurable function that is continuous at  $b$ , then*

$$\lim_{a \rightarrow 0} \mathbf{E}[f(Y_a + b)] = f(b)$$

**Proof:**

$$\begin{aligned} \limsup_{a \rightarrow 0^+} |\mathbf{E}[f(Y_a + b)] - f(b)| &= \limsup_{a \rightarrow 0^+} |\mathbf{E}[f(-aN + b) - f(b)]| \\ &\leq \limsup_{a \rightarrow 0^+} \mathbf{E}[|f(-aN + b) - f(b)|] \\ &= 0 \quad (\text{DominatedConvergenceTheorem}) \end{aligned}$$

We apply this lemma to the case at hand as follows. Integrate the righthand side of (5.3) from  $-\infty$  to  $x$ , where  $x$  is a point of continuity of  $F$ , and apply Theorem 72 and then Theorem 70:

$$\begin{aligned} \frac{1}{2\pi} \left( \int_{-\infty}^x \frac{1}{\sqrt{2a^2\pi}} \int_{\mathbf{R}} \exp\left(-\frac{(s-t)^2}{2a^2}\right) dF(s) \right) dt &= \int_{\mathbf{R}} \left( \int_{-\infty}^{x-s} \phi_a(z) dz \right) dF(s) \\ &= \Pr(aN + X \leq x) \\ &= \mathbf{E}[F(x - aN)] \end{aligned}$$

and this last expression converges to  $F(x)$  as  $a \rightarrow 0^+$  by the lemma. Thus we have proven the

**Theorem 90 (Fourier Inversion Formula for Distribution Functions)** *Let  $F$  be a distribution function with characteristic function  $\phi$ . Then at every point of continuity of  $F$ ,*

$$\lim_{a \rightarrow 0^+} \frac{1}{2\pi} \int_{-\infty}^x \left( \int_{\mathbf{R}} \phi(u) \exp\left(-\frac{a^2 u^2}{2}\right) du \right) dt = F(x).$$

Since a distribution function is determined by its values at its points of continuity, Theorem 88 follows from Theorem 90. **QED**

The conclusion of the Inversion Theorem can be strengthened if more is known about the characteristic function  $\phi$ . For example, if  $\phi$  is itself integrable on  $\mathbf{R}$ , and  $F$  has a density,  $f$ , then

$$f(t) = \frac{1}{2\pi} \int_{\mathbf{R}} \exp(-itu)\phi(u) du.$$

For more information about Fourier inversion, see Rudin, *Real and Complex Analysis*.

**Theorem 91 (The Convergence Theorem for Characteristic Functions)** *Let  $\{F_n\}_{n=1}^{\infty}$  be a sequence of distribution functions which converge weakly to the distribution function  $F$ . Let  $\phi_n$  be the characteristic function of  $F_n$  and let  $\phi$  be the characteristic function of  $F$ . Then for each real number  $t$ ,  $\phi_n(t) \rightarrow \phi(t)$  as  $n \rightarrow \infty$ .*

**Proof:** Fix a value of  $t$  and let  $h_t(x) = \exp(itx)$ . Now apply Theorem 76. **QED**

**Theorem 92 (Levy's Continuity Theorem)** *Let  $\{F_n\}_{n=1}^{\infty}$  be a sequence of distribution functions, and let  $\phi_n$  be the characteristic function of  $F_n$ . Suppose that for each real number  $t$ , the sequence  $\phi_n(t)$  converges as  $n \rightarrow \infty$ . Let  $\phi(t)$  denote the limit, and suppose that the function  $\phi$  is continuous at 0. Then  $\phi$  is a characteristic function, and  $\{F_n\}_{n=1}^{\infty}$  converges weakly to the distribution function whose characteristic function is  $\phi$ .*

**Remark:** This is a useful converse to the Convergence Theorem. The proof given here may be due to W. Feller.

**Proof:** From the Helly Selection Theorem (Theorem 79) there is a subsequence of  $\{F_n\}_{n=1}^{\infty}$  that converges to a subdistribution function  $G$ , at the points of continuity of  $G$ , which form a dense subset of  $\mathbf{R}$ . Let  $c = G(\infty) - G(-\infty)$ . Since  $G$  is monotone increasing,  $c$  is the total variation of  $G$ . Let  $\{G_n\}_{n=1}^{\infty}$  be a subsequence of  $\{F_n\}_{n=1}^{\infty}$  that converges to  $G$ .

It then follows from (5.3) (in the proof of Theorem 88) applied to each  $G_n$ , the dominated convergence theorem, and the proof of Theorem 76 that

$$\frac{1}{2\pi} \int_{\mathbf{R}} \exp(-itu)\phi(u) \exp\left(-\frac{a^2u^2}{2}\right) du = \frac{1}{\sqrt{2a^2\pi}} \int_{\mathbf{R}} \exp\left(-\frac{(t-s)^2}{2a^2}\right) dG(s)$$

for any  $a > 0$ .

Next, observe that we can write  $G(s) = cH(s)$  where  $H$  is a distribution function. We need to prove that  $c = 1$ . We have

$$\frac{1}{2\pi} \int_{\mathbf{R}} \exp(-itu)\phi(u) \exp\left(\frac{-a^2u^2}{2}\right) du = c \frac{1}{\sqrt{2a^2\pi}} \int_{\mathbf{R}} \exp\left(-\frac{(t-s)^2}{2a^2}\right) dH(s),$$

which is equivalent to

$$\frac{a}{\sqrt{2\pi}} \int_{\mathbf{R}} \exp(-itu)\phi(u) \exp\left(\frac{-a^2u^2}{2}\right) du = c \int_{\mathbf{R}} \exp\left(-\frac{(t-s)^2}{2a^2}\right) dH(s).$$

Since the integral on the righthand side of this equation is between 0 and 1 we have

$$\begin{aligned} 1 &\geq c \\ &\geq \frac{a}{\sqrt{2\pi}} \int_{\mathbf{R}} \exp(-itu) \phi(u) \exp\left(-\frac{a^2 u^2}{2}\right) du \\ &= \frac{1}{\sqrt{2\pi}} \int_{\mathbf{R}} \exp\left(-\frac{itv}{a}\right) \phi\left(\frac{v}{a}\right) \exp\left(-\frac{v^2}{2}\right) dv \end{aligned}$$

Let  $a \rightarrow \infty$  in this double inequality, and it follows from the dominated convergence theorem that

$$1 \geq c \geq \phi(0) = 1,$$

where we know that  $\phi(0) = 1$  since  $\phi_n(0) = 1$  for all  $n$ . Therefore  $\{G_n\}_{n=1}^{\infty}$  converges to the distribution function  $H$  at all points of continuity of  $H$ , and  $\phi$  is the characteristic function of  $H$ , by the Uniqueness Theorem and the Correspondence Theorem.

Now, if we have any other weakly convergent subsequence of  $\{F_n\}_{n=1}^{\infty}$  it must also converge to  $H$ , since the preceding argument shows that its limit is a distribution function whose characteristic function is  $\phi$ . To finish the proof, we need the following theorem:

**Theorem 93** *Let  $\{F_n\}_{n=1}^{\infty}$  be a sequence of distribution functions, and suppose that any weakly convergent subsequence of this sequence converges to the same subdistribution function  $F$ . Then the sequence converges weakly to  $F$ .*

**Proof:** Suppose not. Let  $x_0$  be a point of continuity of  $F$  such that  $F_n(x_0)$  does not converge to  $F(x_0)$  as  $n \rightarrow \infty$ . The sequence of real numbers  $\{F_n(x_0)\}_{n=1}^{\infty}$  is bounded, so it has a convergent subsequence whose limit,  $L$ , is not  $F(x_0)$ . Therefore there is a subsequence of  $\{F_n\}_{n=1}^{\infty}$  which converges to  $L$  when evaluated at  $x_0$ . This subsequence has a weakly convergent subsequence. Since this subsequence is a subsequence of the original sequence, it must converge to  $F(x_0)$  when evaluated at  $x_0$ . This is a contradiction. **QED.**

### 5.5.1 Differentiability of characteristic functions and moments

**Theorem 94** *Let  $X$  be a real valued random variable, let  $\phi$  be its characteristic function, and let  $k$  be a positive integer. If  $E[|X|^k] < \infty$  then  $\phi$  is  $k$  times differentiable, and the  $k^{\text{th}}$  is uniformly continuous. In particular,  $i^k E[X^k] = \phi^{(k)}(0)$ .*

**Proof:** Since

$$\int_0^x i \exp(iu) du = \exp(ix) - 1$$

we have  $|\exp(ix) - 1| \leq |x|$ , so

$$\left| \frac{\exp(i(x+h)X) - \exp(ixX)}{h} \right| \leq |X|$$

Therefore, since

$$\frac{\phi(x+h) - \phi(x)}{h} = E \left[ \frac{\exp(i(x+h)X) - \exp(ixX)}{h} \right]$$

it follows from the dominated convergence theorem that

$$\phi'(x) = i\mathbb{E}[X \exp(ixX)].$$

This argument may be repeated, up to the  $k^{\text{th}}$  derivative, giving for  $j \leq k$ ,

$$\phi^{(j)}(x) = i^j \mathbb{E}[X^j \exp(ixX)]$$

which is a uniformly continuous function of  $x$ . (See Theorem 35 and its proof.)

**QED**

Now, there is also a partial converse.

**Theorem 95** *Suppose that  $X$  is a real valued random variable with characteristic function  $\phi$ . If*

$$\frac{\phi'(-t) - \phi'(t)}{t}$$

*is bounded for  $t \in (0, A)$  for some  $A > 0$  then  $\mathbb{E}[X^2] < \infty$ .*

**Proof:** Let  $s_k \in (0, A)$  be a sequence which converges to 0. It follows from the generalized mean value theorem that there is a sequence  $t_k \in (0, A)$  tending to 0 such that

$$\frac{2 - \phi(-s_k) - \phi(s_k)}{s_k^2} = \frac{\phi'(-t_k) - \phi'(t_k)}{2t_k}.$$

Therefore

$$\frac{2 - \phi(-s_k) - \phi(s_k)}{s_k^2}$$

is a bounded sequence. Observe that

$$X^2 = \lim_{u \rightarrow 0^+} \frac{2 - \exp(-iuX) - \exp(iuX)}{u^2}.$$

Hence by Fatou's lemma,

$$\begin{aligned} \mathbb{E}[X^2] &= \mathbb{E} \left[ \lim_{k \rightarrow \infty} \frac{2 - \exp(-is_k X) - \exp(is_k X)}{s_k^2} \right] \\ &\leq \liminf_{k \rightarrow \infty} \frac{2 - \phi(-s_k) - \phi(s_k)}{s_k^2} \\ &< \infty \end{aligned}$$

as desired.

**QED**

**Corollary 96** *Suppose that  $X$  is a real valued random variable with characteristic function  $\phi$ . If  $\phi$  is twice differentiable at 0 then  $X$  has a finite second moment, and  $\phi''$  exists and is uniformly continuous on  $\mathbf{R}$ .*

**Proof:** Since  $\phi$  is twice differentiable at 0 the hypotheses of the preceding theorem are satisfied.

**QED**

**Remark:** This corollary may be generalized for higher even moments:

**Theorem 97** *Suppose that  $X$  is a real valued random variable with characteristic function  $\phi$  and  $n$  is a positive integer. If  $\phi$  is  $2n$  times differentiable at 0 then  $\mathbb{E}[X^{2n}] < \infty$ , and  $\phi^{(2n)} = (-1)^n \mathbb{E}[X^{2n} \exp(itX)]$ .*

We now can give the classical proof of the Central Limit Theorem. Suppose that  $\{X_n\}_{n=1}^\infty$  are independent and identically distributed with  $E[X_1] = 0$  and  $E[X_1^2] = 1$ . Let  $\phi$  be the common characteristic function for the  $X_n$ . Then

$$\left(\phi\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

gives the characteristic function of  $n^{-1/2}(X_1 + \cdots + X_n)$  for each real number  $t$ . By Theorem 92 it is sufficient to show that

$$\lim_{n \rightarrow \infty} \left(\phi\left(\frac{t}{\sqrt{n}}\right)\right)^n = \exp(-t^2/2)$$

for each real number  $t$ . From Theorem 94 and Taylor's theorem we know that

$$\phi(u) = 1 + \frac{u^2}{2}\phi''(\theta(u)u)$$

where  $\theta(u) \in (0, 1)$ . Therefore

$$\phi\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + (1 + \phi''((t/\sqrt{n})\theta(t/\sqrt{n})))\frac{t^2}{n}$$

Since  $\phi''$  is continuous at 0, we can write

$$\phi\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n}(1 + \epsilon(n))$$

where  $\epsilon(n) \rightarrow 0$  as  $n \rightarrow +\infty$ . Therefore, when  $|t^2| < 2n$ , by factoring differences of  $n$ th powers,

$$\begin{aligned} \left| \left(1 - \frac{t^2}{2n}\right)^n - \left(\phi\left(\frac{t}{\sqrt{n}}\right)\right)^n \right| &\leq \left| \frac{t^2\epsilon(n)}{n} \right| \sum_{j=0}^{n-1} \left| \phi\left(\frac{t}{\sqrt{n}}\right) \right|^j \\ &\leq |t^2\epsilon(n)| \end{aligned}$$

so

$$\lim_{n \rightarrow \infty} \left(\phi\left(\frac{t}{\sqrt{n}}\right)\right)^n = \exp(-t^2/2)$$

as desired. **QED**

## 5.6 Miscellaneous Facts About Characteristic Functions

We would like necessary and sufficient conditions for a function to be a characteristic function.

**Theorem 98 (S. Bochner, 1932)** *Let  $\phi : \mathbf{R} \rightarrow \mathbf{C}$ . Then  $\phi$  is a characteristic function if and only if*

- $\phi$  is continuous at 0;
- $\phi(0) = 1$ ;

- $\phi$  is positive semi-definite, that is, for all positive integers  $n$ , complex numbers  $a_j$  and real numbers  $t_j$ ,

$$\sum_{j=1}^n \sum_{k=1}^n a_j \bar{a}_k \phi(t_j - t_k) \geq 0.$$

**Proof:** For sufficiency, see Chung pages 179 to 181, Theorem 6.5.2. For necessity, we have already discussed the first two items. For the positive definite condition,

$$\begin{aligned} \sum_{j=1}^n \sum_{k=1}^n a_j \bar{a}_k \phi(t_j - t_k) &= \sum_{j=1}^n \sum_{k=1}^n a_j \bar{a}_k \mathbb{E}[\exp((t_j - t_k)iX)] \\ &= \mathbb{E}\left[\sum_{j=1}^n \sum_{k=1}^n a_j \bar{a}_k \exp((t_j - t_k)iX)\right] \\ &= \mathbb{E}\left[\left|\sum_{j=1}^n a_j \exp(it_j X)\right|^2\right] \end{aligned}$$

and this last quantity is clearly non-negative. **QED**

We also can manufacture new characteristic functions from old ones. One easy, yet important way to do this is to multiply existing characteristic functions together. Theorem 94 tells that if we take infinite products, if the limit exists and is continuous at 0 then it is a characteristic function. The same conclusion can be reached by using Theorem 98. Of course, such products come from the addition of independent random variables.

The second way to form new characteristic functions from old ones is by convex combinations. This corresponds to forming new distributions from mixtures of old ones.

**Proposition 99** Let  $\{a_k\}_{k=1}^{\infty}$  be a sequence of positive numbers whose sum is 1. Let  $\{\phi_k\}_{k=1}^{\infty}$  be a sequence of characteristic functions. Then

$$\sum_{k=1}^{\infty} a_k \phi_k$$

is a characteristic function.

**Proof:** Since  $|\phi_k(t)| \leq 1$  for any  $t$  and  $k$ ,

$$\phi(t) \equiv \sum_{k=1}^{\infty} a_k \phi_k(t)$$

defines a function  $\phi : \mathbf{R} \rightarrow \mathbf{C}$ . This function is the limit of positive semi-definite functions, so it is positive semi-definite. Clearly  $\phi(0) = 1$ . Finally, it is continuous at 0 by applying the dominated convergence theorem. Therefore, it follows from Theorem 98 that  $\phi$  is a characteristic function. **QED**

**Remark:** This proposition can also be proven using Theorem 94 by noting that

$$\phi(t) = \lim_{n \rightarrow \infty} \sum_{k=1}^n \frac{a_k}{a_1 + \cdots + a_n} \phi_k(t)$$

and therefore is the limit of characteristic functions.

Characteristic functions of this type arise as a result of conditioning.

Application: Let  $\phi$  be any characteristic function, and let  $\psi$  be given by

$$\psi(t) = \sum_{k=0}^{\infty} \exp(-m) \frac{m^k}{k!} (\phi(t))^k.$$

Such a characteristic function  $\psi$  arises if we have  $\{X_k\}_{k=1}^{\infty}$  are independent and identically distributed random variables with characteristic function  $\phi$  and  $N$  is a Poisson random variable with mean  $m$ , independent of the  $X_k$ . If we put

$$Y = \sum_{k=1}^N X_k$$

then  $\psi$  is the characteristic function of  $Y$ .

There are more easily verified sufficient conditions for a real valued function to be a characteristic function:

**Theorem 100 (Polya)** *Suppose  $\phi : \mathbf{R} \rightarrow \mathbf{R}$  is even, convex on  $(0, \infty)$ , is continuous at 0 and has  $\phi(0) = 1$  and  $\phi(+\infty) = 0$ . Then  $\phi$  is a characteristic function.*

For  $a > 0$ , the distribution function

$$\int_{-\infty}^x \frac{1 - \cos(at)}{\pi at^2} dt$$

has characteristic function  $\phi_a$  given by

$$\phi_a(t) = \max\{1 - \frac{|t|}{a}, 0\}$$

Since  $\phi$  converges to 0 at infinity it can be uniformly approximated by convex combinations of the  $\phi_a$ . Hence it can be represented a limit of a sequence of characteristic functions, so it, too is a characteristic function as a consequence of Theorem 94. **QED**

For example,  $\phi(t) = \exp(-|t|^a)$  is a characteristic function if  $a \in (0, 1]$ .

## 5.7 Self-Similar Random Variable; Stable Distributions

We know that  $\exp(-|t|)$  is the characteristic function of the Cauchy distribution. Let  $S_n$  be the sum of  $n$  independent Cauchy random variables. It is easy to see that  $n^{-1}S_n$  also has  $\exp(-|t|)$  as its characteristic function, so that it, too, has the Cauchy distribution.

**Definition 101 (Self-Similar Random Variables)** *Let  $\{X_n\}_{n=1}^{\infty}$  be a sequence of independent and identically distributed random variables on  $(\Omega, \mathcal{F}, \text{Pr})$ . We say that  $X_1$  is **self-similar of index**  $a$  if and only if  $n^{-1/a}(X_1 + \dots + X_n)$  has the same distribution as  $X_1$  for every positive integer  $n$ .*

For example, standard normal random variables are self-similar of index 2. Note also that it possible using calculus alone to verify this and that Cauchy random variables are self-similar. Thus the following theorem allows us to compute their characteristic functions without resorting to direct calculation.

**Theorem 102** *If  $X$  is symmetric and self-similar of index  $a$  then the characteristic function  $\phi$  of  $X$  is given by*

$$\phi(t) = \exp(-c|t|^a)$$

for some  $c \geq 0$ .

**Proof:** Fix  $a > 0$ , and let  $\phi$  be the characteristic function in question. Self-similarity tells us that

$$\left(\phi(tn^{-1/a})\right)^n = \phi(t)$$

for all non-negative integers  $n$  and all real numbers  $t$ . Since  $X$  is symmetric we know that  $\phi$  is real valued, and by taking  $n = 2$  we see that  $\phi$  is non-negative.  $\phi(t) \neq 0$  for any  $t$ , since if  $\phi(t_0) = 0$  then  $\phi(t_0/n^{1/a}) = 0$  for every  $n$ , and by continuity,  $\phi(0) = 0$ , a contradiction. Now, let  $u = t^a$ , and let  $\psi(u) = \phi(u^{1/a})$ . Then we have  $\psi(u) = (\psi(u/n))^n$  for all positive integers  $n$  and all real numbers  $u$ . Hence  $\psi(n) = (\psi(1))^n$  for all positive integers  $n$ . This gives us  $\psi(r) = (\psi(1))^r$  for all positive rational numbers  $r$ . By continuity,  $\psi(u) = \exp(u \log(\psi(1)))$  for all non-negative real numbers. Therefore,  $\phi(t) = \exp(-c|t|^a)$  where  $c = -\log(\psi(1))$ . **QED**

To determine that  $c = 1$  for the Cauchy distribution, one may employ the inversion formula derived in Theorem 90. One may also argue directly using the

**Fourier Integral Formula:** *Let  $f : (-\infty, \infty) \rightarrow (-\infty, \infty)$  be piecewise smooth and absolutely integrable. Then*

$$\frac{f(x^+) + f(x^-)}{2} = \frac{1}{\pi} \int_0^\infty \int_{-\infty}^\infty f(y) \cos(s(y-x)) dy ds$$

Set  $f(u) = 1/(1+u^2) \equiv \pi f_C(u)$  and  $x = 0$  to get

$$1 = \int_0^\infty \int_{-\infty}^\infty f_C(y) \cos(sy) dy ds = \int_0^\infty e^{-cs} ds = c.$$

**Theorem 103** *For each  $a \in (0, 2]$ ,  $\phi(t) = \exp(-|t|^a)$  defines a characteristic function.*

**Proof:** This proof is due to Lévy. We already know this assertion to be true when  $a = 2$  as we would have the characteristic function of the normal distribution with mean 0 and variance 2. So let us now consider  $a \in (0, 2)$ . Let  $p$  be the density function given

$$p(x) = \frac{a}{2|x|^{a+1}}$$

for  $|x| \geq 1$  and 0 otherwise. Let  $\psi$  be the characteristic function for this density. Since  $p$  is even, we have

$$\begin{aligned} 1 - \psi(t) &= \int_1^\infty \frac{1 - \cos(tx)}{x^{a+1}} \\ &= a|t|^a \left[ \int_0^\infty \frac{1 - \cos(x)}{x^{a+1}} - \int_0^{|t|} \frac{1 - \cos(x)}{x^{a+1}} \right] \\ &= c_a |t|^a - a|t|^a \int_0^{|t|} \frac{1 - \cos(x)}{x^{a+1}} dx. \end{aligned}$$

where  $c_a > 0$ . From L'Hôpital's rule we see that

$$\lim_{t \rightarrow 0^+} \frac{\int_0^t \frac{1 - \cos(x)}{x^{a+1}} dx}{t^{2-a}} = \frac{1}{4 - 2a}$$

so we may write

$$\psi(t) = 1 - c_a |t|^a + \eta(t) \tag{5.4}$$

where  $\eta(t)/t^2$  converges to a constant as  $t \rightarrow 0$ . On the other hand, for each positive integer  $n$ ,  $\phi_n$  defined by

$$\phi_n(t) = (\psi(t/n^{1/a}))^n$$

is a characteristic function, and it follows from (5.4) that  $\phi_n(t)$  converges to  $\exp(-c_a |t|^a)$  as  $n \rightarrow \infty$ . It then follows from Theorem 92 (due to Lévy too!) that  $\exp(-c_a |t|^a)$  is a characteristic function. Since  $c_a > 0$  we may change scale to get the conclusion of the theorem. **QED**

**Theorem 104** *If  $a > 2$  then  $\exp(-|t|^a)$  is not a characteristic function.*

**Proof:** If  $\phi(t) = \exp(-|t|^a)$  were a characteristic function for  $a > 0$  then  $\phi$  would be twice differentiable at 0. In fact we would have  $\phi'(0) = \phi''(0) = 0$  so the corresponding distribution would have mean and variance equal to 0. This would mean that the distribution would be that of the constant 0 random variable, which has characteristic function constantly equal to 1, a contradiction. **QED**

**Theorem 105** *Let  $X$  be a random variable. Then  $X$  is self-similar and symmetric if and only if it has a characteristic function of the form  $\exp(-c|t|^a)$  for some  $c \geq 0$  and  $a \in (0, 2]$ .*

**Proof:** Combine Theorems 102, 103 and 104. **QED**

Stable distributions arise as the limits of iid sums suitably normalized:

**Theorem 106** *Let  $\{X_k\}_{k=1}^\infty$  be a sequence of independent and identically distributed symmetric random variables on  $(\Omega, \mathcal{F}, \Pr)$ , and that for some  $a \in (0, 2]$  the sequence*

$$\frac{1}{n^{1/a}} \sum_{k=1}^n X_k$$

*converges in distribution to the distribution  $F$ . Then  $F$  is the distribution function of a self-similar of index  $a$  and symmetric.*

**Proof:** By hypothesis, for any positive integer  $k$ , the sequence of random variables

$$\frac{1}{(nk)^{1/a}} \sum_{j=1}^{nk} X_j$$

converges in distribution as  $n \rightarrow \infty$  to the same distribution  $F$ . However, with a little algebra we see

$$\frac{1}{(nk)^{1/a}} \sum_{j=1}^{nk} X_j = \sum_{b=0}^{k-1} \frac{1}{k^{1/a}} \sum_{j=1}^n \frac{1}{n^{1/a}} X_{kn+b}$$

so the limiting distribution is self-similar of index  $a$ , and is clearly symmetric. **QED**

## 5.8 The Central Limit Theorem in Higher Dimensions

**Theorem 107** Let  $\{\vec{X}_k\}_{k=1}^{\infty}$  be an independent and identically distributed sequence of  $d$ -dimensional random variables with mean  $\vec{0}$  and covariance quadratic form  $\Sigma(\vec{\lambda}) = E[(\vec{\lambda} \cdot \vec{X}_k)^2]$ . Then

$$n^{-1/2} \sum_{k=1}^n \vec{X}_k$$

converges in distribution to a multivariate normal random variable  $\vec{X}$  with the same mean and covariance quadratic form.

**Proof:** The characteristic function of  $X$  is

$$\phi(\vec{\lambda}) = E[\exp(i\vec{\lambda} \cdot \vec{X})] = \exp(-\Sigma(\vec{\lambda})/2).$$

So it follows from Theorem 94 that it is sufficient to prove that the characteristic function of

$$n^{-1/2} \sum_{k=1}^n \vec{X}_k$$

converges to  $\exp(-\Sigma(\vec{\lambda})/2)$  as  $n \rightarrow \infty$ . To do so, choose  $\vec{\lambda} \in \mathbf{R}^n$  and let  $Y_j = \vec{\lambda} \cdot \vec{X}_j$ . The  $Y_j$  are independent and identically distributed real valued random variables with mean 0 and variance  $\Sigma(\vec{\lambda})$ . Hence it follows from the version of the Central Limit Theorem already proven that for any real number  $t$ ,

$$\lim_{n \rightarrow \infty} E\left[\exp\left(ian^{-1/2} \sum_{j=1}^n Y_j\right)\right] = \exp(-a^2 \Sigma(\vec{\lambda})/2).$$

Setting  $a = 1$  proves the theorem. **QED**

Note: In a more general setting, we can consider normal random variables taking values in Hilbert or Banach Spaces. Suppose, for example,  $H$  is a Hilbert Space with inner product  $\langle \cdot | \cdot \rangle$ . Then if  $X : (\Omega, \mathcal{F}, \Pr) \rightarrow H$  is a random variable, we say that  $X$  is normal if and only if  $\langle X | h \rangle$  is normally distributed for every  $h \in H$ . If we have a Banach Space  $B$  instead, we simply require that  $b^*(X)$  is normal for every  $b^* \in B^*$ .

## 5.9 An application:

Suppose that  $\{X_j\}_{j=1}^{\infty}$  are independent and identically distributed random variables taking values in  $\mathbf{R}^2$ . Specifically, suppose these random variables are uniformly distributed on  $\{(1, 0), (-1, 0), (0, 1), (0, -1)\}$ . Then

$$\Sigma = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}.$$

Put  $S_n = X_1 + \cdots + X_n$ . We have shown that  $E[||S_n||^2] = n$ . We shall now show that

$$\lim_{n \rightarrow \infty} \frac{E[||S_n||]}{\sqrt{n}} = c \in (0, \infty).$$

Applying the Central Limit Theorem in two dimensions we have  $\sqrt{2/n}S_n$  converges in distribution to the bivariate normal distribution with  $\Sigma(\vec{\lambda}) = \lambda_1^2 + \lambda_2^2$ . That means the density has the form  $f(x, y) = \exp(-(x^2 + y^2)/2)/(2\pi)$ . By computing in polar coordinates we expect that

$$\mathbb{E}[\sqrt{2/n}\|S_n\|] \approx \int_0^{2\pi} \int_0^\infty r^2 \exp(-r^2/2)/(2\pi) dr d\theta = \sqrt{\pi/2}.$$

To make this a convincing argument we need to know if it is true that

$$\lim_{n \rightarrow \infty} \int_{\mathbf{R}} x^p dF_n(x) = \int_{\mathbf{R}} x^p dF(x)$$

if there are finite  $p^{\text{th}}$  moments. Unfortunately, no. But it is true if for some  $d > 0$  the  $p + d$  moments are uniformly bounded. This would justify our proof by taking  $d = 1$ .

**Theorem 108** *Let  $\{F_n\}_{n=1}^\infty$  be a sequence of distribution functions which converges weakly to the distribution function  $F$ . Let  $X_n$  be a random variable with distribution function  $F_n$  and let  $X$  be a random variable with distribution function  $F$ .*

*Let  $g : [0, \infty) \rightarrow [0, \infty)$  be a non-decreasing function such that  $g(x) \rightarrow \infty$  as  $x \rightarrow \infty$ , and let  $h : \mathbf{R} \rightarrow \mathbf{R}$  be a continuous function. If for some  $M > 0$  we have  $\mathbb{E}[g(|X|)|h(X)] < M$  and  $\mathbb{E}[g(|X_n|)|h(X_n)] < M$  for all positive integers  $n$ , then*

$$\lim_{n \rightarrow \infty} \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)].$$

**Proof:** Fix  $A > 0$  such that  $g(A) > 0$ . We have

$$\begin{aligned} |\mathbb{E}[h(X_n)] - \mathbb{E}[h(X)]| &\leq |\mathbb{E}[h(X_n); |X_n| \leq A] - \mathbb{E}[h(X); |X_n| \leq A]| \\ &\quad + |\mathbb{E}[h(X_n); |X_n| > A] - \mathbb{E}[h(X); |X| > A]| \\ &\leq |\mathbb{E}[h(X_n); |X_n| \leq A] - \mathbb{E}[h(X); |X_n| \leq A]| \\ &\quad + \frac{1}{g(A)} (\mathbb{E}[g(A)|h(X_n); |X_n| > A] + \mathbb{E}[g(A)|h(X); |X| > A]) \\ &\leq |\mathbb{E}[h(X_n); |X_n| \leq A] - \mathbb{E}[h(X); |X_n| \leq A]| \\ &\quad + \frac{1}{g(A)} (\mathbb{E}[g(|X_n|)|h(X_n); |X_n| > A] + \mathbb{E}[g(|X|)|h(X); |X| > A]) \\ &\leq |\mathbb{E}[h(X_n); |X_n| \leq A] - \mathbb{E}[h(X); |X_n| \leq A]| + \frac{2M}{g(A)} \end{aligned}$$

By the corollary to Theorem 76 we see that

$$0 \leq \limsup_{n \rightarrow \infty} |\mathbb{E}[h(X_n)] - \mathbb{E}[h(X)]| \leq \frac{2M}{g(A)}.$$

Since  $A$  may be taken to be as large as we like, and  $g(A)$  is divergent, we see that

$$0 \leq \limsup_{n \rightarrow \infty} |\mathbb{E}[h(X_n)] - \mathbb{E}[h(X)]| = 0,$$

which is sufficient to prove our claim. **QED**

Sometimes moments are used to prove limit theorems. For example, suppose that  $\{X_n\}_{n=1}^\infty$  is a sequence of random variables on  $(\Omega, \mathcal{F}, \Pr)$  and  $F_n$  is the distribution function of  $X_n$ . Furthermore, suppose it is known that for each positive integer  $p$  we have

$$\lim_{n \rightarrow \infty} E[X_n^p] = \mu^{(p)} \in (-\infty, \infty).$$

Is this enough to say that the  $F_n$  converge weakly to a distribution function  $F$  with the  $\mu^{(p)}$  the moments of  $F$ ? Unfortunately, no. See, for example, the discussion of the Helly-Bray Theorem in Loève. The answer, however, is affirmative if we know that there is a unique distribution function with the  $\mu^{(p)}$  as its moments. A sufficient condition for this is that the formal power series

$$1 + \sum_{p=1}^{\infty} \frac{\mu^{(p)}}{p!} z^p$$

has a positive radius of convergence. See the discussion of the moment problem in Chung for more details.

## Chapter 6

# The General Theory of Conditional Probability

Suppose that  $(\Omega, \mathcal{F}, \Pr)$  is a probability space, and  $\mathcal{G}$  is a subsigma algebra of  $\mathcal{F}$ . If  $X$  is a random variable on  $(\Omega, \mathcal{F}, \Pr)$  with  $E[|X|] < \infty$  then according to the Radon-Nikodym Theorem (see Rudin, *Real and Complex Analysis*) there is a unique random variable  $Y$  such that

- $Y$  is  $\mathcal{G}$  measurable and  $E[|Y|] < \infty$ ;
- For each  $G \in \mathcal{G}$ ,

$$E[YI_G] = E[XI_G].$$

To see why, we define a function  $Q : \mathcal{G} \rightarrow (-\infty, \infty)$  by the rule  $Q(G) = E[XI_G]$ . This function  $Q$  is a signed measure on  $\mathcal{G}$  which is absolutely continuous with respect to  $P$ , so

$$Y = \frac{dQ}{dP}.$$

This random variable  $Y$  is called the **conditional expectation of  $X$  with respect to  $\mathcal{G}$**  and is denoted by  $E[X|\mathcal{G}]$ .

**Theorem 109** *As defined above,  $E[X|\mathcal{G}]$  is well-defined.*

**Proof:** As we said, existence is guaranteed by the Radon-Nikodym Theorem. For uniqueness, if there were two such random variables,  $Y$  and  $Z$ , then  $G = \{Z > Y\} \in \mathcal{G}$  so

$$E[XI_G] = E[ZI_G] \geq E[YI_G] = E[XI_G]$$

so  $E[ZI_G] = E[YI_G]$ . This tells us that  $\Pr(G) = 0$ . Now reverse the roles of  $Z$  and  $Y$ , and we see that  $\Pr(Z \neq Y) = 0$ .

## 6.1 Connections with conditional probability of events

We know that if  $A$  and  $B$  are events in  $(\Omega, \mathcal{F}, \Pr)$  and  $\Pr(B) > 0$  that we have defined the conditional probability of  $A$  given  $B$ , denoted  $\Pr(A|B)$ , to be

$$\Pr(A|B) := \frac{\Pr(A \cap B)}{\Pr(B)}.$$

On the other hand, if we let  $\mathcal{G} = \{\Omega, \emptyset, B, B^c\}$  we can try to compute  $E[I_A|\mathcal{G}]$  by observing that any  $\mathcal{G}$  measurable function must be a linear combination of  $I_B$  and  $I_{B^c}$ :

$$E[I_A|\mathcal{G}] = b_1 I_B + b_2 I_{B^c}.$$

Since  $E[I_A I_G] = E[E[I_A|\mathcal{G}] I_G]$  for each  $G \in \mathcal{G}$  we have

$$\begin{aligned} \Pr(A \cap B) &= b_1 \Pr(B) + b_2 \cdot 0 \\ \Pr(A \cap B^c) &= b_1 \cdot 0 + b_2 \\ \Pr(B^c) & \end{aligned}$$

so

$$b_1 = \Pr(A|B),$$

and if  $\Pr(B) < 1$  we also have

$$b_2 = \Pr(A|B^c)$$

Thus if we agree to say that  $\Pr(A|B) = 0$  if  $\Pr(B) = 0$  we have

$$E[I_A|\{\Omega, \emptyset, B, B^c\}] = \Pr(A|B)I_B + \Pr(A|B^c)I_{B^c},$$

so that  $E[I_A|\{\Omega, \emptyset, B, B^c\}](\omega) = \Pr(A|B)$  if  $\omega \in B$  and  $E[I_A|\{\Omega, \emptyset, B, B^c\}](\omega) = \Pr(A|B^c)$  if  $\omega \in B^c$ .

**Definition 110 (Conditional Probability:)** Suppose that  $\mathcal{G}$  is a subsigma algebra of  $(\Omega, \mathcal{F}, \Pr)$ . Then for each  $F \in \mathcal{F}$ , the **conditional probability of  $F$  given  $\mathcal{G}$** , denoted  $\Pr[F|\mathcal{G}]$ , is the random variable  $E[I_F|\mathcal{G}]$ .

We will see below that in a certain sense,  $\Pr(\cdot|\mathcal{G})$  has all the properties of a measure save that it is not real valued.

**Proposition 111** Suppose that  $\{B_n\}_{n=1}^{\infty}$  is a measurable partition of  $(\Omega, \mathcal{F}, \Pr)$ . Let  $\mathcal{G}$  be the sigma algebra generated by this partition. Then for each  $F \in \mathcal{F}$  we have

$$\Pr[F|\mathcal{G}] = \sum_{n=1}^{\infty} \Pr(F|B_n) I_{B_n}.$$

and for every random variable  $X$  with  $E[|X|] < \infty$ ,

$$E[X|\mathcal{G}] = \sum_{n=1}^{\infty} \frac{E[X I_{B_n}]}{\Pr(B_n)} I_{B_n}.$$

**Proof:** The proof is straightforward. In each case the righthand side satisfies the definition for the lefthand side. **QED**

## 6.2 Connection with vector projection

There is a geometric interpretation of all of this if we add the condition that  $E[X^2] < \infty$ . If  $\mathcal{G}$  is a subsigma algebra of  $(\Omega, \mathcal{F}, \Pr)$  we can consider the two Hilbert Spaces  $H_{\mathcal{G}} := \mathcal{L}^2(\Omega, \mathcal{G}, \Pr)$  and  $H_{\mathcal{F}} := \mathcal{L}^2(\Omega, \mathcal{F}, \Pr)$ . We have  $H_{\mathcal{G}}$  is a closed subspace of  $H_{\mathcal{F}}$ . Therefore for each  $X \in H_{\mathcal{F}}$  there is a unique  $Y \in H_{\mathcal{G}}$  called the projection of  $X$  onto  $H_{\mathcal{G}}$  with the property that  $E[(X - Y)V] = 0$  for all  $V \in H_{\mathcal{G}}$ . If we take  $V = I_G$  this shows that  $Y = E[X|\mathcal{G}]$ . This interpretation can be extended to  $X$  satisfying  $E[|X|] < \infty$  only by truncation and consideration of positive and negative parts.

## 6.3 Connection with conditioning on random variables

**Definition 112** *If  $Y$  is any random variable, the the conditional expectation of  $X$  given  $Y$ , denoted  $E[X|Y]$  is defined to be  $E[X|\sigma(Y)]$  where  $\sigma(Y)$  is the sigma algebra generated by  $Y$ . (Of course, we must have  $E[|X|] < \infty$ .)*

**Theorem 113** *Let  $X$  be a random variable on  $(\Omega, \mathcal{F}, \Pr)$  with finite expectation, and let  $Y$  be another random variable on  $(\Omega, \mathcal{F}, \Pr)$ . Let  $\mu$  be the probability measure on the Borel sets of  $\mathbf{R}$  induced by  $Y$ :*

$$\mu(B) = \Pr(\{\omega : Y(\omega) \in B\}).$$

*Let  $\lambda$  be the signed measure on the Borel subsets of  $\mathbf{R}$  defined by*

$$\lambda(B) = E[X; \{\omega : X(\omega) \in B\}].$$

*Put*

$$\phi = \frac{d\lambda}{d\mu}(Y).$$

$$E[X|Y] = \phi(Y).$$

**Proof:** To prove this theorem we need to show that  $\phi(Y)$  satisfies the two conditions defining conditional probability. It is clear that it is measurable in  $\sigma(Y)$ . We need only show that for each  $G \in \sigma(Y)$ ,

$$E[X; G] = E[\phi(Y); G].$$

From the definition of  $\sigma(Y)$ , if  $G \in \sigma(Y)$  then there is some Borel subset  $B$  of  $\mathbf{R}$  such that  $G = \{\omega : Y(\omega) \in B\}$ . Now we use the change of variables theorem:

$$\begin{aligned} E[\phi(Y); G] &= \int_G \phi(Y) dPr \\ &= \int_{\Omega} \phi(Y) I_G dPr \\ &= \int_{\Omega} \phi(Y) I_B(Y) dPr \\ &= \int_{\mathbf{R}} \phi(u) I_B(u) d\mu \\ &= \lambda(B) \\ &= E[X; G] \end{aligned}$$

which proves the theorem.

**QED** Here

is an example. Let  $\Omega = \{(x, y) \in \mathbf{R}^2 : x \geq 0, y \geq 0, x + y \leq 1\}$ , let  $\mathcal{F}$  be the Borel sets, and let  $\Pr$  be normalized Lebesgue measure. Let  $X((x, y)) = x$  and let  $Y(x, y) = y$ . We want to compute  $E[X|Y]$  by applying the preceding theorem. First, we need the distribution function of  $Y$ :

$$\Pr(Y \leq y) = \begin{cases} 0 & \text{if } y \leq 0 \\ 1 - (1 - y)^2 & \text{if } y \in [0, 1] \\ 1 & \text{if } y \geq 1 \end{cases}$$

Therefore, using the notation of the theorem:

$$d\mu(y) = \begin{cases} 0 \, dy & \text{if } y < 0 \\ 2(1 - y) \, dy & \text{if } y \in [0, 1] \\ 0 \, dy & \text{if } y \geq 1 \end{cases}$$

Next, for any set  $B = (-\infty, b]$  we have

$$\begin{aligned} \lambda(B) &= E[XI_{\{Y \leq b\}}] \\ &= \begin{cases} 0 & \text{if } b \leq 0 \\ \frac{1}{3}(1 - (1 - b)^3) & \text{if } b \in [0, 1] \\ \frac{1}{3} & \text{if } b \geq 1 \end{cases} \end{aligned}$$

so

$$d\lambda(b) = \begin{cases} 0 \, db & \text{if } b < 0 \\ (1 - b)^2 \, db & \text{if } b \in [0, 1] \\ 0 \, db & \text{if } b \geq 1 \end{cases}$$

Therefore

$$\frac{d\lambda}{d\mu}(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{1-y}{2} & \text{if } y \in [0, 1] \\ 0 & \text{if } y \geq 1 \end{cases}$$

A more general example is the following. Suppose that  $f : \mathbf{R}^2 \rightarrow [0, \infty)$  is a density function, and for each  $y$ ,

$$\int_{\mathbf{R}} |x|f(x, y) \, dy < \infty$$

and

$$\int_{\mathbf{R}^2} |x|f(x, y) \, dy \, dx < \infty.$$

Put

$$g(y) = \int_{\mathbf{R}} f(x, y) \, dx$$

and define  $h : \mathbf{R} \rightarrow \mathbf{R}$  by

$$h(y) = \frac{\int_{\mathbf{R}} xf(x, y) \, dx}{\int_{\mathbf{R}} f(x, y) \, dx}.$$

Let  $\mathcal{B}$  be the Borel subsets of  $\mathbf{R}$ . Then  $\mu : \mathcal{B} \rightarrow [0, 1]$  defined by

$$\mu(B) = \int_B g(y) \, dy$$

is a probability measure on  $\mathcal{B}$ , and  $\lambda : \mathcal{B} \rightarrow \mathbf{R}$  defined by

$$\lambda(B) = \int_B h(y) dy$$

defines a real signed measure of bounded variation on  $\mathcal{B}$ . If we define  $\phi : \mathbf{R} \rightarrow \mathbf{R}$  by

$$\phi(y) = \begin{cases} 0 & \text{if } g(y) = 0 \\ \frac{h(y)}{g(y)} & \text{if } g(y) > 0 \end{cases}$$

then for each Borel set  $B$ ,

$$\lambda(B) = \int_B \phi(y)g(y) dy$$

since  $\mu(B) = 0$  implies  $\lambda(B) = 0$ . Therefore  $\phi$  is a version of  $d\lambda/d\mu$ , and by the preceding theorem,  $E[X|Y] = \phi(Y)$  if  $X$  and  $Y$  have the joint density given by  $f$ .  $\phi(t)$  is the ‘‘recipe’’ given for ‘‘ $E[X|Y = t]$ ’’ in elementary textbooks. I will leave it as an exercise to derive by these means the formula for  $E[X|Y]$  when  $X$  and  $Y$  have a joint discrete distribution.

## 6.4 More elementary theorems

One simple theorem is:

**Theorem 114** *Suppose that  $(\Omega, \mathcal{F}, \Pr)$  is a probability space,  $X$  and  $Y$  are random variables on  $(\Omega, \mathcal{F}, \Pr)$  with finite expected values,  $\mathcal{H} \subset \mathcal{G}$  are subsigma algebras of  $\mathcal{F}$ , and  $c$  is a scalar. Then*

1.  $E[cX + Y|\mathcal{G}] = cE[X|\mathcal{G}] + E[Y|\mathcal{G}]$ ;
2.  $E[E[X|\mathcal{G}]|\mathcal{H}] = E[X|\mathcal{H}]$ ;
3.  $E[E[X|\mathcal{H}]|\mathcal{G}] = E[X|\mathcal{H}]$ ;
4. If  $X \geq 0$  then  $E[X|\mathcal{G}] \geq 0$ ;
5. If  $X \leq Y$  then  $E[X|\mathcal{G}] \leq E[Y|\mathcal{G}]$ ;
6.  $|E[X|\mathcal{G}]| \leq E[|X| |\mathcal{G}]$ .
7. If  $X$  is a simple random variable and  $X$  is  $\mathcal{G}$  measurable then  $E[XY|\mathcal{G}] = XE[Y|\mathcal{G}]$ .
8. If  $\sigma(X)$  is independent of  $\mathcal{G}$  then  $E[X|\mathcal{G}] = E[X]$ .

**Proof:** The first three assertions follow by observing that in each case the function on the left satisfies the conditions defining the function on the left. The fourth assertion follows from the observation that if  $X \geq 0$  then  $E[XI_G] \geq 0$  for every  $G \in \mathcal{G}$ , so  $E[X|\mathcal{G}]$  is the Radon-Nikodym derivative of a positive measure with respect to a positive measure, and, therefore, must be positive ae[Pr]. The fifth assertion follows from the fourth and the first by considering  $Y - X$ , and the sixth from the fifth using the double inequality  $-|X| \leq X \leq |X|$ . For the seventh assertion, first consider  $X = I_H$  where  $H \in \mathcal{G}$ . Then  $|XY| \leq |Y|$  so  $|XY|$  has a finite expectation. Now for every  $G \in \mathcal{G}$ ,  $G \cap H \in \mathcal{G}$  so

$$E[YI_HI_G] = E[YI_{H \cap G}] = E[E[Y|\mathcal{G}]I_{H \cap G}] = E[E[Y|\mathcal{G}]I_HI_G]$$

and  $E[Y|\mathcal{G}]I_H$  is  $\mathcal{G}$  measurable. Thus the seventh assertion is true in this elementary case. The general case follows from the first assertion. The last assertion follows by observing that if  $\sigma(X)$  is independent of  $\mathcal{G}$  then for each  $G \in \mathcal{G}$  we have  $X$  and  $I_G$  are independent random variables, so the constant random variable  $E[X]$  meets the defining conditions for  $E[X|\mathcal{G}]$ . **QED**

There are also versions of the monotone and dominated convergence theorems:

**Theorem 115** *Suppose that  $(\Omega, \mathcal{F}, \Pr)$  is a probability space, and  $Y$ ,  $X$  and  $\{X_n\}_{n=1}^{\infty}$  are random variables on  $(\Omega, \mathcal{F}, \Pr)$  with finite expected values. Then for each subsigma algebra  $\mathcal{G} \subset \mathcal{F}$*

1. *If  $X_1 \leq X_2 \leq \dots \leq X$  and  $X_n \rightarrow X$  then  $E[X_n|\mathcal{G}] \rightarrow E[X|\mathcal{G}]$ ;*
2. *If  $X_1 \geq X_2 \geq \dots \geq X$  and  $X_n \rightarrow X$  then  $E[X_n|\mathcal{G}] \rightarrow E[X|\mathcal{G}]$ ;*
3. *If  $|X_n| \leq Y$  for all  $n$  then  $E[X_n|\mathcal{G}] \rightarrow E[X|\mathcal{G}]$ .*

**Proof:** To prove the first assertion, put  $Y_n = X_n - X_1$ . Note that the  $Y_n$  satisfy the hypotheses of the ordinary monotone convergence theorem. Put

$$Z = \lim_{n \rightarrow \infty} E[X_n|\mathcal{G}].$$

That  $Z$  is well-defined follows from Theorem 114.  $Z$  is  $\mathcal{G}$  measurable. Let  $G \in \mathcal{G}$  be given. Then

$$\begin{aligned} E[ZI_G] &= E[(Z - E[X_1|\mathcal{G}])I_G] + E[X_1I_G] \\ &= E[\lim_{n \rightarrow \infty} E[X_n - X_1|\mathcal{G}]I_G] + E[X_1I_G] \\ &= \lim_{n \rightarrow \infty} E[E[X_n - X_1|\mathcal{G}]I_G] + E[X_1I_G] \\ &= \lim_{n \rightarrow \infty} E[(X_n - X_1)I_G] + E[X_1I_G] \\ &= \lim_{n \rightarrow \infty} E[(X_nI_G)] \\ &= E[XI_G], \end{aligned}$$

so  $Z = E[X|\mathcal{G}]$ .

The second assertion follows from the first by multiplying by  $(-1)$  in the appropriate places and using Theorem 114.

For the third assertion, we imitate the usual proof of the dominated convergence theorem. Put

$$Y_n = \sup_{k \geq n} |X - X_k|.$$

The random variables  $Y_n$  are positive, form a sequence that decreases to 0 almost surely and  $Y_1 \leq 2Y$ . From Theorem 114 we have for each positive integer  $n$

$$|E[X_n|\mathcal{G}] - E[X|\mathcal{G}]| \leq E[|X - X_n| |\mathcal{G}] \leq E[Y_n|\mathcal{G}].$$

$E[Y_n|\mathcal{G}] \rightarrow 0$  almost surely as  $n \rightarrow \infty$  from our second assertion, giving us the third assertion. **QED**

**Theorem 116** *Suppose that  $(\Omega, \mathcal{F}, \Pr)$  is a probability space, and  $X$  and  $Y$  are random variables on  $(\Omega, \mathcal{F}, \Pr)$  with  $E[|XY|] < \infty$  and  $E[|X|] < \infty$ . For each subsigma algebra  $\mathcal{G} \subset \mathcal{F}$ , if  $X$  is  $\mathcal{G}$  measurable then  $E[XY|\mathcal{G}] = XE[Y|\mathcal{G}]$ .*

**Proof:** : This is a direct consequence of the preceding theorems. **QED** To prove the analog of Jensen's inequality, we need the following lemma:

**Lemma 117** Suppose that  $(a, b)$  is an interval, and  $\phi : (a, b) \rightarrow (-\infty, \infty)$  is convex. Then

- $\phi$  is continuous;
- For each  $c \in (a, b)$  there is a constant  $m_c$  such that for all  $x \in (a, b)$ ,

$$\phi(x) \geq \phi(c) + m_c(x - c)$$

for all  $x \in (a, b)$ .

- If  $D$  is dense in  $(a, b)$  then

$$\phi(x) = \sup_{c \in D} \phi(c) + m_c(x - c)$$

**Proof:** : The first assertion is found in Rudin, *Real and Complex Analysis*, as is the inequality

$$\frac{\phi(v) - \phi(u)}{v - u} \leq \frac{\phi(w) - \phi(v)}{w - v}$$

whenever  $a < u < v < w < b$ . Thus  $\phi$  has a right and left hand derivative at any point  $c \in (a, b)$ . Let  $m_c$  denote the right hand derivative at  $c$ .  $m_c$  is increasing in  $c$ , and it follows from the definition of Riemann integrals that for  $a < c \leq x < b$ ,

$$\phi(x) - \phi(c) = \int_c^x m_u \, du \geq \int_c^x m_c \, du = m_c(x - c),$$

while if  $a < x \leq c < b$

$$\phi(c) - \phi(x) = \int_x^c m_u \, du \leq \int_x^c m_c \, du = m_c(c - x).$$

Together, these inequalities give us the second assertion.

For the last assertion, for each  $x \in (a, b)$ , put  $\phi_D(x) = \sup_{c \in D} \{\phi(c) + m_c(x - c)\}$ . We have  $\phi(x) \geq \phi_D(x)$  for each  $x \in (a, b)$  and  $\phi(x) = \phi_D(x)$  for  $x \in D$ . Since  $\phi_D : \mathbf{R} \rightarrow \mathbf{R}$  is continuous,  $\phi = \phi_D$  on  $(a, b)$ .

**QED**

Jensen's inequality:

**Theorem 118** Suppose  $\phi$  is a convex function on an open interval  $(a, b)$ , and  $X$  is a random variable on  $(\Omega, \mathcal{F}, \Pr)$  taking values in  $(a, b)$ . If  $E[|X|] < \infty$  and  $E[|\phi(X)|] < \infty$  then  $E[\phi(X)|\mathcal{G}] \geq \phi(E[X|\mathcal{G}])$ .

**Proof:** : Since  $a < X < b$  we have from Theorem 114 that  $a < E[X|\mathcal{G}] < b$  a.s. Therefore  $\phi(E[X|\mathcal{G}])$  is defined up to a set of probability 0. Define it to be 0 otherwise, so that the theorem statement makes sense.

Now, let  $D = (a, b) \cup \mathbf{Q}$ , so that  $D$  is a countable dense subset of  $(a, b)$ . For each  $d \in D$  there is a constant  $m_d$  such that

$$\phi(x) \geq \phi(d) + m_d(x - d).$$

Therefore, for each  $d \in D$  and all  $\omega \in \Omega$ ,

$$\phi(X) \geq \phi(d) + m_d(X - d).$$

Since  $D$  is countable, it follows from Theorem 114 that for some  $\Omega' \in \mathcal{F}$  with  $\Pr(\Omega') = 1$ , for all  $\omega \in \Omega'$  and all  $d \in D$ ,

$$\mathbb{E}[\phi(X)|\mathcal{G}](\omega) \geq \phi(d) + m_d(\mathbb{E}[X|\mathcal{G}](\omega) + d),$$

and  $a < \mathbb{E}[X|\mathcal{G}](\omega) < b$ , so

$$\mathbb{E}[\phi(X)|\mathcal{G}](\omega) \geq \sup_{d \in D} \{\phi(d) + m_d(\mathbb{E}[X|\mathcal{G}](\omega) + d)\} = \phi(\mathbb{E}[X|\mathcal{G}](\omega))$$

as desired.

**QED**

# Chapter 7

## Martingales

Suppose now that we have a probability space  $(\Omega, \mathcal{F}, \Pr)$  and sigma algebras  $\mathcal{F}_n \subset \mathcal{F}$  where  $\mathcal{F}_n \subset \mathcal{F}_{n+1}$  for  $n \in \{0, 1, \dots\} \equiv \mathbf{N}$ . Such an increasing sequence of subsigma algebras is called a *filtration*. In many cases,  $\mathcal{F}_0 = \{\Omega, \emptyset\}$ . Typically, we imagine as sequence of random variables  $\{Y_n, n \in \mathbf{N}\}$  and  $\mathcal{F}_n = \sigma(Y_1, \dots, Y_n)$  for  $n \in \mathbf{N}$ .

A sequence of random variables  $\{X_n, n \in \mathbf{N}\}$  is called a *martingale relative to the filtration*  $\{\mathcal{F}_n, n \in \mathbf{N}\}$  if for each  $n \in \mathbf{N}$ ,  $X_n$  is measurable with respect to  $\mathcal{F}_n$  and  $E[X_{n+1}|\mathcal{F}_n] = X_n$ . Note, then, that this requires that each  $|X_n|$  be integrable.

### 7.1 Two important examples

First, let  $\{Y_k\}_{k=1}^{\infty}$  be a sequence of independent random variables, each with mean 0. Put  $\mathcal{F}_0 = \{\Omega, \emptyset\}$  and  $\mathcal{F}_n = \sigma(Y_1, \dots, Y_n)$  for  $n \in \mathbf{N}$ . Put  $X_0 = 0$  and  $X_n = X_{n-1} + Y_n$  for  $n \in \mathbf{N}$ . Since  $E[X_n|\mathcal{F}_n] = X_n$  and  $E[Y_{n+1}|\mathcal{F}_n] = E[Y_{n+1}] = 0$ ,

$$E[X_{n+1}|\mathcal{F}_n] = E[X_n + Y_{n+1}|\mathcal{F}_n] = E[X_n|\mathcal{F}_n] + E[Y_{n+1}|\mathcal{F}_n] = X_n + 0$$

so  $\{X_n, n \in \mathbf{N}\}$  is a martingale relative to the filtration  $\{\mathcal{F}_n, n \in \mathbf{N}\}$ . In particular, the partial sums of independent and identically distributed random variables give martingales relative to the obvious filtration.

Second example: Same set-up, but now assume that for any real number  $t$  we have  $M_n(t) := E[\exp(tY_n)] < \infty$ . Put  $X_0 \equiv 1$ ,  $b_0 = 0$ , and  $X_n = \exp(Y_1 + \dots + Y_n - b_n)$  for  $n = 1, 2, \dots$ . Then

$$\begin{aligned} E[X_{n+1}|\mathcal{F}_n] &= E[X_n \exp(Y_{n+1} + b_n - b_{n+1})] \\ &= X_n E[\exp(Y_{n+1} + b_n - b_{n+1})] \\ &= X_n M_{n+1}(1) \exp(b_n - b_{n+1}). \end{aligned}$$

We will have a martingale if for every  $n \in \{0, 1, 2, \dots\}$  we have  $b_{n+1} = b_n + \log(M_{n+1}(1))$ .

## 7.2 Submartingales and Supermartingales

If the  $X_n$  satisfy the condition  $E[X_{n+1}|\mathcal{F}_n] \leq X_n$  for all  $n$  instead of the condition  $E[X_{n+1}|\mathcal{F}_n] = X_n$  for all  $n$  we call the resulting process a **supermartingale**.

If the  $X_n$  satisfy the condition  $E[X_{n+1}|\mathcal{F}_n] \geq X_n$  for all  $n$  instead of the condition  $E[X_{n+1}|\mathcal{F}_n] = X_n$  for all  $n$  we call the resulting process a **submartingale**.

Sub- and supermartingales are intimately related to sub- and superharmonic functions.

**Theorem 119** *Suppose that  $\phi : (-\infty, \infty) \rightarrow (-\infty, \infty)$ .*

*If  $\phi$  is convex and  $\{X_n\}$  is a martingale relative to the filtration  $\{\mathcal{F}_n\}$  then  $\{\phi(X_n)\}$  is a submartingale relative to the same filtration provided the required conditional expectations exist.*

*If  $\phi$  is convex and non-decreasing, and  $\{X_n\}$  is a submartingale relative to the filtration  $\{\mathcal{F}_n\}$  then  $\{\phi(X_n)\}$  is a submartingale relative to the same filtration provided the required conditional expectations exist.*

**Proof:** We apply Jensen's inequality.

$$\begin{aligned} E[\phi(X_{n+1})|\mathcal{F}_n] &\geq \phi(E[X_{n+1}|\mathcal{F}_n]) \\ &\geq \phi(X_n). \end{aligned}$$

At the last step we have equality if  $\{X_n\}$  is a martingale. If we only have a submartingale we need the additional condition that  $\phi$  is non-decreasing. **QED**

Three common choices of  $\phi$  are  $\phi(x) = \max(x, 0)$ ,  $\phi(x) = |x|$  and  $\phi(x) = x^2$ . For this last choice we need our martingales to be square-integrable.