

**NOTES ON PROBABILITY THEORY FOR ENEE 620**

**Adrian Papamarcou**

(with references to “Probability and Measure” by Patrick Billingsley)

Revised February 2006

## 1. Introduction to random processes

A *random process* is a collection of random variables (r.v.'s for short) that arise in the same probability experiment (the last clause can be replaced by the exact statement “that are defined on a common probability space;” the term *probability space* will be defined in the next section). Thus a random process is mathematically represented by the collection

$$\{X_t, t \in I\} ,$$

where  $X_t$  denotes the  $t^{\text{th}}$  random variable in the process, and the index  $t$  runs over an *index set*  $I$  which is arbitrary.

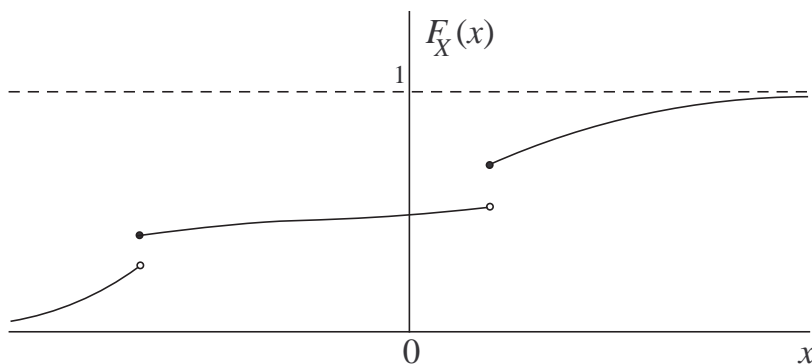
A random process is a mathematical idealization of a set of random measurements obtained in a physical experiment. This randomness can be quantified by a probabilistic or statistical description of the process, and the complexity of this description depends largely on the size of the index set  $I$ . In briefly discussing this issue of complexity, we consider index set sizes, which cover most cases of interest.

(a)  $I$  consists of one index only. In this case we are measuring a single random quantity, represented by the r.v.  $X$ . From elementary probability, we know that a simple way of describing  $X$  statistically is through its *cumulative distribution function* (or *cdf*)  $F_X$ , which is defined by the relationship

$$F_X(x) = \Pr\{X \leq x\}$$

(**Notation.** Throughout this course, random variables will be denoted by upper case letters, and fixed (non-random) numbers by lower case letters.)

$F_X$  is always nondecreasing, continuous from the right, and such that  $F_X(-\infty) = 0$ ,  $F_X(+\infty) = 1$ . Thus typically it looks like this:



We also know that in most cases of interest, we can alternatively specify the statistics of  $X$  by a *probability density function* (or *pdf*)  $f_X$ , which is a nonnegative function that integrates to unity over the entire real line and is related to  $F_X$  by

$$F_X(x) = \int_{-\infty}^x f_X(u) du .$$

This also covers r.v.'s with discrete components, in which case  $f_X$  contains  $\delta$ -functions.

(b)  $I$  consists of  $n$  indices, e.g.  $I = \{1, \dots, n\}$ . In this case the process variables form a random vector in  $\mathbf{R}^n$ , denoted by

$$\mathbf{X} = (X_1, \dots, X_n) .$$

The statistical description of the process can be accomplished by specifying the cdf  $F_{\mathbf{X}}$  (or  $F_{X_1, \dots, X_n}$ ) of the random vector  $\mathbf{X}$ ; this is a real-valued function on  $\mathbf{R}^n$  defined by the relationship

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \Pr\{X_1 \leq x_1, \dots, X_n \leq x_n\} .$$

The dimension of their argument notwithstanding, the functions  $F_{\mathbf{X}}$  and  $F_X$  have quite similar behavior.  $F_{\mathbf{X}}$ , too, is nondecreasing: if  $y_k \geq x_k$  for all values of  $k$ , then

$$F_{\mathbf{X}}(y_1, \dots, y_n) \geq F_{\mathbf{X}}(x_1, \dots, x_n) .$$

Furthermore, in most cases of interest, we can write

$$F_{\mathbf{X}}(x_1, \dots, x_n) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} f_{\mathbf{X}}(u_1, \dots, u_n) du_1 \dots du_n ,$$

for a suitable pdf  $f_{\mathbf{X}}$ , also defined on  $\mathbf{R}^n$ .

Recall that the cdf of any sub-vector of  $X_n$  can be easily determined from the cdf  $F_{\mathbf{X}}$  by setting the redundant arguments of  $F_{\mathbf{X}}$  equal to  $+\infty$ . Thus for example, the cdf of the r.v.  $X_1$  is computed via

$$F_{X_1}(x_1) = F_{\mathbf{X}}(x_1, \infty, \dots, \infty) .$$

This procedure is not reversible: knowledge of the marginal distributions of  $\mathbf{X}$  does not in general suffice to determine  $F_{\mathbf{X}}$ . One important exception is the case where the components of the random vector are *independent*; then the cdf of  $\mathbf{X}$  is given by the product of the cdf's of the individual components, i.e.,

$$F_{\mathbf{X}}(x_1, \dots, x_n) = F_{X_1}(x_1) \dots F_{X_n}(x_n) ,$$

and the same relationship is true if we replace cdf's by pdf's ( $F$  by  $f$ ).

In the last two cases, we consider infinite index sets  $I$ .

(c)  $I$  is countably infinite, say  $I = \mathbf{N}$  (the set of positive integers or natural numbers). Here the process is equivalent to a *sequence* of random variables

$$X_1, X_2, \dots .$$

The problem of describing the statistics of infinitely many random variables is most economically solved by specifying the so-called *finite-dimensional distributions*, namely the

distributions of all finite-dimensional vectors that can be formed with these variables. In this case, the stated procedure amounts to specifying the cdf

$$F_{X_{t_1}, \dots, X_{t_n}}$$

for every choice of  $n$  and (distinct) integers  $t_1, \dots, t_n$ .

Although the above specification of finite-dimensional distributions suffices to describe statistically what happens in the random process over any finite index- (or time-) window, it is not clear whether it also determines properties of the process that effectively involve the entire index set (or discrete-time axis). Consider for example the random variable

$$\bar{X}_\infty = \lim_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n},$$

which gives the asymptotic value of the time average of the random observations.  $\bar{X}_\infty$  is clearly a property of the process, yet its value is not determined by any finite number of variables of the process. Thus

$$\bar{X}_\infty \neq g(X_{t_1}, \dots, X_{t_n})$$

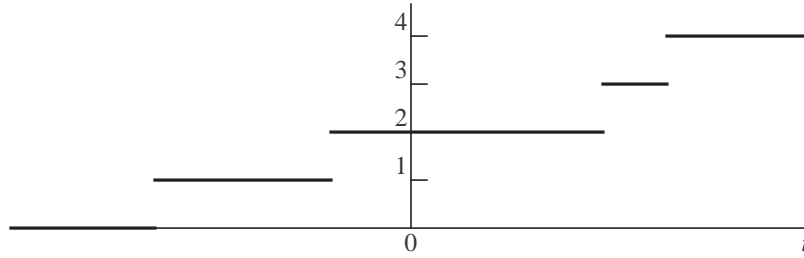
for any choice of  $g$  and arguments  $t_1, \dots, t_n$ , and we cannot use a single finite-dimensional distribution of the process to determine the cdf of  $\bar{X}_\infty$ .

As it turns out (this is a rather profound fact in probability theory), *most* infinitary properties of the process are determined by the set of all finite-dimensional distributions. Such properties include random quantities such as limits of time averages, and thus the statistics of  $\bar{X}_\infty$  are in principle deducible from the finite-dimensional distributions (in practice, the task is usually formidable!). Put differently, if two distinct random processes  $\{X_k, k \in \mathbf{N}\}$  and  $\{Y_k, k \in \mathbf{N}\}$  have identical finite-dimensional distributions, then the variables  $\bar{X}_\infty$  and  $\bar{Y}_\infty$  will also have identical statistics.

In summary, augmentation of a finite index set to a countably infinite one necessitates the specification of an infinite set of finite-dimensional distributions. This entails a considerable jump in complexity, but ensures that all important properties of the process (including asymptotic ones) are statistically specified.

**(d)** In this last case we consider an uncountably infinite index set, namely  $I = \mathbf{R}$ . If we think of the process as evolving in time, then we are effectively dealing with a continuous-time process observed at all times.

In continuing our previous discussion on finite-dimensional distributions, we note another rather profound fact: finite-dimensional distributions *no longer suffice* to determine all salient characteristics of the process. As an example, suppose one wishes to model the number of calls handled by a telephone exchange up to time  $t$ , where  $t \in \mathbf{R}$ . A typical graph of this random time-varying quantity would be



It is now possible to construct two models  $\{X_t, t \in \mathbf{R}\}$  and  $\{Y_t, t \in \mathbf{R}\}$  that have *identical* finite dimensional distributions, yet differ in the following important aspect:  $\{X_t, t \in \mathbf{R}\}$  (almost) always gives observations of the above typical form, whereas  $\{Y_k, k \in \mathbf{Z}\}$  is not known to do the same. More precisely,

$$\Pr\{X_t \text{ is integer-valued and nondecreasing for all } t\}$$

equals unity, whereas the quantity

$$\Pr\{Y_t \text{ is integer-valued and nondecreasing for all } t\}$$

cannot be defined, and hence does not exist.\* Of the two processes, only  $\{X_t, t \in \mathbf{R}\}$  is (possibly) suitable for modeling the random physical system in hand.

The reason for the above discrepancy is that the two random processes  $\{X_t, t \in \mathbf{R}\}$  and  $\{Y_t, t \in \mathbf{R}\}$  are constructed in entirely different ways. This illustrates the general principle that random processes are not fully characterized by distributions alone; their construction amounts to the specification of a family of random variables on the same probability space. Precise understanding of the concepts *probability space* and *random variable* is therefore essential.

## 2. A simple stochastic process

Billingsley, Sec. 1, *The unit interval*.

Consider the probability experiment in which we choose a point  $\omega$  at random from the unit interval  $(0, 1]$ .

(**Notation.** A parenthesis implies that the endpoint lies *outside* the interval; a square bracket that it lies *inside*.)

We assume that the selection of  $\omega$  is *uniform*, in that

$$\Pr\{\omega \in (a, b]\} = b - a .$$

---

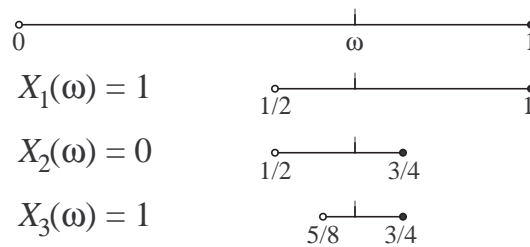
\* The pivotal difference between the statements “ $\bar{X}_\infty \leq 0$ ” in (c) and “ $X_t$  is integer-valued and nondecreasing for all  $t$ ” in (d) is that the former involves a countable infinity of time indices, whereas the latter involves an uncountable one. Agreement of two processes over finite-dimensional distributions implies agreement over “countably expressible” properties, but does not guarantee the same for “uncountably expressible” ones.

As expected,  $\Pr\{\omega \in (0, 1]\} = 1$ .

We now consider the binary expansion of the random point  $\omega$ . We can write

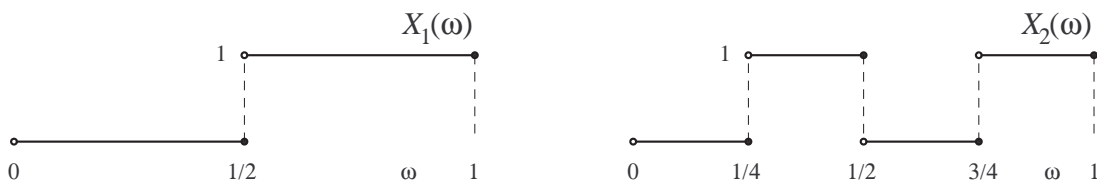
$$\omega = .X_1(\omega)X_2(\omega)\dots = \sum_{k=1}^{\infty} \frac{X_k(\omega)}{2^k},$$

where  $X_k(\omega)$  stands for the  $k^{\text{th}}$  digit in the binary expansion. An iterative algorithm for deriving these digits is as follows. We divide the unit interval into two equal subintervals, and set the first digit equal to 0 if  $\omega$  falls in the left-hand subinterval, 1 otherwise. On the subinterval containing  $\omega$  we perform a similar division to obtain the second digit; and so forth. This is illustrated in the figure below.



**(Notation.** Endpoints marked “o” lie outside, those marked “•” inside, the set or curve depicted)

The variation of the  $k^{\text{th}}$  digit  $X_k(\omega)$  with  $\omega$  has the following distinctive feature. Starting from the left,  $X_k$  alternates in value between 0 and 1 on adjacent intervals of length  $2^{-k}$ : Thus the graphs of  $X_1$  and  $X_2$  look like this:



From the above observation we deduce that the vector  $\mathbf{X} = (X_1, \dots, X_n)$  has the following behavior as  $\omega$  varies:

$$\begin{array}{lll} \omega \in (0, 2^{-n}] & : & \mathbf{X}(\omega) = 000\dots000 \\ \omega \in (2^{-n}, 2 \cdot 2^{-n}] & : & \mathbf{X}(\omega) = 000\dots001 \\ \omega \in (2 \cdot 2^{-n}, 3 \cdot 2^{-n}] & : & \mathbf{X}(\omega) = 000\dots010 \\ & \vdots & \vdots \\ \omega \in ((2^n - 1)2^{-n}, 1] & : & \mathbf{X}(\omega) = 111\dots111 \end{array}$$

Thus each of the  $2^n$  binary words of length  $n$  is obtained over an interval of length  $2^{-n}$ . In terms of probabilities (here length=probability), all binary words of length  $n$  are equally likely candidates for the truncated expansion of a point drawn uniformly at random from the unit interval. Noting also that any fixed digit  $X_k$  is 0 or 1 with equal probability, we conclude that for a binary word  $(a_1, \dots, a_n)$ ,

$$\Pr\{X_1 = a_1, \dots, X_n = a_n\} = 2^{-n} = \Pr\{X_1 = a_1\} \cdots \Pr\{X_n = a_n\} .$$

Now compare the above with the situation in which  $Y_1, Y_2, \dots$  are the outcomes of a sequence of independent tosses of a fair coin labeled 0 and 1. By independence, we have

$$\Pr\{Y_1 = a_1, \dots, Y_n = a_n\} = 2^{-n} = \Pr\{Y_1 = a_1\} \cdots \Pr\{Y_n = a_n\} .$$

Thus we have two sequences of random quantities with identical probabilistic descriptions (it is easy to verify that the processes  $\{X_k, k \in \mathbf{N}\}$  and  $\{Y_k, k \in \mathbf{N}\}$  have the same finite-dimensional distributions). Since drawing a point from an interval is in a sense simpler than tossing a coin infinitely many times, we can use the process  $\{X_k, k \in \mathbf{N}\}$  instead of  $\{Y_k, k \in \mathbf{N}\}$  to model the outcomes of independent coin tosses. This choice has the interesting implication that one can construct infinitely many random quantities without explicit reference to their (joint) statistical description by defining these quantities as *functions* of the outcome of a single probability experiment.

The above leads to the following interpretation, which will prevail in this course: a random variable is a real-valued *function* of the outcome of a probability experiment. A random process is a collection of such functions, all of which are defined in terms of the *same* probability experiment.

### 3. The notion of a probability space

For references, see subsequent sections.

A **probability space** is a mathematical model for a random experiment (or probability experiment). It consists of three entities.

(i) An abstract set of points, called **sample space**, and usually denoted by  $\Omega$ . The points, or elements, of  $\Omega$  are usually denoted by  $\omega$ .

**Interpretation:**  $\Omega$  is the set of all possible outcomes of the random experiment.

(ii) A collection of subsets of  $\Omega$ , called **event space**, and usually denoted by an upper case script character such as  $\mathcal{F}$ . The sets that constitute the event space are called **events**.

**Interpretation:** The event space essentially represents all possible modes of *observing* the experiment. A subset  $A$  of  $\Omega$  is an event if we can set up an observation mechanism to detect whether the outcome  $\omega$  of the experiment lies in  $A$  or not, i.e., whether  $A$  *occurs* or not.

(iii) A function  $P$ , called **probability measure**, which is defined on the event space and takes values in the interval  $[0, 1]$ .

**Interpretation:** For every event  $A$ ,  $P(A)$  provides a numerical assessment of the likelihood that  $A$  occurs; the quantity  $P(A)$  is the **probability** of  $A$ .

The standard representation of a probability space is a triple with the above three entities in their respective order, i.e.,

$$(\Omega, \mathcal{F}, P) .$$

The pair  $(\Omega, \mathcal{F})$  is referred to as a **measurable space**. It describes the outcomes and modes of observation of the experiment without reference to the likelihood of the observables. In general, the same measurable space can give rise to many different probability spaces.

#### 4. Event spaces and fields

Billingsley, Sec. 2, *Spaces and Classes of Sets*.

From a mathematical viewpoint, the sample space  $\Omega$  is entirely unconstrained; it is an arbitrary set of points. Constraints on  $\Omega$  are imposed only by modeling considerations:  $\Omega$  should be “rich” enough to represent all outcomes of the physical experiment that we wish to model. This does not mean that a point  $\omega$  should be of the same form as the outcome of the physical experiment; it merely suggests that one should be able to set up a correspondence between the actual outcomes and the points in  $\Omega$ . Thus in the Example of Sec. 2 above, the sample space  $\Omega = (0, 1]$  adequately represented the outcomes of a sequence of coin tosses, in spite of the fact that the points in  $\Omega$  were not themselves binary sequences. This was because it was possible to identify every  $\omega$  with a distinct binary sequence by taking its binary expansion (conversely, every binary sequence that does not converge to 0 can be identified with a distinct point in  $(0, 1]$ ).

In contrast to the above, the mathematical constraints on the event space  $\mathcal{F}$  are rigid; they stem from the earlier interpretation of events as sets of outcomes that are observable by available mechanisms. Three such constraints are given below.

1.  $\emptyset \in \mathcal{F}, \Omega \in \mathcal{F}$ .

This is reasonable in view of the fact that no observation is needed to determine whether the outcome lies in  $\emptyset$  (impossible) or  $\Omega$  (certain).

2.  $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$  (*closure under complementation*)

Obvious, since the same observation mechanism is used for both  $A$  and  $A^c$ .

3.  $A \in \mathcal{F}, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$  (*closure under union*)

By combining the two observation mechanisms ( $A$  versus  $A^c$  and  $B$  versus  $B^c$ ), one obtains a single observation mechanism for  $A \cup B$  versus  $(A \cup B)^c$ .

**Definition.** An **algebra** or **field** is a collection of subsets of  $\Omega$  satisfying conditions (1)–(3) above.

#### Examples of fields.

- (i)  $\Omega$  arbitrary.  $\mathcal{F} = \{\emptyset, \Omega\}$ .

$\mathcal{F}$  is easily seen to satisfy conditions (1)–(3). It is the smallest field that can be built from a sample space  $\Omega$ , and is often referred to as the *trivial* field. Clearly, no useful observations can be made in the experiment represented here.

(ii)  $\Omega$  arbitrary.  $\mathcal{F}$  = **power set** of  $\Omega$  = the collection of all subsets of  $\Omega$ .

Again  $\mathcal{F}$  is easily seen to satisfy conditions (1)–(3): by convention, the empty set is a subset of every set, and set operations on subsets of  $\Omega$  always yield subsets of  $\Omega$ . In the experiment modeled here, every subset of  $\Omega$  can be tested for occurrence; we thus have the exact opposite of example (i).

(**Notation.** The power set of  $\Omega$  is denoted by  $2^\Omega$ .)

(iii) Here  $\Omega$  is again arbitrary, and we consider sets  $C_1, \dots, C_M$  that form a finite **partition** or **decomposition** of  $\Omega$ ; that is,

$$(\forall i, j \text{ s.t. } i \neq j) C_i \cap C_j = \emptyset \quad \text{and} \quad \bigcup_{i=1}^M C_i = \Omega .$$

The sets  $C_i$  are referred to as **cells** or **atoms** of the partition. The definition of  $\mathcal{F}$  is as follows:

$$\mathcal{F} = \left\{ A : A = \bigcup_{i \in I} C_i, I \subset \{1, \dots, M\} \right\} .$$

Thus  $\mathcal{F}$  consists of all unions of sets  $C_i$ ; by convention, we let

$$\bigcup_{i \in \emptyset} C_i = \emptyset .$$

To see whether  $\mathcal{F}$  is a field, we check conditions (1)–(3).

$$(1) \quad \emptyset = \bigcup_{i \in \emptyset} C_i \in \mathcal{F}, \quad \Omega = \bigcup_{i \in \{1, \dots, M\}} C_i \in \mathcal{F};$$

$$(2) \quad A = \bigcup_{i \in I} C_i \Rightarrow A^c = \bigcup_{i \in \{1, \dots, M\} - I} C_i \in \mathcal{F};$$

$$(3) \quad A = \bigcup_{i \in J} C_i, B = \bigcup_{i \in K} C_i \Rightarrow A \cup B = \bigcup_{i \in J \cup K} C_i \in \mathcal{F},$$

and thus  $\mathcal{F}$  is a field.

(iv) Here we take  $\Omega = (0, 1]$ , and we define  $\mathcal{F}$  as the collection consisting of the empty set and all finite unions of semi-open subintervals of  $(0, 1]$ , i.e.,

$$\mathcal{F} = \{\emptyset\} \cup \left\{ A : A = \bigcup_{i=1}^M (a_i, b_i], M < \infty, (a_i, b_i] \subset (0, 1] \right\} .$$

Here condition (1) is easily seen to be satisfied:  $\emptyset$  is explicitly included in  $\mathcal{F}$ , and the choice  $M = 1, a_1 = 0, b_1 = 1$  yields  $\Omega \in \mathcal{F}$ . The same is true of condition (3), since the union of two finite unions of intervals is itself a finite union of intervals.

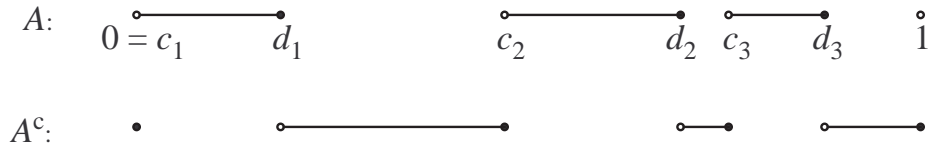
To check condition (2), we first note that if two intervals  $(a_1, b_1]$  and  $(a_2, b_2]$  overlap, their union is a single semi-open interval  $(c, d]$ . Based on this observation, we can use a simple inductive argument to show that a finite union of semi-open intervals can be expressed as a finite union of *non-overlapping* semi-open intervals. In other words,

$$\bigcup_{i=1}^M (a_i, b_i] = \bigcup_{i=1}^N (c_i, d_i],$$

where  $0 \leq c_1 < d_1 < c_2 < d_2 < \dots < c_N < d_N \leq 1$  and  $N \leq M$ . Now

$$\left( \bigcup_{i=1}^N (c_i, d_i] \right)^c = (0, c_1] \cup (d_1, c_2] \cup \dots \cup (d_{N-1}, c_N] \cup (d_N, 1],$$

where both  $(0, 0]$  and  $(1, 1]$  are taken to be the empty set. This equality (illustrated in the figure below) verifies condition (2), thereby proving that  $\mathcal{F}$  is a field.



## Two further properties of fields

**(4) Closure under intersection:**  $A \in \mathcal{F}, B \in \mathcal{F} \Rightarrow A \cap B \in \mathcal{F}$ .

To see this, recall **de Morgan's law**:

$$(A \cap B)^c = A^c \cup B^c.$$

Suppose now that  $A$  and  $B$  lie in  $\mathcal{F}$ . By axioms (2) and (3), the same is true of the sets  $A^c, B^c, A^c \cup B^c$  and  $(A^c \cup B^c)^c$ . The last set is precisely  $A \cap B$ .

**(5) Closure under finite unions:**  $A_1, \dots, A_n \in \mathcal{F} \Rightarrow A_1 \cup \dots \cup A_n \in \mathcal{F}$ .

We prove this by an easy induction: suppose the statement is true for any  $n$  sets  $A_1, \dots, A_n \in \mathcal{F}$ , and that  $A_{n+1}$  is also a set in  $\mathcal{F}$ . Then by axiom (3), we have that

$$A_1 \cup \dots \cup A_{n+1} = (A_1 \cup \dots \cup A_n) \cup A_{n+1}$$

also lies in  $\mathcal{F}$ , which proves that the statement is true for any  $n + 1$  sets in  $\mathcal{F}$ . As the statement is obviously true in the case  $n = 1$ , the induction is complete.

**Remark.** From (4) and (5) it easily follows that every field is closed under finite intersections.

## 5. Event spaces and sigma-fields

Billingsley, Sec. 1., *Classes of Sets*.

### Unions and intersections over arbitrary index sets

Suppose  $\{A_i, i \in I\}$  is a collection of subsets of  $\Omega$ ; here the index set  $I$  is entirely arbitrary.

The **union** of the sets  $A_i$  over  $I$  is defined as the set of points  $\omega$  that lie in *at least one* of the sets in the collection; i.e.,

$$\bigcup_{i \in I} A_i = \{\omega : (\exists i \in I) \omega \in A_i\} .$$

The **intersection** of the sets  $A_i$  over  $I$  is defined as the set of points  $\omega$  that lie in *every one* of the sets in the collection, i.e.,

$$\bigcap_{i \in I} A_i = \{\omega : (\forall i \in I) \omega \in A_i\} .$$

(**Notation.** The symbol  $\forall$  reads “for all,” and  $\exists$  reads “there exists one.”)

### Fields and countable unions

We saw that fields are closed under the operation of taking unions of finitely many constituent sets. However, closure does not always hold if we take unions of infinitely many such sets. Thus if we have a sequence  $A_1, A_2, \dots$  of sets in a field  $\mathcal{F}$ , the union

$$\bigcup_{i=1}^{\infty} A_i \stackrel{\text{def}}{=} \bigcup_{i \in \mathbf{N}} A_i$$

will not always lie in  $\mathcal{F}$ .

To see an instance where such a countable union lies outside the field, consider Example (iv) introduced earlier. If we take

$$A_i = \left(0, \frac{1}{2} - \frac{1}{3i}\right] ,$$

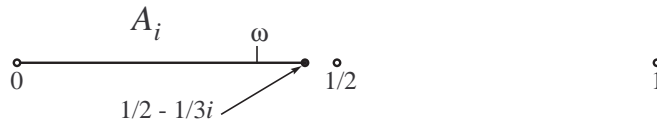
then the countable union

$$\bigcup_{i=1}^{\infty} A_i$$

will be a subset of  $(0, 1/2)$ , since each of the  $A_i$ 's is a subset of that open interval. We claim that this union is actually equal to  $(0, 1/2)$ . Indeed, if  $\omega$  is any point in  $(0, 1/2)$ , then for a sufficiently large value of  $i$  we will have

$$\omega \leq \frac{1}{2} - \frac{1}{3i} ,$$

and thus  $\omega$  will lie in  $A_i$  for that value of  $i$ . This is illustrated in the figure below.



We have therefore shown that the union of all  $A_i$ 's is given by the open interval  $(0, 1/2)$ , which cannot be expressed as a finite union of semi-open intervals and hence lies outside  $\mathcal{F}$ .

**Remark.** An often asked question is: what happens for  $i = \infty$ ? The answer is,  $i$  never takes infinity as a value, and the inclusion of  $\infty$  in the symbol for the above countable union is purely a matter of convention (just as in the case of an infinite series). Thus the definition of the sequence  $A_1, A_2, \dots$  does *not* encompass a set such as  $A_\infty$ , which could be naïvely taken as  $(0, 1/2 - 1/\infty] = (0, 1/2]$ .

### The definition of a sigma-field

A  $\sigma$ -field or  $\sigma$ -algebra is a field that is closed under countable unions. Thus a collection  $\mathcal{F}$  of subsets of  $\Omega$  is a  $\sigma$ -field if it satisfies the following axioms.

1.  $\emptyset \in \mathcal{F}, \Omega \in \mathcal{F}$ .
2.  $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$  (closure under complementation).
- 3'.  $(\forall i \in \mathbf{N}) A_i \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$  (closure under countable unions).

**Remark.** *Countable* means either *finite* or *countably infinite*; a set is countably infinite if its elements can be arranged in the form of an infinite sequence, or equivalently, put in a one-to-one correspondence with the natural numbers. Thus strictly speaking, (3') should be labeled *closure under countably infinite unions*. Yet the distinction is unimportant, since a finite union is a countably infinite union where all but finitely many sets are empty. In particular, (3') readily implies axiom (3) in the definition of a field (closure under union), as well as property (5) of the previous section (closure under finite unions).

The following statement is a direct consequence of the above considerations.

**Corollary.** If a field consists of finitely many sets, it is also a  $\sigma$ -field.

### Examples of sigma fields

Let us again consider the examples of fields given in Section 4.

(i)  $\mathcal{F} = \{\emptyset, \Omega\}$  is a  $\sigma$ -field by the above Corollary.

(ii)  $\mathcal{F} = 2^\Omega$  is a  $\sigma$ -field since any set operation (including taking countable unions) yields a subset of  $\Omega$ .

(iii) In this case  $\mathcal{F}$  consists of all unions of cells in a finite partition of  $\Omega$ . Clearly  $\mathcal{F}$  is finite (if there are  $M$  cells, then there are  $2^M$  sets in  $\mathcal{F}$ ) and thus by the earlier corollary,  $\mathcal{F}$  is a  $\sigma$ -field.

If we take a *countably infinite* partition of  $\Omega$  into sets  $C_1, C_2, \dots$ , then the collection

$$\mathcal{F}' = \left\{ A : A = \bigcup_{i \in I} C_i, I \subset \mathbf{N} \right\},$$

will also be a  $\sigma$ -field. It is easy to check the first two axioms; for closure under countable unions, we note that given any sequence of sets  $A_1, A_2, \dots$  in  $\mathcal{F}'$  such that

$$A_k = \bigcup_{i \in I_k} C_i,$$

we can write

$$\bigcup_{k=1}^{\infty} A_k = \bigcup_{i \in I} C_i,$$

where  $I = I_1 \cup I_2 \cup \dots$ .

**Remark.** A measurable space  $(\Omega, \mathcal{F})$  in which  $\mathcal{F}$  is a  $\sigma$ -field consisting of all unions of atoms in a countable partition of  $\Omega$  is called **discrete**.

(iv). In this example,  $\mathcal{F}$  consisted of the empty set and all finite unions of semi-open subintervals of  $(0, 1]$ . As we saw earlier in this section, there exists sequence of semi-open intervals in  $\mathcal{F}$ , the countable union of which is an open interval lying outside  $\mathcal{F}$ . Thus  $\mathcal{F}$  is not closed under countable unions, and hence it is **not** a  $\sigma$ -field.

## 6. Generated sigma-fields and the Borel field

Billingsley, Sec. 1, *Classes of Sets*.

### The sigma-field generated by a collection

As we saw in the previous section, a field of subsets of  $\Omega$  is not always a  $\sigma$ -field. A question that arises naturally is in what ways such a field (or more generally, an arbitrary collection of subsets of  $\Omega$ ) can be augmented so as to form a  $\sigma$ -field.

It is easy to see that this is always possible, since the power set  $2^\Omega$  is a  $\sigma$ -field which contains (as a subcollection) every collection  $\mathcal{G}$  of subsets of  $\Omega$ . A far more interesting fact is that there also exists a *minimal* such augmentation: that is, given any  $\mathcal{G}$ , there exists a unique  $\sigma$ -field of subsets of  $\Omega$  that both contains  $\mathcal{G}$  and is contained in every  $\sigma$ -field containing  $\mathcal{G}$ .

(**Notation.** The term “contains” can mean either “contains as a subset” or “contains as an element;” which meaning is pertinent depends on the context.)

Before proving the existence of a minimal such  $\sigma$ -field, it is worth giving a simple example. Suppose  $\Omega = (0, 1]$ , and define the collection  $\mathcal{G}$  by

$$\mathcal{G} = \left\{ (0, 1/3], (2/3, 1] \right\}.$$

As pointed out above, the power set

$$\mathcal{F}_1 = 2^\Omega$$

is (trivially) a  $\sigma$ -field that contains  $\mathcal{G}$ . To find the smallest  $\sigma$ -field with this property, we reason as follows.

The sets  $\emptyset$  and  $(0, 1]$  clearly lie in every  $\sigma$ -field containing  $\mathcal{G}$ , as do  $(0, 1/3]$  and  $(2/3, 1]$ . Hence the union

$$(0, 1/3] \cup (2/3, 1]$$

also lies in every such  $\sigma$ -field, and so does its complement  $(1/3, 2/3]$ . By closure under union, the same is true of the sets  $(0, 2/3]$  and  $(1/3, 1]$ . Thus every  $\sigma$ -field containing  $\mathcal{G}$  must also contain the collection

$$\mathcal{F}_2 = \left\{ \emptyset, \Omega, (0, 1/3], (1/3, 2/3], (2/3, 1], (0, 2/3], (1/3, 1], (0, 1/3] \cup (2/3, 1] \right\} .$$

Since  $\mathcal{F}_2$  is itself a  $\sigma$ -field, we conclude that  $\mathcal{F}_2$  is the smallest  $\sigma$ -field containing  $\mathcal{G}$ .

In the case of  $\mathcal{G}$  consisting of infinitely many sets, the construction of a minimal  $\sigma$ -field containing  $\mathcal{G}$  is often impossible. In contrast, the proof of its existence is quite straightforward, and relies on the simple fact that the intersection of an arbitrary class of  $\sigma$ -fields is itself a  $\sigma$ -field.

**(Remark.** In taking the intersection of two collections of subsets of  $\Omega$ , we identify those subsets of  $\Omega$  that are common to both collections; we do **not** take intersections of subsets of  $\Omega$ . For example, if  $\Omega = (0, 1]$  and  $\mathcal{G}, \mathcal{F}_1$  and  $\mathcal{F}_2$  are defined as above, then

$$\mathcal{G} \cap \mathcal{F}_2 = \mathcal{G} \cap \mathcal{F}_1 = \mathcal{G}, \quad \mathcal{F}_2 \cap \mathcal{F}_1 = \mathcal{F}_2 .$$

If also  $\mathcal{F}_3 = \{(0, 1/2]\}$ , then

$$\mathcal{F}_3 \cap \mathcal{G} = \mathcal{F}_3 \cap \mathcal{F}_2 = \emptyset, \quad \mathcal{F}_3 \cap \mathcal{F}_1 = \mathcal{F}_3 .$$

Analogous statements can be made for every set operation and relation applied to collections. Thus for example,

$$\mathcal{G} \subset \mathcal{F}_2 \subset \mathcal{F}_1 ,$$

while  $\mathcal{F}_3$  is a subset of neither  $\mathcal{G}$  nor  $\mathcal{F}_2$ .)

To show that an arbitrary intersection of  $\sigma$ -fields is itself a  $\sigma$ -field, consider

$$\mathcal{F}_\cap = \bigcap_{k \in K} \mathcal{F}_k ,$$

where each  $\mathcal{F}_k$  is a  $\sigma$ -field of subsets of  $\Omega$  and the index set  $K$  is arbitrary. We check each of the three axioms in turn.

(1) Both  $\emptyset$  and  $\Omega$  lie in every  $\mathcal{F}_k$ , thus also in  $\mathcal{F}_\cap$ .

(2) If  $A$  lies in  $\mathcal{F}_k$  for some  $k$ , then  $A^c$  also lies in that  $\mathcal{F}_k$ . If  $A$  lies in every  $\mathcal{F}_k$ , then so does  $A^c$ , i.e.,  $A^c \in \mathcal{F}_\cap$ .

(3') The argument here is essentially the same as above: if  $A_1, A_2, \dots$  lie in every  $\mathcal{F}_k$ , then by closure of each  $\mathcal{F}_k$  under countable unions, the same will be true of

$$\bigcup_{i=1}^{\infty} A_i .$$

With the above fact in mind, we can easily show that the minimal  $\sigma$ -field containing a collection  $\mathcal{G}$  is the intersection of all  $\sigma$ -fields containing  $\mathcal{G}$ . Indeed, the said intersection

- (i) is itself a  $\sigma$ -field (by the above fact);
- (ii) is contained in every  $\sigma$ -field containing  $\mathcal{G}$ ; and
- (iii) contains  $\mathcal{G}$  .

We summarize the above information in the following definition.

Given a collection  $\mathcal{G}$  of subsets of  $\Omega$ , the **minimal**  $\sigma$ -field containing  $\mathcal{G}$ , or equivalently the  $\sigma$ -field **generated** by  $\mathcal{G}$ , is defined as the intersection of all  $\sigma$ -fields containing  $\mathcal{G}$ , and is denoted by  $\sigma(\mathcal{G})$ .

One last remark before proceeding to the next topic is the following: it is possible for two or more distinct collections to generate the same  $\sigma$ -field . A simple illustration of this fact can be given in terms of our earlier example, where

$$\mathcal{G} = \left\{ (0, 1/3], (2/3, 1] \right\} .$$

If we now take

$$\mathcal{G}' = \left\{ (0, 1/3], (1/3, 2/3], (2/3, 1] \right\} \quad \text{and} \quad \mathcal{G}'' = \left\{ (0, 1/3], (0, 2/3] \right\} ,$$

then one can easily verify that

$$\sigma(\mathcal{G}) = \sigma(\mathcal{G}') = \sigma(\mathcal{G}'') .$$

## The Borel field

As noted above, the  $\sigma$ -fields generated by finite collections of subsets of  $\Omega$  are rather easy to construct; such  $\sigma$ -fields are simply described in terms of finite partitions of  $\Omega$  (cf. Section 4, Example (iii)). For a countably infinite  $\mathcal{G}$ , one might be tempted to extrapolate that  $\sigma(\mathcal{G})$  will consist of unions of cells in a countable partition of  $\Omega$ . This will always be true if the sample space  $\Omega$  is countably infinite, but not so if  $\Omega$  is uncountable.

To elaborate on the last statement, we return to the uncountable space

$$\Omega = (0, 1] .$$

As we saw in Section 4, the field

$$\mathcal{F} = \{\emptyset\} \cup \left\{ A : A = \bigcup_{i=1}^M (a_i, b_i], M < \infty, (a_i, b_i] \subset (0, 1] \right\}$$

is not a  $\sigma$ -field. By the foregoing discussion,  $\mathcal{F}$  can be augmented to a minimal  $\sigma$ -field  $\sigma(\mathcal{F})$ . This  $\sigma$ -field is called the **Borel field** of the unit interval, and is denoted by  $\mathcal{B}((0, 1])$ . Thus

$$\mathcal{B}((0, 1]) \stackrel{\text{def}}{=} \sigma(\mathcal{F}) .$$

The Borel field is of crucial importance in probability theory, being the basis for the definition of a random variable and its distribution. It contains, among others, all sets that arise from intervals by countably many set operations. It does not, however, contain every subset of the unit interval; it is possible to give (admittedly contrived) counterexamples to that effect.

What other collections of subsets of  $(0, 1]$  generate  $\mathcal{B}((0, 1])$ ? The answer is many, including some that are easier to describe than  $\mathcal{F}$ . In what follows we give an example of such an alternative collection, principally in order to illustrate a general method of proving that two given collections generate the same  $\sigma$ -field.

We claim that the  $\sigma$ -field generated by the collection

$$\mathcal{G} = \left\{ A : A = (0, a), a < 1 \right\} ,$$

is  $\mathcal{B}((0, 1])$ , i.e., that  $\sigma(\mathcal{G}) = \sigma(\mathcal{F})$ .

To prove the above equality, we must prove each of the inclusions  $\sigma(\mathcal{G}) \subset \sigma(\mathcal{F})$  and  $\sigma(\mathcal{F}) \subset \sigma(\mathcal{G})$ . For the former inclusion, it is sufficient to show that  $\mathcal{G}$  is contained in  $\sigma(\mathcal{F})$ . This is because any  $\sigma$ -field that contains  $\mathcal{G}$  will, by definition of  $\sigma(\cdot)$ , also contain  $\sigma(\mathcal{G})$ . The same argument can be made with  $\mathcal{F}$  and  $\mathcal{G}$  interchanged, and thus we conclude that

$$\mathcal{G} \subset \sigma(\mathcal{F}) \text{ and } \mathcal{F} \subset \sigma(\mathcal{G}) \quad \Rightarrow \quad \sigma(\mathcal{G}) = \sigma(\mathcal{F}) .$$

To prove  $\mathcal{G} \subset \sigma(\mathcal{F})$ : By the method given under *Fields and countable unions* (Section 5), we can write an arbitrary interval  $(0, a)$  in  $\mathcal{G}$  as

$$(0, a) = \bigcup_{i=1}^{\infty} (0, (1 - (i + 1)^{-1})a] .$$

Each of the sets in the above union lies in  $\mathcal{F}$ . Thus  $(0, a)$  is expressible as a countable union of sets in  $\mathcal{F}$ , and hence lies in  $\sigma(\mathcal{F})$ .

To prove  $\mathcal{F} \subset \sigma(\mathcal{G})$ : The empty set trivially lies in  $\sigma(\mathcal{G})$ . It thus remains to prove that every finite union of semi-open intervals  $(a_i, b_i]$  lies in  $\sigma(\mathcal{G})$ ; this is equivalent to proving that every single semi-open interval  $(a, b]$  lies in  $\sigma(\mathcal{G})$ .

We express  $(a, b]$  as

$$(a, b] = (a, 1] - (b, 1] = (a, 1] \cap (b, 1]^c ,$$

so that it suffices to show that every  $(a, 1]$  lies in  $\sigma(\mathcal{G})$ . We now write  $(a, 1]$  as

$$(a, 1] = \bigcup_{i=1}^{\infty} [a + (1 - a)i^{-1}, 1] .$$

Each of the sets in the above union is a complement of a set in  $\mathcal{G}$ , and hence lies in  $\sigma(\mathcal{G})$ . Thus  $(a, 1]$  also lies in  $\sigma(\mathcal{G})$ , and the proof is complete.

We emphasize again that the choice of the alternative generating collection  $\mathcal{G}$  is not unique; one can easily show that substitution of the generic set  $(0, a)$  by any of the intervals  $(0, a]$ ,  $[a, b]$ , etc., still yields the Borel field. More importantly,  $\mathcal{G}$  can be replaced by a (sub-)collection of intervals  $(0, a)$  such that  $a$  is a *rational* number (expressible as a ratio of integers). Since any real number can be written as an increasing or decreasing sequence of rationals, we can easily adapt the above proof to suit the modified  $\mathcal{G}$  by using rational endpoints in the appropriate unions. And since the set of rationals is countable, this implies that the Borel field can be generated by a *countable* collection of intervals.

We can now justify our earlier statement that  $\sigma$ -fields generated by countable collections on uncountable sample spaces are not always described in terms of countable partitions. We do so by noting that the Borel field contains (among others) all sets that consist of single points on the unit interval; these sets alone form an *uncountable* partition of that interval.

The Borel field of the entire real line can be defined in a similar fashion:

$$\mathcal{B}(\mathbf{R}) \stackrel{\text{def}}{=} \sigma \left( \left\{ (-\infty, a] : a \in \mathbf{R} \right\} \right) .$$

Here again the choice of generating intervals is not unique, and rational endpoints are fully acceptable.

We can also define the Borel field of an arbitrary subset  $\Omega$  of the real line by

$$\mathcal{B}(\Omega) \stackrel{\text{def}}{=} \{A : A = C \cap \Omega, C \in \mathcal{B}(\mathbf{R})\} .$$

An interesting exercise is to prove that in the case of the unit interval, the above definition of  $\mathcal{B}((0, 1])$  is consistent with the one given originally.

## 7. Definition of the event space

Billingsley, Sec. 4, *Limit sets*.

As we argued in Section 4, it is desirable that every event space contain  $\emptyset$  and  $\Omega$ , and be closed under complementation and finite unions. Thus every event space should at least be a field. That it should also be a  $\sigma$ -field is not so obvious. The axiom of

closure under countable unions implies the following: if we have a sequence of observation mechanisms  $\mathcal{M}_1, \mathcal{M}_2, \dots$  (where  $\mathcal{M}_i$  observes the occurrence of event  $A_i$ ), then we can effectively combine these mechanisms into one that will decide whether the union

$$\bigcup_{i=1}^{\infty} A_i$$

occurred or not. This implication is not always true in practice, as the example given in this section illustrates. Despite this shortcoming, the structure of the  $\sigma$ -field is chosen for the event space because it allows us to use the powerful mathematical machinery associated with the probability measure (which will be formally defined in the following section).

**For the remainder of this course, the event space  $\mathcal{F}$  will always be a  $\sigma$ -field.**

Consider now the following example which involves an infinite sequence of events  $A_1, A_2, \dots$  in a measurable space  $(\Omega, \mathcal{F})$ . We are interested in descriptions of the set  $A^*$  of sample points  $\omega$  that lie in *infinitely many* (but not necessarily all) of the sets  $A_i$ . Thus

$$A^* = \{\omega : \omega \in A_i \text{ for infinitely many } i\} .$$

In deriving an alternative description of  $A^*$ , we argue as follows. If a point  $\omega$  lies in *finitely* many sets  $A_i$ , then there exists an index  $k$  such that  $\omega$  does not lie in any of the sets  $A_k, A_{k+1}, \dots$ . Conversely, if a point  $\omega$  lies in *infinitely many*  $A_i$ 's, then for *every*  $k$  that point will lie in at least one of the sets  $A_k, A_{k+1}, \dots$ . Thus

$$\begin{aligned} A^* &= \{\omega : (\forall k) \omega \in \text{at least one of } A_k, A_{k+1}, \dots\} \\ &= \left\{ \omega : (\forall k) \omega \in \bigcup_{i \geq k} A_i \right\} . \end{aligned}$$

Writing  $B_k$  for  $\bigcup_{i \geq k} A_i$ , we have

$$A^* = \{\omega : (\forall k) \omega \in B_k\} = \bigcap_{k \geq 1} B_k = \bigcap_{k \geq 1} \bigcup_{i \geq k} A_i .$$

To show that  $A^*$  is an event, i.e.,  $A^* \in \mathcal{F}$ , we argue as follows. Every  $B_k$  is an event, since it can be written as a countable union of events  $A_i$ ; and thus  $A^*$ , which is the intersection of the  $B_k$ 's, is also an event.

**(Remark.** De Morgan's law is true for arbitrary collections of events, and thus closure under countable unions is equivalent to closure under countable intersections.)

The event  $A^*$  is called the **limit superior** (lim sup) of the sequence  $A_1, A_2, \dots$ , and is often described in words as " $A_i$  occurs infinitely often (i.o.)." Thus

$$\limsup_i A_i \stackrel{\text{def}}{=} \{A_i \text{ i.o.}\} \stackrel{\text{def}}{=} \bigcap_{k \geq 1} \bigcup_{i \geq k} A_i .$$

We can think of the above situation in terms of a random experiment whose actual outcome  $\omega$  is unknown to us. Our information about  $\omega$  is limited to a sequence of partial observations of the experiment: for every  $i$  we know whether  $\omega \in A_i$  or  $\omega \in A_i^c$ , i.e., whether the event  $A_i$  has occurred or not. Since the set  $A^*$  of outcomes is expressible in terms of the sequence  $A_1, A_2, \dots$ , it is reasonable to assume that we can process our observations so as to determine whether or not  $\omega \in A^*$  i.e., whether or not infinitely many of the events  $A_i$  have occurred.

Unfortunately, this is easier said than done. Consider for instance the case in which the observations are made sequentially in discrete time. If we assume that the  $A_i$ 's are such that every intersection of the form

$$\bigcap_{i \geq 1} C_i, \quad (C_i \text{ is either } A_i \text{ or } A_i^c),$$

is nonempty, then we have no means of determining in finite time whether infinitely many of the  $A_i$ 's have occurred. Thus the set of outcomes  $A^* = \limsup_i A_i$  does not correspond to any "real" observation of the experiment; it is an event only because  $\mathcal{F}$  is a  $\sigma$ -field.

**Remark.** In a similar manner we can define the **limit inferior** ( $\liminf$ ) of the sequence  $A_1, A_2, \dots$  as the event that " $A_i$  occurs eventually," or equivalently, " $A_i$  occurs for all but finitely many values of  $i$ ." It is easy to check that this definition is consistent with the representation

$$\liminf_i A_i = \bigcup_{k \geq 1} \bigcap_{i \geq k} A_i,$$

and that  $(\liminf_i A_i)^c = \limsup_i A_i^c$ .

## 8. Probability measures

Billingsley, Sec. 2, *Probability Measures*.

### Definition

A **probability measure** on the measurable space  $(\Omega, \mathcal{F})$  is a real-valued function  $P$  defined on  $\mathcal{F}$  that satisfies the following axioms:

(P1) *Nonnegativity:*  $(\forall A \in \mathcal{F}) P(A) \geq 0$ .

(P2) *Normalization:*  $P(\Omega) = 1$ .

(P3) *Countable Additivity:* if  $A_1, A_2, \dots$  are pairwise disjoint events (i.e.,  $A_i \cap A_j = \emptyset$  for  $i \neq j$ ), then

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

Note that in (P3) above, the union of the  $A_i$ 's lies in  $\mathcal{F}$  by the assumption that the event space is a  $\sigma$ -field. (P1–3) are also known as the *Kolmogorov axioms*.

### Simple properties

From (P1) and (P2) we can deduce that  $P$  is finitely additive and that the probability of the empty set is 0. Indeed, if  $A_1, \dots, A_n$  are pairwise disjoint events, we can write

$$\bigcup_{i=1}^n A_i = \bigcup_{i=1}^{\infty} A_i ,$$

where  $A_{n+1} = A_{n+2} = \dots = \emptyset$ . The sequence  $A_1, A_2, \dots$  still consists of disjoint events, so we can apply (P3) to obtain

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

The infinite sum on the right-hand side consists of terms equal to  $P(\emptyset)$ , and hence it will be equal to 0 if  $P(\emptyset) = 0$ , and  $+\infty$  if  $P(\emptyset) > 0$ . As the probability measure cannot take infinity as a value (it is assumed to be a real-valued function), it must be that

$$\sum_{i=n+1}^{\infty} P(A_i) = P(\emptyset) = 0 .$$

Therefore

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

We thus obtain

$$\text{(P4)} \quad P(\emptyset) = 0;$$

and

**(P5)** *Finite additivity*: if  $A_1, \dots, A_n$  are pairwise disjoint events,

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

**(Remark.** We can obtain (P4) without assuming that  $P(A) < \infty$  by applying (P3) to a sequence where  $A_1 = \Omega$  and the remaining  $A_i$ 's equal to  $\emptyset$  ((P1) and (P2) are also needed here). Having obtained (P4), we can apply (P3) and (P1) as before to obtain (P5).)

$$\text{(P6)} \quad \text{For all } A \in \mathcal{F}, P(A) + P(A^c) = 1 .$$

This follows from (P2) and (P5) since  $A \cap A^c = \emptyset$ ,  $A \cup A^c = \Omega$ .

$$\text{(P7)} \quad \textit{Monotonicity under inclusion: } B \supset A \Rightarrow P(B) \geq P(A) .$$

This is because we can write  $B$  as  $A \cup (B \cap A^c)$ , which is a disjoint union. Hence by (P5) and (P1),

$$P(B) = P(A) + P(B \cap A^c) \geq P(A) .$$

In particular,  $\Omega \supset A$  for every event  $A$ , so

**(P8)** For all  $A \in \mathcal{F}$ ,  $P(A) \leq 1$ .

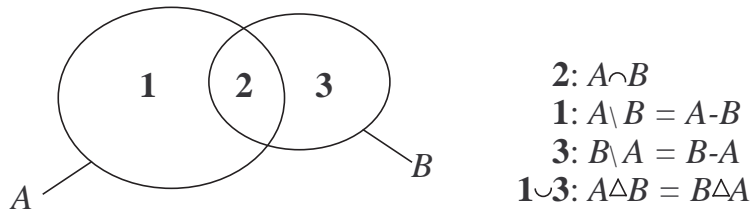
**Notation.** The *set difference*  $A - B$  or  $A \setminus B$  is defined by

$$A - B \stackrel{\text{def}}{=} A \setminus B \stackrel{\text{def}}{=} A \cap B^c .$$

The *symmetric set difference*  $A \triangle B$  is defined by

$$A \triangle B = B \triangle A = (A \cap B^c) \cup (B \cap A^c) .$$

These operations are illustrated in the figure below.



We say that a sequence of events  $(A_n)_{n \in \mathbf{N}}$  is **increasing** if

$$A_1 \subset A_2 \subset \dots ;$$

it is **decreasing** if

$$A_1 \supset A_2 \supset \dots .$$

For such sequences (which are also called *monotone*) we can define a limiting event  $\lim_n A_n$  as follows: if  $(A_n)_{n \in \mathbf{N}}$  is increasing,

$$\lim_n A_n \stackrel{\text{def}}{=} \bigcup_{n=1}^{\infty} A_n ;$$

whereas if  $(A_n)_{n \in \mathbf{N}}$  is decreasing,

$$\lim_n A_n \stackrel{\text{def}}{=} \bigcap_{n=1}^{\infty} A_n .$$

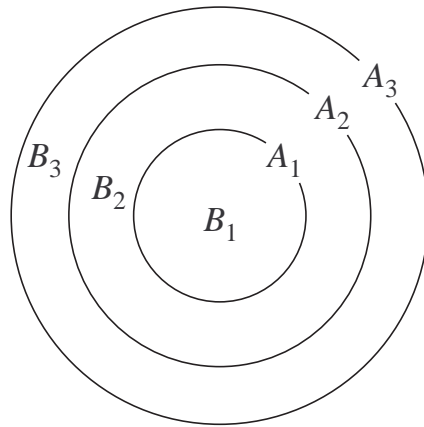
For brevity we write  $A_n \uparrow A$  and  $A_n \downarrow A$  (respectively), where  $A = \lim_n A_n$ .

Probability measures are continuous on monotone sequences of events; in other words,

$$P(\lim_n A_n) = \lim_n P(A_n) .$$

To prove this for an increasing sequence  $A_1, A_2, \dots$ , we generate a sequence of *pairwise disjoint* events  $B_1, B_2, \dots$  as follows:

$$\begin{array}{ll}
 B_1 = A_1 & A_1 = B_1 \\
 B_2 = A_2 - A_1 & A_2 = B_1 \cup B_2 \\
 \vdots & \vdots \\
 B_n = A_n - A_{n-1} & A_n = B_1 \cup \dots \cup B_n
 \end{array}$$



From the above construction, it is easy to see that

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i .$$

This is because any  $\omega$  that lies in one of the  $A_i$ 's will also lie in one of the  $B_i$ 's, and vice versa. Hence

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right) ,$$

and since the  $B_i$ 's are disjoint, countable additivity gives

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(B_i) = \lim_n \sum_{i=1}^n P(B_i) .$$

Now since  $A_n$  is the union of the first  $n$   $B_i$ 's, we also have

$$P(A_n) = \sum_{i=1}^n P(B_i) ,$$

and hence

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_n P(A_n) .$$

The analogous result for decreasing sequences of sets follows easily. If  $A_1, A_2, \dots$  is decreasing, then  $A^c_1, A^c_2, \dots$  is increasing; by De Morgan's law and the above result we then have

$$\begin{aligned} P\left(\bigcap_{i=1}^{\infty} A_i\right) &= P\left(\bigcup_{i=1}^{\infty} A_i^c\right)^c = 1 - \lim_n P(A_n^c) \\ &= 1 - [1 - \lim_n P(A_n)] = \lim_n P(A_n) . \end{aligned}$$

We have thus obtained

**(P9)** *Monotone continuity from below:* If  $A_n \uparrow A$ , then  $\lim_n P(A_n) = P(A)$ .

**(P10)** *Monotone continuity from above:* If  $A_n \downarrow A$ , then  $\lim_n P(A_n) = P(A)$ .

**Remark.** As we have seen, (P9) and (P10) follow directly from the Kolmogorov axioms (P1–3). It is not difficult to show that under the assumption of nonnegativity (P1), the countable additivity axiom (P3) is actually equivalent to the two axioms of finite additivity (P5) and monotone continuity from below (P9) combined. Thus an alternative to the Kolmogorov axioms is the set of axioms consisting of (P 1,2,5) and either (P9) or (P10):

$$\left. \begin{array}{l} \text{Nonnegativity} \\ \text{Normalization} \\ \text{Countable additivity} \end{array} \right\} \iff \left\{ \begin{array}{l} \text{Nonnegativity} \\ \text{Normalization} \\ \text{Finite additivity} \\ \text{Continuity from above or below} \end{array} \right.$$

### Convex mixtures of probability measures

Let  $P_1, P_2, \dots$  be probability measures on the *same* measurable space  $(\Omega, \mathcal{F})$ . We say that the set function  $P$  is a *convex mixture* (or *convex combination*) of these measures if it can be expressed as a weighted sum of the  $P_i$ 's with nonnegative weights that add to unity. In other words, for every  $A \in \mathcal{F}$ ,  $P(A)$  is defined as

$$P(A) = \sum_{i=1}^{\infty} \lambda_i P_i(A) ,$$

where the real coefficients  $\lambda_i$  satisfy

$$(\forall i) \quad \lambda_i \geq 0, \quad \sum_{i=1}^{\infty} \lambda_i = 1 .$$

**Claim.**  $P$  is a probability measure on  $(\Omega, \mathcal{F})$ .

**Proof.** Nonnegativity of  $P$  follows directly from that of the  $\lambda_i$ 's and  $P_i$ 's. Normalization is also easily established via

$$P(\Omega) = \sum_{i=1}^{\infty} \lambda_i P_i(\Omega) = \sum_{i=1}^{\infty} \lambda_i = 1 .$$

To prove countable additivity, consider a sequence of disjoint events  $A_1, A_2, \dots$ . We have

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{i=1}^{\infty} \lambda_i P_i\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{i=1}^{\infty} \lambda_i \sum_{j=1}^{\infty} P_i(A_j) ,$$

where the last equality follows by countable additivity of the measures  $P_i$ . The iterated sum has positive summands, so we can change the order of summation to obtain

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} \lambda_i P_i(A_j) = \sum_{j=1}^{\infty} P(A_j) .$$

Thus  $P$  is a probability measure.

## 9. Specification of probability measures

Billingsley, Sec. 3, *Lebesgue measure on the unit interval*; Sec. 4.

### Discrete spaces

As we saw in Section 5, a measurable space  $(\Omega, \mathcal{F})$  is discrete if the  $\sigma$ -field  $\mathcal{F}$  is generated by a countable partition of  $\Omega$  into atoms

$$C_1, C_2, \dots .$$

Then for every event  $A$  there exists an index set  $I \subset \mathbf{N}$  such that

$$A = \bigcup_{i \in I} C_i ,$$

and if  $P$  is a probability measure on  $(\Omega, \mathcal{F})$ , we have

$$P(A) = \sum_{i \in I} P(C_i) .$$

The above demonstrates that in order to define a probability measure  $P$  on  $(\Omega, \mathcal{F})$ , it suffices to specify the quantity

$$p_i = P(C_i)$$

for every atomic event  $C_i$ . Clearly, the  $p_i$ 's satisfy

$$(\forall i) \quad p_i \geq 0, \quad \sum_{i=1}^{\infty} p_i = 1 .$$

That any sequence  $(p_i)_{i \in \mathbf{N}}$  satisfying this nonnegativity/normalization condition generates a probability measure is not difficult to see: if we let

$$P(A) = \sum_{i \in I} p_i$$

with  $A$  and  $I$  defined as before, then the set function  $P$  is nonnegative and such that  $P(\Omega) = 1$ . To establish countable additivity, we simply note that disjoint events can be expressed as unions of cells over likewise disjoint index sets.

**Definition.** A **probability mass function (pmf)** is a sequence of nonnegative numbers whose sum equals unity. In the context of a given discrete probability space, the pmf is the function that assigns probability to each atomic event.

**Example.**  $\Omega = \{0, 1, \dots\}$ ,  $\mathcal{F} = 2^\Omega$ .

It is easy to see in this case that  $\mathcal{F}$  is generated by the one point sets (or **singletons**)  $\{k\}$ . To define the measure  $P$ , we use the *Poisson* pmf:

$$p_k = e^{-\lambda} \frac{\lambda^k}{k!},$$

where  $k = 0, 1, \dots$ . We have

$$\sum_{k=0}^{\infty} p_k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^\lambda = 1$$

as required. If we let  $P\{k\} = p_k$ , then the probability that the outcome of the experiment is odd is

$$P\{1, 3, 5, \dots\} = e^{-\lambda} \left( \lambda + \frac{\lambda^3}{3!} + \frac{\lambda^5}{5!} + \dots \right) = e^{-\lambda} \left( \frac{e^\lambda - e^{-\lambda}}{2} \right) = \frac{1 - e^{-2\lambda}}{2}.$$

## Non-discrete spaces

As we saw above, we can concisely specify a probability measure on a discrete space by quoting the probabilities of the atomic events. This is not possible in the case of a non-discrete space, since the event space is no longer generated by a countable partition of  $\Omega$  and thus there is no countable family of “minimal” events. Yet a concise specification of measures on non-discrete spaces is absolutely necessary if such spaces are to be used in probability models. Without such specification, the task of explicitly defining a set function on *all* (uncountably many) events and subsequently testing it for the Kolmogorov axioms becomes impracticable.

A method often used for defining a probability measure on a non-discrete  $\sigma$ -field  $\mathcal{F}$  involves the construction of a preliminary set function  $Q$  on a field  $\mathcal{F}_0$  that generates  $\mathcal{F}$ . If the function  $Q$ , as constructed on  $\mathcal{F}_0$ , satisfies certain conditions similar (but not identical) to the Kolmogorov axioms, then it is possible to extend  $Q$  to a *unique* probability

measure  $P$  on  $\mathcal{F} = \sigma(\mathcal{F}_0)$ . Thus under these conditions, specification of  $Q$  on  $\mathcal{F}_0$  suffices to determine the probability measure  $P$  on  $\mathcal{F}$  without ambiguity. (One word of caution: that  $Q$  uniquely determines  $P$  does not imply that for an arbitrary event  $A \notin \mathcal{F}_0$  one can easily compute  $P(A)$  based on the values taken by  $Q$ ; in many cases, this computation will be highly complex or even infeasible.)

The above method of defining measures is based on the following theorem, whose proof can be found in Billingsley, Section 3.

**Theorem.** Let  $\mathcal{F}_0$  be a field of subsets of  $\Omega$ , and let  $Q$  be a nonnegative countably additive set function on  $\mathcal{F}_0$  such that  $Q(\Omega) = 1$ . Then there exists a unique probability measure on  $\sigma(\mathcal{F}_0)$  such that  $P \equiv Q$  on  $\mathcal{F}_0$ .

**Remark.** As  $\mathcal{F}_0$  will not in general be closed under countable unions, the statement “ $Q$  is countably additive on  $\mathcal{F}_0$ ” is understood as “if  $A_1, A_2, \dots$  lie in  $\mathcal{F}_0$  and  $\bigcup_i A_i$  also lies in  $\mathcal{F}_0$ , then

$$Q\left(\bigcup_i A_i\right) = \sum_i P(A_i) .”$$

### The Lebesgue measure on the unit interval

To illustrate the extension technique outlined in the previous subsection, we briefly consider the problem of defining a probability measure  $P$  on the Borel field of the unit interval such that

$$P(a, b] = b - a$$

for every  $(a, b] \subset (0, 1]$ .

We first need to identify a field  $\mathcal{F}_0$  that generates the Borel field, i.e., such that  $\sigma(\mathcal{F}_0) = \mathcal{B}((0, 1])$ . As we saw in Section 6, one such choice of  $\mathcal{F}_0$  consists of the empty set and all finite *disjoint* unions of semi-open intervals

$$\bigcup_{i=1}^N (c_i, d_i] ,$$

where  $0 \leq c_1 < d_1 < c_2 < d_2 < \dots < c_N < d_N \leq 1$ . Since we would like the probability of any semi-open interval to be equal to its length, we *must* define the set function  $Q$  on  $\mathcal{F}_0$  by

$$Q(\emptyset) = 0, \quad \left(\bigcup_{i=1}^N (c_i, d_i]\right) = \sum_{i=1}^N (d_i - c_i) .$$

The set function  $Q$  is clearly nonnegative and satisfies  $Q(0, 1] = 1$ . It is not difficult to show that  $Q$  is finitely additive: if two sets in  $\mathcal{F}_0$  are disjoint, then their constituent intervals are non-overlapping. Countable additivity of  $Q$  on  $\mathcal{F}_0$  can be also established, but the proof is slightly more involved (see Billingsley, Sec. 2, *Lebesgue measure on the unit interval*).

By the extension theorem of the previous subsection, there exists a unique probability measure  $P$  on  $\sigma(\mathcal{F}_0) = \mathcal{B}((0, 1])$  such that  $P = Q$  on  $\mathcal{F}_0$ .  $P$  is called the **Lebesgue**

**measure** on the unit interval. It is the only probability measure on  $\mathcal{B}((0, 1])$  that assigns to every semi-open interval a probability equal to its length.

What is the probability of a singleton  $\{x\}$  under the Lebesgue measure? Intuitively, it should be zero (a point has no length). This is easily verified by writing  $\{x\}$  as the limit of a decreasing sequence of semi-open intervals,

$$\{x\} = \bigcap_{i=1}^{\infty} ((1 - (i + 1)^{-1})x, x] ,$$

and invoking monotone continuity from above:

$$P\{x\} = \lim_n P((1 - (n + 1)^{-1})x, x] = \lim_n (n + 1)^{-1}x = 0 .$$

Thus the Lebesgue measure of any interval (whether open, semi-open or closed) equals the interval length.

**Question.** Do countable subsets of the unit interval (e.g., the set of rationals in that interval) lie in its Borel field? What is the Lebesgue measure of such sets?

## 10. Definition of random variable

Billingsley, Sec. 5, *Definition*; Sec. 13, *Measurable Mappings, Mappings into  $R^k$* .

### Preliminaries

In this section we consider real-valued functions  $f$  defined on a sample space  $\Omega$ , i.e.,

$$f : \Omega \mapsto \mathbf{R} ,$$

We recall the definition of the **image** of a set  $A \subset \Omega$  under  $f$  as

$$f(A) \stackrel{\text{def}}{=} \{x \in \mathbf{R} : (\exists \omega \in A) f(\omega) = x\} ;$$

the **inverse image** of a set  $H \subset \mathbf{R}$  under  $f$  is defined by

$$f^{-1}(H) \stackrel{\text{def}}{=} \{\omega \in \Omega : f(\omega) \in H\} .$$

In developing the concept of the random variable, we will employ images and inverse images (in particular) extensively. The following simple property will be quite useful: if  $\{H_i, i \in I\}$  is an arbitrary collection of subsets of  $\mathbf{R}$ , then

$$f^{-1}\left(\bigcup_{i \in I} H_i\right) = \bigcup_{i \in I} f^{-1}(H_i) .$$

To see this, let  $\omega$  lie in the inverse image of the union. Then  $f(\omega) \in H_i$  for some  $i = i(\omega)$ , and thus  $\omega \in f^{-1}(H_i)$ . Conversely, if  $\omega'$  lies in the union of the inverse images, then  $f(\omega') \in H_j$  for some  $j = j(\omega')$ , and thus  $f(\omega')$  also lies in the union of all  $H$ 's.

By similar reasoning, we can show that

$$f^{-1}\left(\bigcap_{i \in I} H_i\right) = \bigcap_{i \in I} f^{-1}(H_i),$$

and for (forward) images,

$$f\left(\bigcup_{i \in I} A_i\right) = \bigcup_{i \in I} f(A_i), \quad f\left(\bigcap_{i \in I} A_i\right) \subset \bigcap_{i \in I} f(A_i).$$

### Definition

A **random variable** (r.v.) on a measurable space  $(\Omega, \mathcal{F})$  is a real-valued function  $X = X(\cdot)$  on  $\Omega$  such that for all  $a \in \mathbf{R}$ , the set

$$X^{-1}(-\infty, a] = \{\omega : X(\omega) \leq a\}$$

lies in  $\mathcal{F}$  (i.e., is an event).

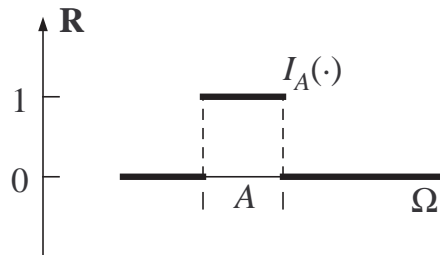
We can think of  $X(\omega)$  as the result of a measurement taken in the course of a random experiment: if the outcome of the experiment is  $\omega$ , we obtain a reading  $X(\omega)$  which provides partial information about  $\omega$ . A more precise interpretation will be given in Section 11. For the moment, we consider some simple examples of random variables.

### Examples of random variables

(i) Let  $(\Omega, \mathcal{F})$  be arbitrary, and  $A \in \mathcal{F}$ . The **indicator function**  $I_A(\cdot)$  of the event  $A$  is defined by

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

We illustrate this definition in the figure below, where for the sake of simplicity, the sample space  $\Omega$  is represented by an interval on the real line.



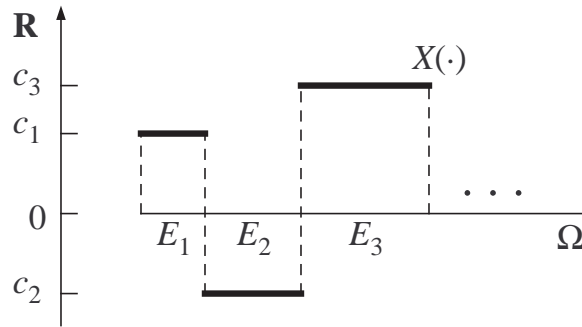
To see that  $X = I_A$  defines a random variable on  $(\Omega, \mathcal{F})$ , note that as  $a$  ranges over the real line, the set  $X^{-1}(-\infty, a] = \{\omega : X(\omega) \leq a\}$  is given by

$$X^{-1} = \begin{cases} \emptyset, & \text{if } a < 0; \\ A^c & \text{if } 0 \leq a < 1; \\ \Omega & \text{if } a \geq 1. \end{cases}$$

Thus  $X^{-1}(-\infty, a]$  is always an event, and hence  $X = I_A$  is a r.v. by the above definition. If  $B \notin \mathcal{F}$ , then also  $B^c \notin \mathcal{F}$ , and thus  $I_B$  is **not** a r.v.

(ii)  $(\Omega, \mathcal{F})$  is arbitrary,  $E_1, E_2, \dots$  is a countable partition of  $\Omega$ , and  $(c_j)_{j \in \mathbf{N}}$  is a sequence of *distinct* real numbers. We define  $X(\cdot)$  as the function that takes the constant value  $c_j$  on every event  $E_j$ . In terms of indicator functions,

$$X(\omega) = \sum_{j=1}^{\infty} c_j I_{E_j}(\omega) .$$



We then have for  $a \in \mathbf{R}$

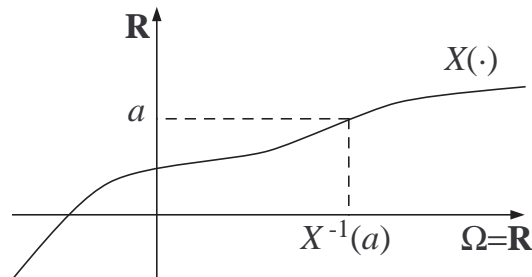
$$X^{-1}(-\infty, a] = \bigcup_{j \in J_a} E_j ,$$

where  $J_a = \{j : a \geq c_j\}$ . As the above inverse image is a countable union of events, it is itself an event and thus  $X$  is a r.v. on  $(\Omega, \mathcal{F})$ .

**Remark.**  $X$  as defined above takes the general form of a **discrete** random variable on the measurable space  $(\Omega, \mathcal{F})$ . Thus a discrete r.v. is one whose range is countable. A **simple** random variable is one whose range is finite.

(iii) Let  $(\Omega, \mathcal{F}) = (\mathbf{R}, \mathcal{B}(\mathbf{R}))$  and  $X(\cdot)$  be a continuous increasing real-valued function such that

$$\lim_{\omega \downarrow -\infty} X(\omega) = -\infty . \quad \lim_{\omega \uparrow +\infty} X(\omega) = +\infty .$$



Then  $X^{-1}(-\infty, a] = (-\infty, X^{-1}(a)]$ . Since every interval lies in  $\mathcal{B}(\mathbf{R})$ ,  $X$  is a r.v.

**Remark.** There are many possible variations on this example. If we merely assume that  $X$  is nondecreasing (as opposed to strictly increasing) the inverse images are again of the form  $(-\infty, b]$  for suitable values of  $b$ . If we remove the constraint that the limits as  $\omega \rightarrow \pm\infty$  be infinite, we admit  $\emptyset$  and  $\mathbf{R}$  as possible inverse images. Finally, if we drop the continuity assumption, then the inverse images will take the form  $(-\infty, b]$  or  $(-\infty, b)$ . In all these cases,  $X$  remains a r.v. on  $(\Omega, \mathcal{F})$ .

## 11. The sigma field generated by a random variable

Billingsley, Sec. 5, *Definition*; Sec. 13, *Measurable Mappings, Mappings into  $R^k$* ; Sec. 20, *Subfields*.

As noted in the previous section, we can think of the random variable  $X$  as a measurement taken in the course of the random experiment  $(\Omega, \mathcal{F}, P)$ . Thus if the outcome of the experiment is  $\omega$ , we can observe  $X(\omega)$  with the aid of a measuring instrument  $\mathcal{M}_X$ .

In introducing the important concept of the  $\sigma$ -field *generated* by  $X$ , we will assume that the instrument  $\mathcal{M}_X$  can be set in an infinity of *observation modes* indexed by the real numbers  $a$  and denoted by  $\mathcal{M}_{X,a}$ . In mode  $\mathcal{M}_{X,a}$ , the instrument determines whether  $X(\omega) \leq a$  or  $X(\omega) > a$ ; that is, it decides which of the complementary events

$$X^{-1}(-\infty, a] = \{\omega : X(\omega) \leq a\} \quad \text{and} \quad X^{-1}(a, \infty) = \{\omega : X(\omega) > a\}$$

occurs. (Note that the above subsets of  $\Omega$  are events by definition of the random variable; this is also consistent with our earlier interpretation of an event as a set of outcomes whose occurrence can be determined.)

The class of events observable by  $\mathcal{M}_X$  is not limited to inverse images of intervals. As in our earlier discussion of the event space, we may assume that this class contains the empty set and the sample space, and is closed under complementation and countable unions, i.e., it is a  $\sigma$ -field. Since the “nominal” events observable by  $\mathcal{M}_X$  are the inverse images of the intervals  $(-\infty, a]$ , it is reasonable to postulate that the  $\sigma$ -field associated with the instrument  $\mathcal{M}_X$  is the smallest  $\sigma$ -field that contains these events.

### Definition.

Let  $X$  be a random variable on  $(\Omega, \mathcal{F}, P)$ . We denote by  $\sigma(X)$  the  $\sigma$ -field generated by the events  $\{\omega : X(\omega) \leq a\}$  as  $a$  varies over the real line. Thus

$$\sigma(X) \stackrel{\text{def}}{=} \sigma\left(\left\{X^{-1}(-\infty, a] : a \in \mathbf{R}\right\}\right).$$

$\sigma(X)$  is referred to as the  **$\sigma$ -field generated by  $X$** .

**Corollary.**  $\sigma(X) \subset \mathcal{F}$ .

To see this, note that the generating collection

$$\mathcal{G} = \left\{X^{-1}(-\infty, a] : a \in \mathbf{R}\right\}$$

is contained in  $\mathcal{F}$ . Thus  $\sigma(X) = \sigma(\mathcal{G}) \subset \mathcal{F}$ .

### Examples

Consider the three examples of the previous section.

(i)  $(\Omega, \mathcal{F})$  arbitrary,  $A \in \mathcal{F}$ ,  $X = I_A$ . We have seen that the inverse image  $X^{-1}(-\infty, a]$  is one of the three sets

$$\emptyset, A^c, \Omega.$$

Thus

$$\sigma(X) = \sigma(\{\emptyset, A^c, \Omega\}) = \{\emptyset, \Omega, A, A^c\},$$

and  $I_A$  allows us to determine the occurrence of a single nontrivial event, namely  $A$ .

(ii)  $(\Omega, \mathcal{F})$  is again arbitrary, and

$$X = \sum_{j=1}^{\infty} c_j I_{E_j},$$

where the  $c_j$ 's are distinct and the events  $E_j$  form a countable partition of  $\Omega$ . In this case the inverse images satisfy

$$X^{-1}(-\infty, a] = \bigcup_{j \in J_a} E_j$$

for  $J_a = \{j : a \geq c_j\}$ . Thus every inverse image lies in the  $\sigma$ -field generated by the partition  $\{E_1, E_2, \dots\}$ , and hence

$$\sigma(X) \subset \sigma(\{E_1, E_2, \dots\}).$$

We claim that the reverse inclusion is also true. To show this, it suffices to prove that every atom  $E_j$  of the partition lies in  $\sigma(X)$ . Noting that (by distinctness of the  $c_j$ 's)  $E_j = X^{-1}\{c_j\}$ , and that

$$\{c_j\} = (-\infty, c_j] - \bigcup_{n=1}^{\infty} (-\infty, c_j - n^{-1}],$$

we obtain

$$E_j = X^{-1}(-\infty, c_j] - \bigcup_{n=1}^{\infty} X^{-1}(-\infty, c_j - n^{-1}].$$

Since the above expression involves countably many operations on inverse images of intervals  $(-\infty, a]$ , we have that  $E_j \in \sigma(X)$ . Hence we conclude that

$$\sigma(X) = \sigma(\{E_1, E_2, \dots\}).$$

Thus the  $\sigma$ -field of a discrete r.v. is the  $\sigma$ -field generated by the corresponding countable partition.

(iii)  $(\Omega, \mathcal{F}) = (\mathbf{R}, \mathcal{B}(\mathbf{R}))$ , and  $X(\cdot)$  is continuous, strictly increasing and unbounded both from above and below. In this case an inverse function  $X^{-1}$  exists, and its range equals  $\mathbf{R}$ . Then

$$X^{-1}(-\infty, a] = (-\infty, X^{-1}(a)]$$

and since  $X^{-1}(a)$  takes all possible real values, we have

$$\left\{ X^{-1}(-\infty, a] : a \in \mathbf{R} \right\} = \left\{ (-\infty, a] : a \in \mathbf{R} \right\}.$$

As the collection on the r.h.s. generates the Borel field of the real line, we have

$$\sigma(X) = \mathcal{B}(\mathbf{R}) = \mathcal{F}.$$

Thus the r.v.  $X$  is *completely* informative about the underlying experiment: the occurrence or not of any event can be determined using  $X$ .

**Remark.** As it turns out, the above result is true even without assuming that  $X(\cdot)$  is continuous and unbounded. However, it is essential that  $X(\cdot)$  be *strictly* increasing. If it were not, e.g., if  $X(\cdot)$  were constant over an interval  $(c, d)$ , then the instrument  $\mathcal{M}_X$  would not be able to distinguish between outcomes lying in that interval. Thus no proper subset of  $(c, d)$  could lie in  $\sigma(X)$ , and  $\sigma(X) \neq \mathcal{B}(\mathbf{R})$ .

An alternative, somewhat more direct, representation of  $\sigma(X)$  exists by virtue of the following theorem.

**Theorem.** The  $\sigma$ -field generated by a random variable is the collection of all inverse images of Borel sets on the real line. Thus if  $X$  is a r.v. on  $(\Omega, \mathcal{F}, P)$ ,

$$\sigma(X) = \{X^{-1}(H) : H \in \mathcal{B}(\mathbf{R})\}.$$

To prove this theorem, we use the following simple lemma.

**Lemma.** Let  $\Omega$  be arbitrary, and  $f$  a real-valued function on  $\Omega$ . Then

(i) if  $\mathcal{B}$  is a  $\sigma$ -field of subsets of  $\mathbf{R}$ , the collection  $\{f^{-1}(H) : H \in \mathcal{B}\}$  is a  $\sigma$ -field of subsets of  $\Omega$ ;

(ii) if  $\mathcal{A}$  is a  $\sigma$ -field of subsets of  $\Omega$ , the collection  $\{H \subset \mathbf{R} : f^{-1}(H) \in \mathcal{A}\}$  is a  $\sigma$ -field of subsets of  $\mathbf{R}$ .

The proof of the lemma is straightforward and is left as an exercise. The identity  $f^{-1}(\bigcup_i H_i) = \bigcup_i f^{-1}(H_i)$  is useful in establishing closure under countable unions.

**Proof of Theorem.** For convenience let  $\mathcal{L} = \{X^{-1}(H) : H \in \mathcal{B}(\mathbf{R})\}$ .

$\mathcal{L} \supset \sigma(X)$ : Since  $\mathcal{B}(\mathbf{R})$  is a  $\sigma$ -field of subsets of  $\mathbf{R}$ , the collection  $\mathcal{L}$  is a  $\sigma$ -field of subsets of  $\Omega$  by statement (i) of the above lemma. Since all intervals lie in  $\mathcal{B}(\mathbf{R})$ , we have that

$$\mathcal{L} \supset \left\{ X^{-1}(-\infty, a] : a \in \mathbf{R} \right\},$$

and hence  $\mathcal{L}$  also contains the  $\sigma$ -field generated by the collection on the r.h.s. Thus  $\mathcal{L} \supset \sigma(X)$ .

$\mathcal{L} \subset \sigma(X)$ : Let

$$\mathcal{H} = \{H \subset \mathbf{R} : X^{-1}(H) \in \sigma(X)\}.$$

This collection is a  $\sigma$ -field by statement (ii) of the above lemma. Furthermore, it contains every interval  $(-\infty, a]$  since  $X^{-1}(-\infty, a] \in \sigma(X)$ . Hence  $\mathcal{H}$  also contains the  $\sigma$ -field generated by such intervals, namely the Borel field  $\mathcal{B}(\mathbf{R})$  of the real line. Thus  $X^{-1}(H) \in \sigma(X)$  for every Borel set  $H$ , or equivalently,  $\mathcal{L} \subset \sigma(X)$ .

The above characterization of  $\sigma(X)$  conforms with our intuition about event spaces and random variables. Indeed, the class of subsets  $H$  of the real line for which  $X \in H$  can be tested should be a  $\sigma$ -field. One would also expect this  $\sigma$ -field to be generated by the sets that are directly observable using the instrument  $\mathcal{M}_X$ , namely the intervals  $(-\infty, a]$ . Thus we obtain the Borel field, and infer that the class of events in  $\mathcal{F}$  that can be observed using  $X$  or  $\mathcal{M}_X$  is

$$\{X^{-1}(H) : H \in \mathcal{B}(\mathbf{R})\}.$$

This is precisely the statement of the above theorem.

**Remark.** Note that in Example (ii), the fact that  $E_j \in \sigma(X)$  follows directly from the above theorem by noting that  $\{c_j\} \in \mathcal{B}(\mathbf{R})$ .

**Equivalent definition of random variable.** From the above theorem it follows that  $X : \Omega \mapsto \mathbf{R}$  is a r.v. on  $(\Omega, \mathcal{F})$  if and only if  $X^{-1}(H) \in \mathcal{F}$  for every  $H \in \mathcal{B}(\mathbf{R})$ .

## 12. Operations on random variables

Billingsley, Sec. 13, *Measurable Mappings, Mappings onto  $R^k$* .

**Convention.** We shall reserve the term *random variable* for functions defined on a measurable space  $(\Omega, \mathcal{F})$  which is part of a probability space  $(\Omega, \mathcal{F}, P)$ . For functions which satisfy the definition of the random variable as given in Section 10, but which are not defined on a measurable space associated with a probability experiment, we will use the term **measurable function**.

### Operations on finite sets of random variables

**1.** Every piecewise continuous real-valued function on  $\mathbf{R}$  is a measurable function on  $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$  (proof of this fact can be found in Billingsley). Thus *most* real-valued functions on  $\mathbf{R}$  encountered in practice are measurable functions.

**2.** If  $X$  is a r.v. on  $(\Omega, \mathcal{F})$  and  $g$  is a measurable function on  $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$ , then the composition  $Y = g \circ X$  defined by

$$Y(\omega) = (g \circ X)(\omega) = g(X(\omega))$$

is also a random variable on  $(\Omega, \mathcal{F})$ .

To prove this, we show that  $Y^{-1}(H) \in \mathcal{F}$  for every Borel set  $H$ . Indeed,

$$\begin{aligned} Y^{-1}(H) &= \{\omega : g(X(\omega)) \in H\} = \{\omega : X(\omega) \in g^{-1}(H)\} \\ &= X^{-1}(g^{-1}(H)) . \end{aligned}$$

Since  $g$  is a measurable function on  $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$ , the set  $g^{-1}(H)$  is a Borel set; since  $X$  is a r.v. on  $(\Omega, \mathcal{F})$ , the set  $X^{-1}(g^{-1}(H))$  is an event.

Thus for example, if  $X$  is a r.v. on  $(\Omega, \mathcal{F})$ , so are the functions  $\alpha X$ ,  $e^X$  and  $I_{[a,b]}(X)$ .

**3.** (without proof) If  $X$  and  $Y$  are r.v.'s on the same measurable space, then so are  $X + Y$ ,  $XY$  and  $\max(X, Y)$ .

Thus if  $X_1, \dots, X_n$  are random variables on the same probability space, and

$$g : \mathbf{R}^n \mapsto \mathbf{R}$$

is a function composed by a finite number of additions, multiplications and maximizations, then

$$Y(\omega) = g(X_1(\omega), \dots, X_n(\omega))$$

defines yet another random variable on  $(\Omega, \mathcal{F})$ . As we shall see later, this property extends to most scalar-valued functions on  $\mathbf{R}^n$  encountered in practice.

### Operations on sequences of random variables

Suppose  $X_1, X_2, \dots$  is a sequence of random variables on  $(\Omega, \mathcal{F})$ . If

$$g : \mathbf{R}^{\mathbf{N}} \mapsto \mathbf{R} ,$$

then we can define a function  $Y(\cdot)$  on  $\Omega$  by

$$Y(\omega) = g(X_1(\omega), X_2(\omega), \dots) .$$

Operations such as the above on sequences of r.v.'s often yield infinite values. For example, if we take

$$Y(\omega) = \sum_{i=1}^{\infty} |X_i(\omega)| ,$$

then it is conceivable that  $Y(\omega)$  will be infinite for some  $\omega \in \Omega$ . Thus it seems desirable to extend the definition of a random variable to functions that take  $\pm\infty$  as values, provided these values are “observable” according to the interpretation of the previous section. We do so as follows.

### Definition

An **extended** random variable on  $(\Omega, \mathcal{F})$  is a function  $Y$  on  $\Omega$  into  $\mathbf{R} \cup \{-\infty, +\infty\} = [-\infty, +\infty]$  such that the set

$$Y^{-1}[-\infty, a] = \{\omega : Y(\omega) \leq a\}$$

lies in  $\mathcal{F}$  for every  $a \in [-\infty, \infty]$ .

An equivalent definition is the following:  $Y$  is an extended r.v. on  $(\Omega, \mathcal{F})$  if there exists a r.v.  $X$  on  $(\Omega, \mathcal{F})$  and events  $A_+$  and  $A_-$  in  $\mathcal{F}$  such that

$$X(\omega) = \begin{cases} +\infty & \text{if } \omega \in A_+ ; \\ -\infty & \text{if } \omega \in A_- ; \\ X(\omega) & \text{if } \omega \in (A_+ \cup A_-)^c . \end{cases}$$

For most transformations  $g$  of interest, the function  $Y = g(X_1, X_2, \dots)$  will be a random variable. We show this for the important case in which  $Y$  is the *supremum* (or *infimum*) of the random variables  $X_1, X_2, \dots$

### *Digression on suprema and infima*

We say that a nonempty set  $S$  of real numbers is **bounded from above** if there exists a number  $b \in \mathbf{R}$  such that every element in  $S$  is less than or equal to  $b$ . Such  $b$  is called an **upper bound** of the set  $S$ .

**Example.**  $S_1 = (-\infty, 1)$  is bounded from above by 2, which is an upper bound of  $S_1$ . The set  $S_2 = \mathbf{N}$  is not bounded from above.

**Axiom.** Every nonempty set  $S \subset \mathbf{R}$  that is bounded from above has a least upper bound (**l.u.b.**).

Based on this axiom, we define the **supremum** of a nonempty set  $S$  as follows:

$$\sup S = \begin{cases} \text{l.u.b. of } S, & \text{if } S \text{ is bounded from above;} \\ +\infty, & \text{otherwise.} \end{cases}$$

Listed below are a few elementary properties of the supremum:

**1.** If  $\sup S \in S$ , then  $\sup S = \max S$ .

**2.** If  $S$  is finite, then  $\sup S \in S$ . If  $S$  is infinite, then  $\sup S$  may or may not lie in  $S$ . For example,

$$\sup(0, 1] = 1 \in S, \quad \sup(0, 1) = 1 \notin S.$$

**3.** If  $S$  is bounded from above, then  $\sup S$  is the unique real number  $t$  with the property that for every  $a < t < b$ ,

$$(a, t] \cap S \neq \emptyset, \quad (t, b) \cap S = \emptyset.$$

**4.** For all  $a \in \mathbf{R}$ , the following equivalence is true:

$$\sup S \leq a \iff (\forall x \in S) x \leq a.$$

If  $S$  comprises the terms of an infinite sequence  $x_1, x_2, \dots$ , it is customary to write  $\sup_n x_n$  for  $\sup S$ .

The notion of the **infimum** can be developed along parallel lines. The concepts *bounded from below* and *greatest lower bound* should be transparent; we let  $\inf S$  equal the greatest lower bound if it exists,  $-\infty$  otherwise. Counterparts of properties (1–4) above can be obtained by noting that

$$\inf S = -\sup(-S) .$$

*End of Digression.*

We now claim that the function  $Y(\cdot)$  defined by

$$Y(\omega) = \sup_n X_n(\omega)$$

is an extended random variable on the space  $(\Omega, \mathcal{F})$ . To see this, consider the inverse images  $Y^{-1}[-\infty, a]$  for all finite and infinite values of  $a$ . Clearly

$$Y^{-1}\{-\infty\} = \emptyset , \quad Y^{-1}[-\infty, \infty] = \Omega .$$

For  $a$  finite, we use the equivalence stated in (4) above. Thus

$$\begin{aligned} Y^{-1}[-\infty, a] &= \{\omega : \sup_n X_n(\omega) \leq a\} = \{\omega : (\forall n) X_n(\omega) \leq a\} \\ &= \bigcap_{n \geq 1} \{\omega : X_n(\omega) \leq a\} = \bigcap_{n \geq 1} X_n^{-1}(-\infty, a] . \end{aligned}$$

Since each  $X_n$  is a random variable, the final expression is a countable intersection of events, hence also an event. This concludes our proof.

A similar argument establishes that  $\inf_n X_n$  is also an extended random variable. One can proceed from this point to show that the set  $C$  of sample points  $\omega$  over which  $\lim_n X_n(\omega)$  exists is an event, and the function thus defined on  $C$  is an extended r.v. on the restriction of  $(\Omega, \mathcal{F})$  on  $C$ .

The outline of the proof is as follows: The set  $C$  consists of those  $\omega$  for which

$$\sup_n \inf_{k \geq n} X_k(\omega) = \inf_n \sup_{k \geq n} X_k(\omega) ,$$

the common value being equal to  $\lim_n X_n(\omega)$ . By our earlier result on suprema and infima, both sides of the above equation define extended r.v.'s on  $(\Omega, \mathcal{F})$ , and thus the set  $C$  is an event. Also, for all  $a \in [-\infty, +\infty]$ , we have the equality

$$\{\omega : \lim_n X_n(\omega) \leq a\} = C \cap \{\sup_n \inf_{k \geq n} X_k(\omega) \leq a\} ,$$

which implies that the above set is an event. In this sense, the mapping  $\omega \mapsto \lim_n X_n(\omega)$  defines an extended r.v. on the restriction of  $(\Omega, \mathcal{F})$  on  $C$ .

### 13. The distribution of a random variable

Billingsley, Sec. 12, *Specifying Measures on the Line*; Sec. 14, *Distribution Functions*; Sec. 20, *Distributions*.

In continuing our discussion of random variables, we bring probability measures into play. Indeed, the most marked characteristic of a random variable  $X$  on  $(\Omega, \mathcal{F}, P)$  is that it *induces* a probability measure on  $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$  in the following fashion.

By the definition of random variable, the inverse image of any Borel set  $H$  under  $X$  is an event in  $\mathcal{F}$ . Thus if we let

$$P_X(H) \stackrel{\text{def}}{=} P(X^{-1}(H)) = P\{\omega : X(\omega) \in H\} ,$$

we obtain a well-defined set function  $P_X$  on  $\mathcal{B}(\mathbf{R})$ .

We claim that  $P_X$  is a probability measure on  $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$ . Nonnegativity is self-evident, and normalization follows from

$$P_X(\mathbf{R}) = P\{\omega : X(\omega) \in \mathbf{R}\} = P(\Omega) = 1 .$$

For countable additivity, consider a sequence  $H_1, H_2, \dots$  of pairwise disjoint Borel sets. It is easy to see that the inverse images under  $X$  will also be pairwise disjoint. Recalling that

$$X^{-1}\left(\bigcup_i H_i\right) = \bigcup_i X^{-1}(H_i) ,$$

and that  $P$  is countably additive, we obtain

$$\begin{aligned} P_X\left(\bigcup_i H_i\right) &= P\left[X^{-1}\left(\bigcup_i H_i\right)\right] = P\left[\bigcup_i X^{-1}(H_i)\right] \\ &= \sum_i P(X^{-1}(H_i)) = \sum_i P_X(H_i) . \end{aligned}$$

#### Definition

The **distribution** of a random variable  $X$  on  $(\Omega, \mathcal{F}, P)$  is the probability measure  $P_X$  on  $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$  defined by

$$(\forall H \in \mathcal{B}(\mathbf{R})) . \qquad P_X(H) = P(X^{-1}(H))$$

The interpretation of  $P_X$  is clear if we associate the random variable  $X$  with a *sub-experiment* of  $(\Omega, \mathcal{F}, P)$ . The sample space for this sub-experiment is the real line, the events are the Borel sets, and the probabilities of these events are given by the distribution of  $P_X$ . Thus in describing the random measurement  $X$ , we can restrict ourselves to the probability space  $(\mathbf{R}, \mathcal{B}(\mathbf{R}), P_X)$ .

## Specification of distributions

The **cumulative distribution function (cdf)**  $F_X$  of the random variable  $X$  on  $(\Omega, \mathcal{F}, P)$  is defined for all  $x \in \mathbf{R}$  by

$$F_X(x) \stackrel{\text{def}}{=} P_X(-\infty, x] = P\{\omega : X(\omega) \leq x\} .$$

$F_X$  has three essential properties:

(D1)  $F_X(x)$  is nondecreasing in  $x$ ;

(D2)  $\lim_{x \uparrow \infty} F_X(x) = 1$ ,  $\lim_{x \downarrow -\infty} F_X(x) = 0$ ;

(D3)  $F_X$  is everywhere *right continuous*:  $(\forall x) \lim_{\epsilon \downarrow 0} F_X(x + \epsilon) = F_X(x)$ .

To show (D1), we note that

$$x < x' \Rightarrow (-\infty, x] \subset (-\infty, x'] \Rightarrow P_X(-\infty, x] \leq P_X(-\infty, x'] .$$

Monotone continuity of  $P_X$  yields (D2) and (D3) as follows:

$$\lim_{x \rightarrow \infty} F_X(x) = \lim_{n \rightarrow \infty} P_X(-\infty, n] = P_X\left(\bigcup_{n=1}^{\infty} (-\infty, n]\right) = P_X(\mathbf{R}) = 1 ,$$

$$\lim_{x \rightarrow -\infty} F_X(x) = \lim_{n \rightarrow \infty} P_X(-\infty, -n] = P_X\left(\bigcap_{n=1}^{\infty} (-\infty, -n]\right) = P_X(\emptyset) = 0 ,$$

$$\lim_{\epsilon \downarrow 0} F_X(x + \epsilon) = \lim_{n \rightarrow \infty} P_X(-\infty, x + n^{-1}] = P_X\left(\bigcap_{n=1}^{\infty} (-\infty, x + n^{-1}]\right) = P_X(-\infty, x] .$$

It is worth noting that  $X$  is **not** always left continuous: if we evaluate the quantity  $\lim_{\epsilon \downarrow 0} F_X(x - \epsilon)$  in the above fashion, we obtain

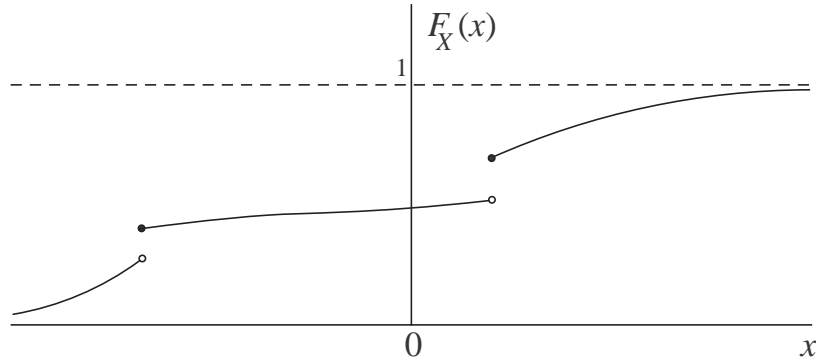
$$\lim_{\epsilon \downarrow 0} F_X(x - \epsilon) = \lim_{n \rightarrow \infty} P_X(-\infty, x - n^{-1}] = P_X\left(\bigcup_{n=1}^{\infty} (-\infty, x - n^{-1}]\right) = P_X(-\infty, x) .$$

Thus

$$\lim_{\epsilon \downarrow 0} F_X(x - \epsilon) = F_X(x) - P_X\{x\} ,$$

and  $F_X$  is continuous at  $x$  (both from the left and the right) if and only if  $P_X\{x\} = 0$ .

The general form of  $F_X$  is illustrated below. The magnitude of each jump is given by the probability of the corresponding abscissa.



Conditions (D1–3) are clearly *necessary* for a function to be a cumulative distribution function. As it turns out, they are also *sufficient*, in that any function possessing these properties is the cdf of some random variable. This is by virtue of the following theorem.

**Theorem.** Let  $F$  be a function on the real line satisfying conditions (D1–3). Then there exists a measure  $P$  on  $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$  such that

$$P(-\infty, a] = F(a)$$

for every  $a \in \mathbf{R}$ . Furthermore, there exists on some probability space  $(\Omega, \mathcal{F}, P)$  a random variable  $X$  such that  $P_X \equiv P$ ,  $F_X \equiv F$ .

To prove this theorem, we follow a procedure similar to the construction of the Lebesgue measure in Section 9. Briefly, we consider the field  $\mathcal{F}_0$  consisting of finite disjoint unions of bounded semi-open intervals on the real line. We define a set function  $Q$  on  $\mathcal{F}_0$  by letting

$$Q(\emptyset) = 0, \quad Q(c, d] = F(d) - F(c),$$

and extending  $Q$  to all members of  $\mathcal{F}_0$  in the obvious manner. It is possible to show that  $Q$  is countably additive on  $\mathcal{F}_0$  and thus possesses a unique extension to a probability measure  $P$  on  $\sigma(\mathcal{F}_0) = \mathcal{B}(\mathbf{R})$ . Then for all  $a \in \mathbf{R}$ ,

$$P(-\infty, a] = \lim_{n \rightarrow \infty} P(-n, a] = Q(a) - \lim_{n \rightarrow \infty} Q(-n) = Q(a).$$

The second statement of the theorem follows immediately from the above construction if we take  $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$  as the probability space, with  $X$  being the identity transformation  $X(\omega) = \omega$ . Indeed for every Borel set  $H$ ,

$$P_X(H) = P\{\omega : X(\omega) \in H\} = P\{\omega : \omega \in H\} = P(H),$$

and thus also  $F_X \equiv F$ .

Thus there is a one-to-one correspondence between probability measures on the real line and functions satisfying (D1–3). Any such measure is completely specified by its cdf, and conversely, any function possessing these three properties defines a probability measure on the real line.

The Lebesgue measure on the unit interval can be obtained from the cdf

$$F(x) = \begin{cases} 1, & \text{if } x > 1; \\ x, & \text{if } 0 < x \leq 1; \\ 0, & \text{if } x \leq 0. \end{cases}$$

Indeed, the corresponding measure  $P$  is such that

$$P(0, 1] = F(1) - F(0) = 1 ,$$

and for every  $(a, b] \subset (0, 1]$ ,

$$P(a, b] = F(b) - F(a) = b - a .$$

Thus the restriction of  $P$  on the Borel subsets of the unit interval is identical to the Lebesgue measure on that interval.

**Notation.** We will also use the symbols  $\mu$  and  $\nu$  to denote measures on the real line or subsets thereof.

## Decomposition of Distributions

Any cdf  $F$  can be decomposed into cdf's of three distinct types: **discrete**, **absolutely continuous** and **singular**. By this we mean that we can write

$$F \equiv \alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_3 ,$$

where

$$\alpha_1, \alpha_2, \alpha_3 \geq 0 , \quad \alpha_1 + \alpha_2 + \alpha_3 = 1 ,$$

and the functions  $F_1$ ,  $F_2$  and  $F_3$  have the following properties.

$F_1$  is a *discrete cdf*. A discrete cdf corresponds to a measure which assigns probability one to a *countable* set  $\{c_1, c_2, \dots\}$  of real numbers. It is easily seen that all discrete random variables have discrete cdf's. Assuming that the probabilities of the  $c_i$ 's are given by the pdf  $(p_i)_{i \in \mathbf{N}}$ , we can write

$$F_1(x) = \sum_{i: c_i \leq x} p_i = \sum_{i \in \mathbf{N}} p_i I_{[c_i, \infty)}(x) ,$$

and thus  $F_1$  is the countable sum of weighted step functions.

$F_2$  is an *absolutely continuous cdf*. An absolutely continuous cdf has no jumps. Its defining property is

$$(\forall x) \quad F_2(x) = \int_{-\infty}^x f(t) dt$$

for a nonnegative function  $f$  called the **probability density function**, or **pdf**, or simply **density** of  $F_2$ .

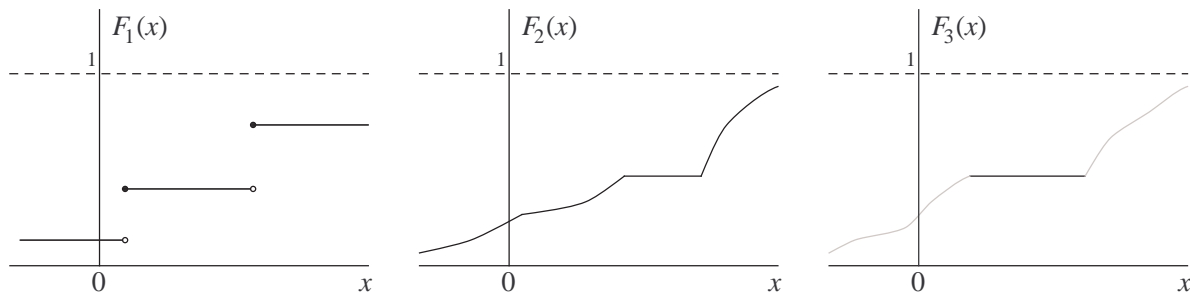
**Remark.** The above integral can be interpreted in the usual (Riemann) sense; this is certainly adequate from the computational viewpoint, as most densities encountered in calculations are piecewise continuous, hence Riemann integrable. For such densities, it is also true that

$$F_2'(x) = f(x)$$

at all points of continuity of  $f$ ; furthermore, the values taken by  $f$  at its points of discontinuity do not affect the resulting cdf  $F_2$ . Thus in most cases of interest, the density of an absolutely continuous cdf is obtained by simple differentiation.

$F_3$  is a *singular cdf*. Singular (or more precisely, continuous and singular) cdf's are seldom encountered in probability models. Such functions combine the seemingly incompatible features of continuity, monotone increase from 0 to 1, and *almost everywhere* existence of a derivative that is equal to zero. The last property is interpreted by saying that any interval of unit length has a Borel subset  $A$  of full (i.e., unit) measure such that  $F_3'(x) = 0$  on  $A$ . Examples of such unusual functions can be found in Billingsley (under *Singular Functions*, Sec. 31) and various textbooks on real analysis (under *Cantor ternary function*). We shall not concern ourselves with explicit forms of singular cdf's.

The general forms of  $F_1$ ,  $F_2$  and  $F_3$  is illustrated below; unfortunately, the microscopic irregularities of  $F_3$  are not discernible on graphs of finite precision and resolution.



Before looking at a numerical example, we note that the above decomposition of cdf's also holds for probability measures on the real line. Thus if  $\mu$  is the measure on  $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$  corresponding to the cdf  $F$ , we can write

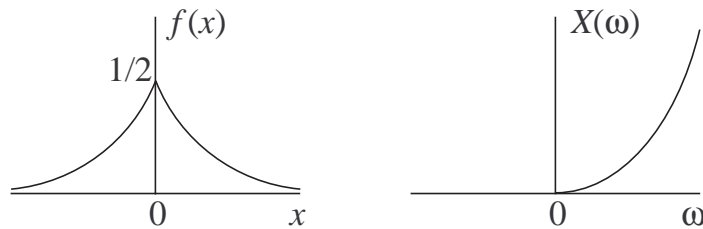
$$\mu \equiv \alpha_1 \mu_1 + \alpha_2 \mu_2 + \alpha_3 \mu_3 ,$$

where for every  $i$ ,  $\mu_i$  corresponds to  $F_i$ .

**Example.** Let  $(\Omega, \mathcal{F}, P) = (\mathbf{R}, \mathcal{B}(\mathbf{R}))$ , where the measure  $\mu$  has the *Laplace density*

$$f(x) = \frac{1}{2} e^{-|x|} \quad (x \in \mathbf{R}) .$$

Define the random variable  $X$  on  $(\Omega, \mathcal{F}, P)$  by  $X(\omega) = \omega^2 I_{[0, \infty)}(\omega)$ .



Since  $X$  is nonnegative, we have for  $x < 0$

$$F_X(x) = 0 .$$

For  $x \geq 0$  we have  $X(\omega) \leq x \Leftrightarrow \omega \leq \sqrt{x}$ , and thus

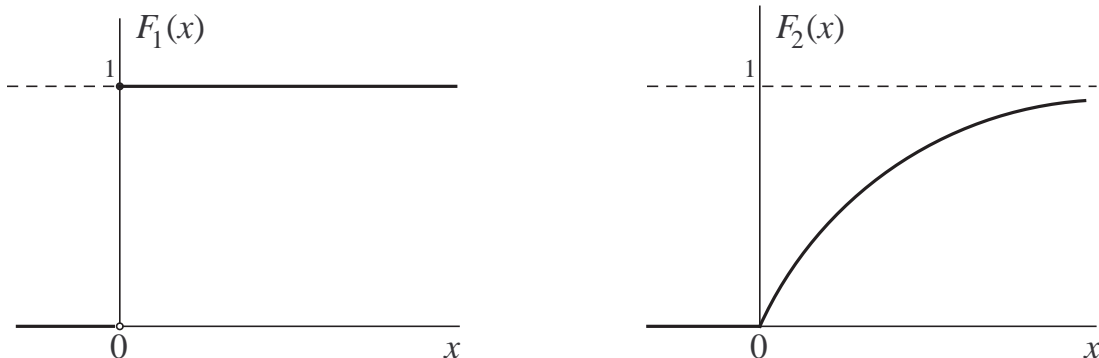
$$\begin{aligned} F_X(x) &= \mu(-\infty, \sqrt{x}] \\ &= \int_{-\infty}^{\sqrt{x}} f(t) dt = 1 - \frac{1}{2} \int_{\sqrt{x}}^{\infty} e^{-t} dt = 1 - \frac{1}{2} e^{-\sqrt{x}} . \end{aligned}$$

Thus

$$F_X(x) = \left(1 - \frac{1}{2} e^{-\sqrt{x}}\right) I_{[0, \infty)}(\omega) = \frac{1}{2} F_1(x) + \frac{1}{2} F_2(x) ,$$

where

$$F_1(x) = I_{[0, \infty)}(x) , \quad F_2(x) = (1 - e^{-\sqrt{x}}) I_{[0, \infty)}(x) .$$



$F_2$  is differentiable everywhere except at the origin:

$$F_2'(x) = 0 \quad (x < 0) , \quad F_2'(x) = \frac{1}{2\sqrt{x}} e^{-\sqrt{x}} \quad (x > 0) .$$

Thus we may take the density of  $F_2$  as

$$f_2(x) = \frac{e^{-\sqrt{x}}}{2\sqrt{x}} I_{(0,\infty)}(x) .$$

#### 14. Random vectors

Billingsley, Sec. 12, *Specifying Measures in  $\mathbf{R}^k$* ; Sec. 13, *Mappings into  $\mathbf{R}^k$* ; Sec. 20, *Random Variables and Vectors, Distributions*.

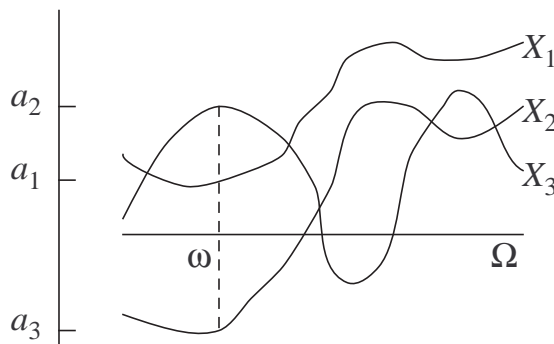
A  $k$ -dimensional **random vector** on  $(\Omega, \mathcal{F}, P)$  is an ordered  $k$ -tuple of random variables on the same probability space. Thus any  $k$  random variables  $X_1, \dots, X_k$  define a random vector

$$\mathbf{X} = (X_1, \dots, X_k) ,$$

and  $\mathbf{X}$  defines a mapping on  $\Omega$  into  $\mathbf{R}^k$  through

$$\mathbf{X}(\omega) = (X_1(\omega), \dots, X_k(\omega)) .$$

The figure below illustrates a three-dimensional random vector  $\mathbf{X} = (X_1, X_2, X_3)$ . At the particular sample point  $\omega$  shown,  $\mathbf{X}(\omega) = (a_1, a_2, a_3)$ .



To interpret the concept of a  $k$ -dimensional random vector  $\mathbf{X}$ , we follow our earlier development on random variables. We postulate the existence of an instrument  $\mathcal{M}_{\mathbf{X}}$  that can be set in an infinity of modes  $\mathcal{M}_{\mathbf{X}, \mathbf{a}}$  indexed by

$$\mathbf{a} = (a_1, \dots, a_k) \in \mathbf{R}^k .$$

For any such  $\mathbf{a}$ , it is possible to determine whether or not

$$X_1(\omega) \leq a_1 , \dots , X_k(\omega) \leq a_k$$

simultaneously, or equivalently, whether or not  $\mathbf{X} \in C$ , where

$$C = (-\infty, a_1] \times \dots \times (-\infty, a_k] .$$

The corresponding event in  $\mathcal{F}$  is

$$\mathbf{X}^{-1}(C) = \{\omega : X_1(\omega) \leq a_1, \dots, X_k(\omega) \leq a_k\} .$$

A set  $C \subset \mathbf{R}^k$  of the above form will be called a  $k$ -dimensional **lower rectangle**. In many respects, lower rectangles are to random vectors what intervals  $(-\infty, a]$  are to random variables (in the one-dimensional case, all lower rectangles are intervals of that form). This will become transparent in what follows.

### The sigma-field generated by a random vector

For a  $k$ -dimensional random vector  $\mathbf{X}$  on  $(\Omega, \mathcal{F}, P)$ , we define  $\sigma(\mathbf{X})$  as the sigma-field generated by the directly observable events, i.e., the inverse images of lower rectangles. This is in direct analogy to our definition of  $\sigma(X)$  in the one-dimensional case. Thus

$$\sigma(\mathbf{X}) \stackrel{\text{def}}{=} \sigma\left(\{\mathbf{X}^{-1}(C_{\mathbf{a}}) : \mathbf{a} \in \mathbf{R}^k\}\right) ,$$

where  $C_{\mathbf{a}}$  denotes the lower rectangle with vertex at the point  $\mathbf{a} = (a_1, \dots, a_n)$ :

$$C_{\mathbf{a}} \stackrel{\text{def}}{=} (-\infty, a_1] \times \dots \times (-\infty, a_k] .$$

Recall that in the one-dimensional case, a direct representation for  $\sigma(X)$  was obtained in terms of the inverse images of all Borel sets in  $\mathbf{R}$ , or sets “generated” by intervals  $(-\infty, a]$ . The same argument can be applied in the  $k$ -dimensional case to show that  $\sigma(\mathbf{X})$  consists of the inverse images of all Borel sets in  $\mathbf{R}^k$ , i.e., sets “generated” by lower rectangles  $C_{\mathbf{a}}$ . We therefore have

$$\sigma(\mathbf{X}) = \{\mathbf{X}^{-1}(H) : H \in \mathcal{B}(\mathbf{R}^k)\} ,$$

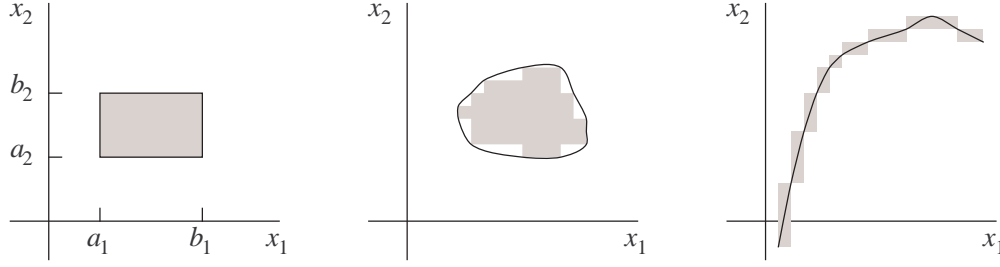
where  $\mathcal{B}(\mathbf{R}^k)$  is the Borel field of the  $k$ -dimensional Euclidean space:

$$\mathcal{B}(\mathbf{R}^k) \stackrel{\text{def}}{=} \sigma\{C_{\mathbf{a}} : \mathbf{a} \in \mathbf{R}^k\} .$$

The collection  $\mathcal{B}(\mathbf{R}^k)$  contains most sets of interest in  $\mathbf{R}^k$ . In particular, smooth shapes and surfaces associated with continuous mappings on  $\mathbf{R}^{k-j}$  into  $\mathbf{R}^j$  are Borel sets. In the two-dimensional case, for example, most “nice” sets are Borel because they are expressible in terms of countably many operations on rectangles  $(a_1, b_1] \times (a_2, b_2]$ ; these are Borel since

$$(a_1, b_1] \times (a_2, b_2] = \left( C_{b_1 b_2} - C_{b_1 a_2} \right) \cap \left( C_{b_1 b_2} - C_{a_1 b_2} \right) .$$

This is illustrated in the figure below.



### The distribution of a random vector

The **distribution** of the random vector  $\mathbf{X} = (X_1, \dots, X_k)$  on  $(\Omega, \mathcal{F}, P)$  is the probability measure  $P_{\mathbf{X}}$  on  $(\mathbf{R}^k, \mathcal{B}(\mathbf{R}^k))$  defined by the following relationship:

$$P_{\mathbf{X}}(H) \stackrel{\text{def}}{=} P(\mathbf{X}^{-1}(H)) = P\{\omega : \mathbf{X}(\omega) \in H\} .$$

That  $P_{\mathbf{X}}$  is indeed a probability measure is easily established as in the one-dimensional case. In describing the sub-experiment that entails observation of the random vector  $\mathbf{X}$ , it suffices to consider the probability space  $(\mathbf{R}^k, \mathcal{B}(\mathbf{R}^k), P_{\mathbf{X}})$ .

The distribution  $P_{\mathbf{X}}$  yields a **cumulative distribution function**  $F_{\mathbf{X}}$ , which gives, for every point  $\mathbf{x} = (x_1, \dots, x_k)$ , the probability of the lower rectangle with vertex at that point:

$$F_{\mathbf{X}}(\mathbf{x}) \stackrel{\text{def}}{=} P_{\mathbf{X}}(C_{\mathbf{x}}) = P\{\omega : X_1(\omega) \leq x_1, \dots, X_k(\omega) \leq x_k\} .$$

We also call  $F_{\mathbf{X}}$  the **joint cdf** of  $X_1, \dots, X_k$ , and use the alternative forms

$$F_{(X_1, \dots, X_k)} \equiv F_{X_1 \dots X_k} \equiv F_{\mathbf{X}} .$$

A  $j$ -dimensional **marginal** cdf of  $\mathbf{X}$  is the cdf of any  $j$ -dimensional subvector of  $\mathbf{X}$ , i.e., any random vector obtained from  $\mathbf{X}$  by eliminating  $k - j$  of its constituent random variables. Any marginal cdf can be easily computed from  $F_{\mathbf{X}}$  using the formula

$$F_{X_1 \dots X_j}(x_1, \dots, x_j) = \lim_{x_{j+1} \uparrow \infty, \dots, x_k \uparrow \infty} F_{X_1 \dots X_k}(x_1, \dots, x_k) .$$

This is so because given any  $k - j$  sequences  $(a_n^{(j+1)})_{n \in \mathbf{N}}, \dots, (a_n^{(k)})_{n \in \mathbf{N}}$  that increase to infinity with  $n$ , we can write

$$\begin{aligned} \{\omega : X_1(\omega) \leq x_1, \dots, X_j(\omega) \leq x_j\} &= \\ &= \lim_n \{\omega : X_1(\omega) \leq x_1, \dots, X_j(\omega) \leq x_j, X_{j+1} \leq a_n^{(j+1)}, \dots, X_k \leq a_n^{(k)}\} \end{aligned}$$

The sought result then follows by monotone continuity of  $P$ .

Counterparts of properties (D1–3) are easily obtained by re-interpreting inequalities and limits *componentwise*. Thus

(1) if  $x_i < y_i$  for all  $i$ , then  $F_{\mathbf{X}}(\mathbf{x}) \leq F_{\mathbf{X}}(\mathbf{y})$ .

(2)  $F_{\mathbf{X}}$  approaches unity when *all* arguments jointly increase to  $+\infty$ ; it approaches zero when *one* or more arguments decrease to  $-\infty$ ;

(3) if  $y_i \downarrow x_i$  for all  $i$ , then  $F_{\mathbf{X}}(\mathbf{y}) \downarrow F_{\mathbf{X}}(\mathbf{x})$ .

It is worth mentioning that unlike the one-dimensional case, the above conditions are **not** sufficient for a  $k$ -dimensional cdf if  $k > 1$ . One needs to strengthen (1) so as to ensure that every rectangle

$$(a_1, b_1] \times \cdots \times (a_k, b_n]$$

is assigned nonnegative probability; in its present form, (1) only guarantees this for proper differences of lower rectangles. Details on the needed modifications and proofs are given in Billingsley, Sec. 18. For our purposes, it suffices to note that any  $k$ -dimensional distribution is completely specified by its cdf.

We say that  $F_{\mathbf{X}}$  is **discrete** if the measure  $P_{\mathbf{X}}$  on  $(\mathbf{R}^k, \mathcal{B}(\mathbf{R}^k))$  assigns probability 1 to a countable set in  $\mathbf{R}^k$ ; this is certainly true if the vector  $\mathbf{X}$  is itself discrete. The general form of  $F_{\mathbf{X}}$  can be obtained as in the one-dimensional case, and can be written as a weighted sum of countably many indicator functions of *upper rectangles*

$$[a_1, \infty) \times \cdots \times [a_k, \infty) .$$

$F_{\mathbf{X}}$  is **absolutely continuous** if it possesses a nonnegative pdf  $f_{\mathbf{X}} = f_{X_1 \dots X_k}$  such that for all  $\mathbf{x} \in \mathbf{R}^k$ ,

$$F_{X_1 \dots X_k}(x_1, \dots, x_k) = \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f_{X_1 \dots X_k}(t_1, \dots, t_k) dt_1 \cdots dt_k .$$

By our earlier result, the marginal cdf for the first  $j$  components is given by

$$F_{X_1 \dots X_j}(x_1, \dots, x_j) = \int_{t_k=-\infty}^{t_k=\infty} \cdots \int_{t_{j+1}=-\infty}^{t_k=\infty} \int_{t_j=-\infty}^{t_j=x_j} \cdots \int_{t_1=-\infty}^{t_1=x_1} f_{X_1 \dots X_k}(t_1, \dots, t_k) dt_1 \cdots dt_k .$$

It follows that  $F_{X_1 \dots X_j}$  is itself absolutely continuous with density

$$f_{X_1 \dots X_j} = \int_{t_k=-\infty}^{t_k=\infty} \cdots \int_{t_{j+1}=-\infty}^{t_{j+1}=\infty} f_{X_1 \dots X_k}(t_1, \dots, t_k) dt_{j+1} \cdots dt_k .$$

In most practical cases, a pdf  $f$  of an absolutely continuous  $F$  can be obtained by taking a mixed partial derivative of  $F$  where such derivative exists, and setting  $f$  equal to 0 elsewhere (usually on a simple boundary  $A \subset \mathbf{R}^k$ ):

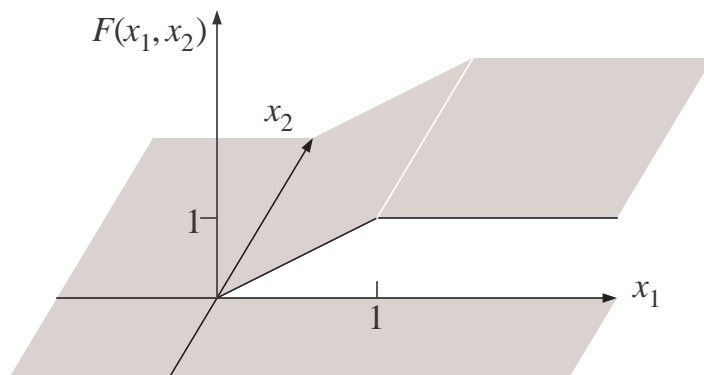
$$f(x_1, \dots, x_k) = \frac{\partial^k F(x_1, \dots, x_k)}{\partial x_1 \cdots \partial x_k} I_{A^c}(x_1, \dots, x_k) .$$

As in the one-dimensional case, it is possible to decompose a  $k$ -dimensional distribution  $F$  into components  $F_1$  (discrete),  $F_2$  (absolutely continuous) and  $F_3$  (singular). One important difference here is that for  $k > 1$ , the singular component  $F_3$  need not be a continuous function if  $k > 1$ .

To see an instance of this, consider the space  $((0, 1], \mathcal{B}((0, 1]), \mu)$  with  $\mu =$  Lebesgue measure, and let  $X_1(\omega) = \omega$ ,  $X_2(\omega) = 0$ . It is easily shown that the cdf of the random vector  $(X_1, X_2)$  is given by

$$F_{X_1 X_2}(x_1, x_2) = \left( x_1 I_{(0,1]}(x_1) + I_{(1,\infty)}(x_1) \right) I_{[0,\infty)}(x_2) .$$

This function is illustrated below, where the discontinuity along the half-line  $x_1 \geq 0$ ,  $x_2 = 0$ , becomes evident.



The mixed partial derivative  $\partial^2 F_{X_1 X_2}(x_1, x_2) / \partial x_1 \partial x_2$  is zero everywhere except on a handful of half-lines. Now lines on a plane are as paltry as points on a line: if we define a Lebesgue measure on the unit square by assigning to each rectangle a measure equal to its area, then any line segment contained in that square will have probability zero. Thus  $\partial^2 F(x_1, x_2) / \partial x_1 \partial x_2$  is “almost everywhere” zero, and by extrapolating from our definition of singularity in the one dimensional-case, we can say that  $F_{X_1 X_2}$  is singular.

It is worth noting here that the discontinuity of  $F_{X_1 X_2}$  in the above example is somewhat coincidental: the corresponding measure  $P_{X_1 X_2}$  is uniformly distributed over an interval of unit length that happens to be aligned with one of the axes. A simple rotation of that interval can remove this discontinuity. Thus if on the above probability space we define

$$X_3(\omega) = X_4(\omega) = \omega / \sqrt{2} ,$$

then  $P_{X_3 X_4}$  is uniform over an interval of unit length, yet

$$F_{X_3 X_4}(x_3, x_4) = \left( \sqrt{2} x_3 \wedge \sqrt{2} x_4 \wedge 1 \right) \vee 0 .$$

(Here “ $\wedge$ ” denotes *minimum* and “ $\vee$ ” *maximum*.) This function is clearly continuous everywhere.

In conclusion, we should note that discrete and singular measures on  $\mathbf{R}^k$  share one essential feature: they assign probability one to Borel sets whose intersection with any  $k$ -dimensional unit cube has zero Lebesgue measure, or “volume.” The key difference is that discrete measures are concentrated on (countably many) points, whereas singular ones are not. This raises a question as to why singletons alone should receive special treatment in higher-dimensional spaces (in the case  $k = 3$ , for example, one could also distinguish between singular measures concentrated on curves and ones concentrated on surfaces). It is difficult to give a satisfactory answer; in fact, in the mathematical literature, all measures assigning unit probability to sets of Lebesgue measure zero are termed singular, and thus discrete measures are subsumed under singular ones.

## 15. Independence

Billingsley, Sec. 4, *Independent Events*; Sec. 20, *Independence*.

### Independent events

The concept of independence stems from elementary considerations of conditional probability. Recall that if  $A$  and  $B$  are two events in a probability space  $(\Omega, \mathcal{F}, P)$  such that  $P(A) > 0$ , then the *conditional probability of  $B$  given  $A$*  is defined by

$$P(B|A) = \frac{P(B \cap A)}{P(A)} .$$

*Independence of  $A$  and  $B$*  is defined by

$$P(A \cap B) = P(A)P(B) .$$

This is easily seen to imply  $P(B|A) = P(B)$  if  $P(A) > 0$ , and  $P(A|B) = P(A)$  if  $P(B) > 0$ . Thus intuitively, two events are independent if knowledge of occurrence of either event does not affect the assessment of likelihood of the other.

A finite collection  $\{A_1, \dots, A_n\}$  of events in  $(\Omega, \mathcal{F}, P)$  is independent if for every set of distinct indices  $\{i_1, \dots, i_r\} \subset \{1, \dots, n\}$ , the following is true:

$$P(A_{i_1} \cap \dots \cap A_{i_r}) = P(A_{i_1}) \dots P(A_{i_r}) .$$

It is worth noting that independence is not affected if we replace one or more events in the collection by their complement(s). To see this, assume independence of the above collection and write

$$\begin{aligned} P(A_{i_1} \cap \dots \cap A_{i_r}) &= P(A_{i_1}) \dots P(A_{i_r}) , \\ P(A_{i_2} \cap \dots \cap A_{i_r}) &= P(A_{i_2}) \dots P(A_{i_r}) . \end{aligned}$$

Upon subtraction, we obtain

$$P(A_{i_1}^c \cap A_{i_2} \cap \dots \cap A_{i_r}) = P(A_{i_1}^c)P(A_{i_2}) \dots P(A_{i_r}) ,$$

and thus independence will still hold if we replace  $A_{i_1}$  by  $A_{i_1}^c$ .

The above observation leads to the following alternative (albeit hardly more economical) definition of independence:  $A_1, \dots, A_n$  are independent if and only if

$$P(B_1 \cap \dots \cap B_n) = P(B_1) \cdots P(B_n)$$

for every choice of events  $B_i$  such that  $B_i = A_i$  or  $B_i = A_i^c$ . The “only if” part was established above. To see that this condition implies independence (the “if” part), we can show that the above product relationship is true for all choices of  $n - 1$  events  $B_i$ , and then proceed to  $n - 2$ ,  $n - 3$ , etc. (or use induction for brevity).

**Remark.** Independence of a collection of events implies *pairwise* independence of events in the collection:

$$\{A_1, \dots, A_n\} \text{ independent} \iff (\forall i \neq j) \{A_i, A_j\} \text{ independent} .$$

It is not difficult to show by counterexample that the reverse implication is **not** true. Thus it should be borne in mind that independence of a collection is equivalent to independence of *all* its subcollections; there is nothing special about subcollections of size two.

### Independent random variables and vectors

A finite collection of random variables  $\{X_1, \dots, X_n\}$  on the same probability space  $(\Omega, \mathcal{F}, P)$  is independent if for every choice of linear Borel sets  $H_1, \dots, H_n$ ,

$$P_{X_1 \dots X_n}(H_1 \times \dots \times H_n) = P_{X_1}(H_1) \cdots P_{X_n}(H_n) ,$$

or equivalently,

$$P\left(X_1^{-1}(H_1) \cap \dots \cap X_n^{-1}(H_n)\right) = P(X_1^{-1}(H_1)) \cdots P(X_n^{-1}(H_n)) .$$

Put differently,  $\{X_1, \dots, X_n\}$  is independent if for every choice of  $A_1 \in \sigma(X_1), \dots, A_n \in \sigma(X_n)$ , the collection  $\{A_1, \dots, A_n\}$  is itself independent (convince yourselves!).

Independence of random variables implies that their joint cdf can be written in a product form:

$$F_{X_1 \dots X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n) . \tag{1}$$

This is seen by substituting  $H_i = (-\infty, x_i]$  in the above definition.

If each cdf  $F_{X_i}$  is absolutely continuous with pdf  $f_{X_i}$ , the above product relationship (1) can be written in the equivalent form

$$f_{X_1 \dots X_n}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) . \tag{2}$$

To see this, consider the integral of the  $n$ -variate function on right-hand side over the lower rectangle with vertex at  $(x_1, \dots, x_n)$ :

$$\begin{aligned} \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f_{X_1}(t_1) \cdots f_{X_n}(t_n) dt_1 \cdots dt_n &= \left( \int_{-\infty}^{x_1} f_{X_1}(t_1) dt_1 \right) \cdots \left( \int_{-\infty}^{x_n} f_{X_n}(t_n) dt_n \right) \\ &= F_{X_1}(x_1) \cdots F_{X_n}(x_n) . \end{aligned}$$

Thus if (1) is true, the joint pdf can be taken to be the product of the marginal pdf's. Conversely, if (2) is true, the joint cdf is the product of the marginal cdf's.

For discrete r.v.'s, we can write a similar product relationship for the pmf's by taking the sets  $H_i$  to be singletons. Thus if  $x_1, \dots, x_n$  are points in the range of  $X_1, \dots, X_n$ , then

$$P_{X_1 \dots X_n} \left\{ (x_1, \dots, x_n) \right\} = P_{X_1} \{x_1\} \cdots P_{X_n} \{x_n\} . \quad (3)$$

The equivalence of (1) and (3) can be established as above; integrals are simply replaced by sums.

In the foregoing discussion we saw that the product relationship (1) for cdf's is necessary for independence. As it turns out, it is also *sufficient*; thus to establish independence of  $X_1, \dots, X_n$ , it suffices to show that

$$F_{X_1 \dots X_n} (x_1, \dots, x_n) = F_{X_1} (x_1) \cdots F_{X_n} (x_n) .$$

This equivalence is proved with the aid of the following important fact

**Theorem.** Let  $B$  be a fixed event and  $\mathcal{A}$  a collection of events that is *closed under intersection* and such that

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

for every  $A$  in  $\mathcal{A}$ . Then the same product relationship also holds for every  $A$  in  $\sigma(\mathcal{A})$ .

For a proof of the above statement, see Billingsley, Sec. 4. We apply it by considering the collection

$$\mathcal{A} = \{X_1^{-1}(-\infty, x_1] : x_1 \in \mathbf{R}\} ,$$

which is both closed under intersection and such that  $\sigma(\mathcal{A}) = \sigma(X_1)$ . We also let

$$B = X_2^{-1}(-\infty, x_2] \cap \cdots \cap X_n^{-1}(-\infty, x_n]$$

for fixed  $x_2, \dots, x_n$ . Assuming that (1) is valid, we can write

$$F_{X_1 \dots X_n} (x_1, \dots, x_n) = F_{X_1} (x_1) F_{X_2 \dots X_n} (x_2, \dots, x_n) ,$$

or equivalently,

$$P(A \cap B) = P(A)P(B) .$$

By the above theorem, this will also be true for all  $A \in \sigma(\mathcal{A}) = \sigma(X_1)$ , and thus for every Borel set  $H_1$ ,

$$P_{X_1 \dots X_n} (H_1 \times C_{x_2 \dots x_n}) = P_{X_1} (H_1) P_{X_2 \dots X_n} (C_{x_2 \dots x_n}) .$$

We continue in the same fashion to replace  $(-\infty, x_2]$  by  $H_2$ , and so on.

**Exercise.** Using a similar technique, prove the following intuitively plausible fact: if the collection  $\{X, Y, Z\}$  is independent and  $g, h$  are Borel measurable functions on  $\mathbf{R}$  and  $\mathbf{R}^2$ , respectively, then the collection  $\{g(X, Y), h(Z)\}$  is also independent.

The concept of independence can be extended to random vectors by replacing univariate distributions by multivariate ones. Thus two random vectors  $\mathbf{X} \in \mathbf{R}^k$  and  $\mathbf{Y} \in \mathbf{R}^l$  are independent (or, more precisely, form an independent pair) if

$$P_{\mathbf{X}, \mathbf{Y}}(G \times H) = P_{\mathbf{X}}(G)P_{\mathbf{Y}}(H)$$

for all  $G \in \mathcal{B}(\mathbf{R}^k)$  and  $H \in \mathcal{B}(\mathbf{R}^l)$ ; equivalently,

$$F_{\mathbf{X}, \mathbf{Y}}(\mathbf{x}, \mathbf{y}) = F_{\mathbf{X}}(\mathbf{x})F_{\mathbf{Y}}(\mathbf{y})$$

for all  $\mathbf{x} \in \mathbf{R}^k$  and  $\mathbf{y} \in \mathbf{R}^l$ . Note that this does **not** imply that  $F_{\mathbf{X}}$  and  $F_{\mathbf{Y}}$  can each be written in product form, i.e., that the individual components of  $\mathbf{X}$  and  $\mathbf{Y}$  are independent. On the other hand, by taking suitable limits on both sides of the last relationship, it is easily seen that any subvector of  $\mathbf{X}$  and any subvector of  $\mathbf{Y}$  will form an independent pair. For example, if  $\mathbf{X} = (X_1, X_2, X_3)$  and  $\mathbf{Y} = (Y_1, Y_2)$  are independent, then so are  $(X_1, X_2)$  and  $Y_1$ ; yet neither of the collections  $\{X_1, X_2, X_3\}$  and  $\{Y_1, Y_2\}$  need be independent. The above definition can be extended to a finite collection of random vectors  $\{\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(n)}\}$  in an obvious way, and product relationships for pdf's and pmf's can also be established.

It is worth noting here that independence of random variables can be used to define independence of events:

$$\{A_1, \dots, A_n\} \text{ independent} \iff \{I_{A_1}, \dots, I_{A_n}\} \text{ independent} .$$

To see this, choose events  $B_1, \dots, B_n$  such that  $B_i$  is either  $A_i$  or  $A_i^c$ ; correspondingly, let  $b_i = 1$  if  $B_i = A_i$ , and  $b_i = 0$  if  $B_i = A_i^c$ . Then

$$\begin{aligned} (\forall i) \quad P(B_i) &= P\{\omega : I_{A_i}(\omega) = b_i\} , \\ P(B_1 \cap \dots \cap B_n) &= P\left\{\omega : I_{A_1}(\omega) = b_1, \dots, I_{A_n}(\omega) = b_n\right\} . \end{aligned}$$

Thus  $\{A_1, \dots, A_n\}$  is an independent collection if and only if the indicator functions  $I_{A_1}, \dots, I_{A_n}$  satisfy the product relationship (3) above, i.e., they are independent.

An *infinite* collection of events, random variables, or random vectors is termed independent if every *finite* subcollection is independent. Thus a sequence of random variables  $X_1, X_2, \dots$  is independent if and only if for every value of  $n$ , the collection  $\{X_1, \dots, X_n\}$  is independent.

**Definition.** A collection of random variables  $X_i$  (or random vectors  $\mathbf{X}^{(i)}$  of the same dimension) is **identically distributed** if every  $X_i$  (or every  $\mathbf{X}^{(i)}$ ) has the same distribution. If, in addition, the collection is independent, the collection is termed **independent and identically distributed (i.i.d.)**.

## Examples

1. Let  $X_1, \dots, X_n$  be independent random variables and define  $Y$  by

$$Y(\omega) = \max\{X_1(\omega), \dots, X_n(\omega)\} .$$

As we saw in Sec. 12,  $Y$  is also a random variable, and

$$\{\omega : Y(\omega) \leq y\} = \{\omega : X_1(\omega) \leq y, \dots, X_n(\omega) \leq y\} .$$

Thus

$$F_Y(y) = F_{X_1 \dots X_n}(y, \dots, y) ,$$

and by invoking independence, we obtain

$$F_Y(y) = F_{X_1}(y) \cdots F_{X_n}(y) .$$

(If the variables are also identically distributed with common cdf  $F_X$ , then  $F_Y = (F_X)^n$ .)

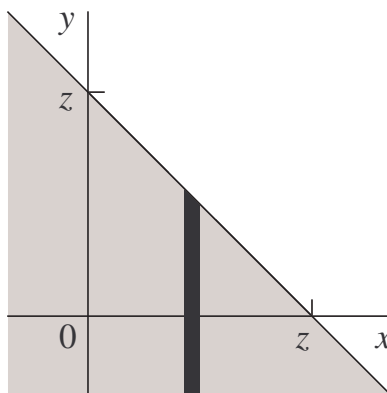
2. *Convolution.* Let  $X$  and  $Y$  be independent with absolutely continuous distributions. We are interested in computing the distribution of

$$Z = X + Y .$$

We have

$$F_Z(z) = P_{XY}\{(x, y) : x + y \leq z\} = \int \int_{x+y \leq z} f_{XY}(x, y) dx dy .$$

The region of integration is shaded in the figure below.



Using independence, we obtain

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{\infty} f_X(x) \left\{ \int_{-\infty}^{z-x} f_Y(y) dy \right\} dx \\ (y = t - x) \quad &= \int_{-\infty}^{\infty} f_X(x) \left\{ \int_{-\infty}^z f_Y(t - x) dt \right\} dx . \\ &= \int_{-\infty}^z \left\{ \int_{-\infty}^{\infty} f_X(x) f_Y(t - x) dx \right\} dt \end{aligned}$$

Thus  $F_Z$  is absolutely continuous with density  $f_Z$  given by the *convolution* of  $f_X$  and  $f_Y$ :

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x) dx ,$$

or  $f_Z = f_X * f_Y$ .

## 16. Expectation

Billingsley, Sections 15, 16, 17; Sec. 21, *Expected Value as Integral, Expected Values and Distributions, Independence and Expected Value.*

### Expectation as integral over $(\Omega, \mathcal{F}, P)$

The expectation of a *discrete nonnegative* random variable  $X$  is the weighted average of the countably many values taken by  $X$ , computed using the pmf of  $X$  as weighting function. Thus if  $X$  takes nonnegative values  $c_1, c_2, \dots$  with probabilities  $p_i = P_X\{c_i\}$ , the expectation of  $X$  is given by

$$EX \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} p_i c_i = \sum_{x: P_X\{x\} > 0} x P_X\{x\} .$$

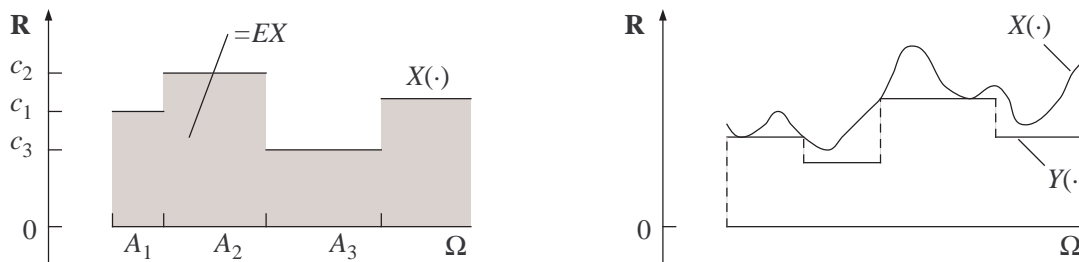
If  $X$  is defined on some probability space  $(\Omega, \mathcal{F}, P)$ , it is possible to express  $X$  in terms of the countable partition  $A_1, A_2, \dots$  that it induces:

$$X(\omega) = \sum_{i=1}^{\infty} c_i I_{A_i}(\omega) .$$

The expectation of  $X$  can then be rewritten as

$$EX = \sum_{i=1}^{\infty} c_i P(A_i).$$

We have often depicted the abstract sample space  $\Omega$  by an interval of the real line. If we go one step further and replace measure by length, then we can interpret the last expression for  $EX$  as the “area” under the graph of  $X(\omega)$ ; this is illustrated in the figure on the left below.



Note also that “area” and “average height” are identical in the above context, since the base is of unit “length.”

In order to extend the notion of expectation, or “area,” to an arbitrary nonnegative random variable  $X$ , we use a discrete approximation: the “area” under  $X(\omega)$  is the least upper bound to the “areas” of all discrete random variables that we can fit under  $X(\omega)$ ; i.e.,

$$EX = \sup\{EY : 0 \leq Y \leq X, Y \text{ discrete}\} .$$

If the  $EY$ 's are unbounded, then  $EX$  takes the value  $+\infty$ . This definition is illustrated in the previous figure, on the right. We also say that  $EX$  equals the **Lebesgue integral** of  $X$  with respect to  $P$ , and we write

$$EX = \int X(\omega) dP(\omega) .$$

The above equation emphasizes the fact that  $EX$  is defined through discrete approximations of  $X$  on the probability space  $(\Omega, \mathcal{F}, P)$ : expectation is an integral, or “area,” under the graph of a measurable function, evaluated using a probability measure in lieu of length on the horizontal axis. As we shall soon see, it is possible to express  $EX$  in terms of the distribution  $P_X$ ; for the moment, however, we can only do this for discrete random variables.

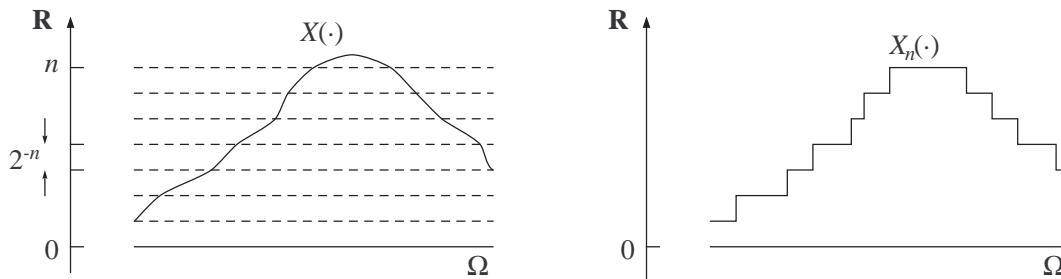
It turns out that for every nonnegative random variable  $X$  there exists a systematic approximation of  $X$  by a nondecreasing sequence of random variables

$$X_1 \leq X_2 \leq \dots \leq X$$

such that

$$EX = \sup_n EX_n = \lim_n EX_n .$$

This approximation is obtained by successive truncations and quantizations of  $X$  as follows. For every  $n$ , we restrict the range space from  $\mathbf{R}$  to  $[0, n]$ , and we partition  $[0, n]$  into  $2^n$  adjoining intervals, each of length  $2^{-n}$ . If  $X(\omega)$  falls in one of these intervals, we set  $X_n(\omega)$  equal to the left endpoint of that interval; otherwise (if  $X(\omega) \geq n$ ) we set  $X_n(\omega)$  equal to  $n$ . This is illustrated in the figure below.



In short, we write

$$X_n(\omega) = \sum_{k=0}^{M(n)} c_k^{(n)} I_{A_k^{(n)}}(\omega) ,$$

where  $M(n) = 2^n$ ;  $c_k^{(n)} = k2^{-n}$  if  $k \leq M(n)$ ,  $c_{M(n)+1}^{(n)} = +\infty$ ; and

$$A_k^{(n)} = X^{-1} \left[ c_k^{(n)}, c_{k+1}^{(n)} \right) .$$

It can be shown (see Billingsley for details) that  $EX = \lim_n EX_n$ , and thus

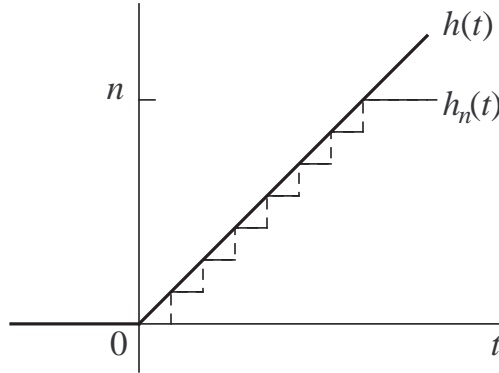
$$EX = \int X(\omega) dP(\omega) = \lim_n \sum_{k=0}^{M(n)} c_k^{(n)} P \left( A_k^{(n)} \right) .$$

### Expectation as integral over $(\mathbf{R}, \mathcal{B}(\mathbf{R}), P_X)$

The above approximation is important in that it allows us to express the expectation of  $X$  as the Lebesgue integral of a measurable function  $h \geq 0$  on the space  $(\mathbf{R}, \mathcal{B}(\mathbf{R}), P_X)$ . The function  $h$  is the ramp function

$$h(t) = tI_{(0, \infty)}(t) ,$$

and can be approximated by piecewise constant functions in the same fashion as  $X$ .



More precisely,

$$h_n(t) = \sum_{k=0}^{M(n)} c_k^{(n)} I_{H_k^{(n)}}(t) ,$$

where  $M(n)$  and  $c_k^{(n)}$  are defined as before, and

$$H_k^{(n)} = h^{-1} \left[ c_k^{(n)}, c_{k+1}^{(n)} \right) = \left[ c_k^{(n)}, c_{k+1}^{(n)} \right) .$$

To see why  $EX$  is the integral of  $h$  with respect to  $P_X$ , we write

$$\begin{aligned} \int h(t) dP_X(t) &= \lim_n \int h_n(t) dP_X(t) \\ &= \lim_n \sum_{k=0}^{M(n)} c_k^{(n)} P_X \left[ c_k^{(n)}, c_{k+1}^{(n)} \right) \\ &= \lim_n \sum_{k=0}^{M(n)} c_k^{(n)} P \left( A_k^{(n)} \right) = \int X(\omega) dP(\omega) = EX . \end{aligned}$$

We conclude that the expected value of a random variable  $X$  on  $(\Omega, \mathcal{F}, P)$  can be expressed solely in terms of the distribution of  $X$ :

$$EX = \int t I_{(0, \infty)}(t) dP_X(t) = \int_{(0, \infty)} t dP_X(t) .$$

(**Notation.** A subscript  $A$  in the Lebesgue integral sign denotes multiplication of the integrand by the indicator function of the measurable set  $A$ .)

In most cases of interest, evaluation of the above integral is quite straightforward. For discrete distributions  $P_X$ , we have already seen that

$$EX = \int_{(0, \infty)} t dP_X(t) = \sum_{x: P_X\{x\} > 0} x P_X\{x\} .$$

For absolutely continuous  $P_X$  with piecewise continuous density  $f_X$ , it can be shown using the approximation  $(h_n)_{n \in \mathbf{N}}$  that the Lebesgue integral with respect to  $P_X$  is given by Riemann integral as follows:

$$EX = \int_{(0, \infty)} t dP_X(t) = \int_0^\infty t f_X(t) dt .$$

(Since  $X$  is nonnegative, we may replace the lower limit in the Riemann integral by  $-\infty$ . For the present, we must keep the subscript  $(0, \infty)$  in the Lebesgue integral, as we have not yet defined this integral for functions that take both positive and negative values.)

For distributions that are mixtures of the above two types, we use the following simple relationship: if  $P_X = \lambda P_1 + (1 - \lambda) P_2$  for  $0 \leq \lambda \leq 1$ , and  $g$  is any nonnegative measurable function, then

$$\int g(t) dP_X(t) = \lambda \int g(t) dP_1(t) + (1 - \lambda) \int g(t) dP_2(t) .$$

## Expectations of functions of random variables

We now turn to expectations of nonnegative functions of random variables. Let  $X$  be an arbitrary random variable (not necessarily nonnegative) on  $(\Omega, \mathcal{F}, P)$ , and let  $g : \mathbf{R} \mapsto \mathbf{R}$  be a nonnegative measurable function on  $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$ . As we saw in section 13, the function

$$(g \circ X)(\omega) = g(X(\omega))$$

is a nonnegative random variable on  $(\Omega, \mathcal{F}, P)$ . By the result of the previous subsection, we can write

$$E[g \circ X] = E[g(X)] = \int g(X(\omega)) dP(\omega) = \int_{(0, \infty)} t dP_{g \circ X}(t). \quad (1)$$

Consider now the function  $g$  as a *random variable* on the probability space  $(\mathbf{R}, \mathcal{B}(\mathbf{R}), P_X)$ . The probability of a Borel set  $H$  under the distribution of  $g$  is given by

$$P_X(g^{-1}(H)) = P(X^{-1}(g^{-1}(H))) = P((g \circ X)^{-1}(H)) = P_{g \circ X}(H).$$

Thus the r.v.  $g$  on  $(\mathbf{R}, \mathcal{B}(\mathbf{R}), P_X)$  has the same distribution as the r.v.  $g \circ X$  on  $(\Omega, \mathcal{F}, P)$ . By the result of the previous subsection, we have

$$\int g(t) dP_X(t) = \int_{(0, \infty)} t dP_{g \circ X}(t). \quad (2)$$

Combining (1) and (2), we obtain

$$E[g(X)] = E[g \circ X] = \int_{(0, \infty)} t dP_X(t) = \int g(t) dP_X(t).$$

The above result is extremely useful in that it allows us to compute the expectation of any nonnegative function  $g$  of a random variable  $X$  in terms of the distribution of  $X$ ; thus the distribution of  $g \circ X$  is not needed in order to evaluate  $E[g(X)]$ .

For  $P_X$  discrete, it is easy to show that

$$E[g(X)] = \sum_{x: P_X\{x\} > 0} g(x) P\{x\}.$$

If  $P_X$  is absolutely continuous with piecewise continuous density  $f$ , and  $g$  itself is piecewise continuous, we can again use a Riemann integral to evaluate  $E[g(X)]$ :

$$E[g(X)] = \int g(t) dP_X(t) = \int_0^\infty g(t) f_X(t) dt.$$

**Example** If  $Y = X^2$ , then

$$EY = \int_{(0, \infty)} t dP_Y(t) = \int t^2 dP_X(t).$$

If  $P_X$  is absolutely continuous with piecewise continuous density, then the same is true of  $P_Y$ , and

$$EY = \int_0^\infty t f_Y(t) d(t) = \int_0^\infty t^2 f_X(t) dt.$$

It is both interesting and useful to note that the proof of the above result is also valid for a random vector  $\mathbf{X}$  replacing the random variable  $X$ . Thus if  $g : \mathbf{R}^k \mapsto \mathbf{R}$  is a nonnegative measurable function on  $(\mathbf{R}^k, \mathcal{B}(\mathbf{R}^k))$ , we can write

$$E[g(\mathbf{X})] = \int g(\mathbf{t}) dP_{\mathbf{X}}(\mathbf{t}) .$$

In the discrete case, evaluation of the above is again straightforward. In the absolutely continuous case, if we assume piecewise continuity of  $f_{\mathbf{X}}$  and  $g$  on  $\mathbf{R}^k$ , we have

$$E[g(\mathbf{X})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(t_1, \dots, t_k) f_{\mathbf{X}}(t_1, \dots, t_k) dt_1 \cdots dt_k .$$

### Expectation of arbitrary random variables

If  $X$  is an arbitrary (i.e., not necessarily nonnegative) random variable on  $(\Omega, \mathcal{F}, P)$ , the expectation of  $X$  is defined through a decomposition of  $X$  into *positive* part  $X^+$  and a *negative* part  $X^-$ :

$$X^+(\omega) \stackrel{\text{def}}{=} X(\omega)I_{\{\omega: X(\omega) > 0\}}(\omega) , \quad X^-(\omega) \stackrel{\text{def}}{=} -X(\omega)I_{\{\omega: X(\omega) < 0\}}(\omega) .$$

Thus  $X^+$  and  $X^-$  are nonnegative random variables such that

$$X = X^+ - X^- , \quad |X| = X^+ + X^- .$$

Furthermore,  $EX^+$  and  $EX^-$  are well-defined by the foregoing discussion.

**Definition.** The expectation of  $X$  is defined by

$$EX \stackrel{\text{def}}{=} EX^+ - EX^-$$

if *both*  $EX^+$  and  $EX^-$  are *finite*, or if *at most one* of  $EX^+$  and  $EX^-$  is *infinite*. If *both*  $EX^+$  and  $EX^-$  are *infinite*, then  $EX$  is *not defined* (i.e., it does not exist).

If  $EX$  exists, we can also express it as a Lebesgue integral:

$$EX = \int X(\omega) dP(\omega) = \int X^+(\omega) dP(\omega) - \int X^-(\omega) dP(\omega) .$$

If both  $EX^+$  and  $EX^-$  are finite, we say that  $X$  is *integrable*; this is also equivalent to saying that  $E|X| < \infty$ , or  $EX \neq \pm\infty$ . If  $E|X|$  is infinite (in which case  $EX$  is itself infinite or does not exist), we say that  $X$  is *non-integrable*.

**Example.** Consider the probability space consisting of the interval  $[-\pi/2, \pi/2]$ , its Borel field, and the measure  $P$  that has uniform density over  $[-\pi/2, \pi/2]$  ( $P$  is a scaled version of the Lebesgue measure). Define the random variables  $X$  and  $Y$  by

$$X(\omega) = \sin \omega , \quad Y(\omega) = \tan \omega .$$

We have

$$EX^+ = \int_{(0, \pi/2)} \sin \omega \, dP(\omega) = \frac{1}{\pi} \int_0^{\pi/2} \sin \omega \, d\omega = \frac{1}{\pi} ,$$

$$EY^+ = \int_{(0, \pi/2)} \tan \omega \, dP(\omega) = \frac{1}{\pi} \int_0^{\pi/2} \tan \omega \, d\omega = +\infty .$$

By symmetry,  $EX^- = EX^+$  and  $EY^- = EY^+$ . Thus  $EX = 0$ , while  $EY$  does not exist.

### Further properties of expectation

(a) *Functions of random variables.* If  $g : \mathbf{R} \mapsto \mathbf{R}$  is Borel measurable (but not necessarily nonnegative) and  $E[g(X)]$  exists, then we can express  $E[g(X)]$  in terms of  $P_X$  as before. To see this, observe that

$$(g \circ X)^+(\omega) = g^+(X(\omega)) , = (g \circ X)^-(\omega) = g^-(X(\omega)) ,$$

and hence by our earlier result,

$$E[(g \circ X)^+] = E[g^+(X)] = \int g^+(t) \, dP_X(t) ,$$

$$E[(g \circ X)^-] = E[g^-(X)] = \int g^-(t) \, dP_X(t) .$$

Our latest extension of the definition of expectation yields

$$E[g(X)] = E[g \circ X] = \int g(t) \, dP_X(t) ,$$

whenever either side of the equation exists. By taking  $g(t) = t$ , we obtain the important relationship

$$EX = \int g(t) \, dP_X(t) .$$

Once again, the Lebesgue integrals in the above relationships are usually evaluated by sums (when  $P_X$  is discrete) or Riemann integrals (when  $P_X$  is absolutely continuous with piecewise continuous density, and  $g$  is itself piecewise continuous). One **must**, however, exercise care in using such sums and integrals; as a rule, one should compute  $E[g^+(X)]$  and  $E[g^-(X)]$  separately before giving the final answer. For instance, if  $g(t) = t$  and  $P_X$  is absolutely continuous, one would need to evaluate both

$$EX^+ = \int_0^\infty t f_X(t) \, dt \quad \text{and} \quad EX^- = - \int_{-\infty}^0 t f_X(t) \, dt .$$

**Example** The random variable  $Y$  of the previous example has the Cauchy density  $f_Y(t) = \pi^{-1}(1+t^2)^{-1}$ . It is tempting to write

$$\int_{-\infty}^\infty t f_Y(t) \, dt = 0$$

by virtue of the fact that the integrand is an odd function of  $t$ . Yet as we have seen,  $EY$  does not exist, and restriction of the above integral to  $(0, \infty)$  and  $(-\infty, 0)$  yields the correct results  $EY^+ = \infty$ , and  $EY^- = \infty$ , respectively. The same error could have been made in the evaluation of  $EY$  in the previous example, as  $Y(\omega)$  is an odd function of  $\omega$ .

As in the previous section, we can extend the above results to measurable functions of random vectors. Again one should exercise care in handling multiple integrals such as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(t_1, \dots, t_k) f_{\mathbf{X}}(t_1, \dots, t_k) dt_1 \cdots dt_k .$$

As a rule, one should first decompose  $g$  into positive and negative parts, and then evaluate the two resulting integrals in the usual iterated fashion. One important exception is the case where integrability of  $g$  can be independently established (e.g. by evaluating the above integral with  $|g|$  replacing  $g$ ); then decomposition into positive and negative parts is not necessary.

(b) *Linearity.* If  $X$  and  $Y$  are integrable random variables and  $\alpha, \beta$  are real constants, then

$$E[\alpha X + \beta Y] = \alpha EX + \beta EY .$$

This is easily shown using linearity of sums and integrals for cases in which the pair  $(X, Y)$  is discrete or absolutely continuous with “nice” density. In the latter case, for instance, we have

$$\begin{aligned} E[\alpha X + \beta Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\alpha x + \beta y) f_{XY}(x, y) dx dy \\ &= \alpha \int_{-\infty}^{\infty} x dx \int_{-\infty}^{\infty} f_{XY}(x, y) dy + \beta \int_{-\infty}^{\infty} y dy \int_{-\infty}^{\infty} f_{XY}(x, y) dx \\ &= \alpha \int_{-\infty}^{\infty} x f_X(x) dx + \beta \int_{-\infty}^{\infty} y f_Y(y) dy = \alpha EX + \beta EY . \end{aligned}$$

(Note that the assumption of integrability of  $X$  and  $Y$  enabled us to decompose the integrals into components without having to worry about indeterminate forms such as  $+\infty - \infty$ .)

The proof of linearity in the general case appears in Billingsley.

(c) *Independence and Expectation.* If  $X_1, \dots, X_n$  are independent and integrable random variables, then the expectation of their product is given by the product of their expectations:

$$E[X_1 \cdots X_n] = E[X_1] \cdots E[X_n] .$$

The proof of this fact for arbitrary random variables is given in Billingsley. For the case of absolutely continuous distributions with piecewise continuous density, we first show the above result for nonnegative independent (but not necessarily integrable) variables  $Y_1, \dots, Y_n$ :

$$\begin{aligned}
E[Y_1 \cdots Y_n] &= \int_0^\infty \cdots \int_0^\infty t_1 \cdots t_n f_{Y_1}(t_1) \cdots f_{Y_n}(t_n) dt_1 \cdots dt_n \\
&= \left( \int_0^\infty t_1 f_{Y_1}(t_1) dt_1 \right) \cdots \left( \int_0^\infty t_n f_{Y_n}(t_n) dt_n \right) \\
&= E[Y_1] \cdots E[Y_n] .
\end{aligned}$$

Thus if  $X_1, \dots, X_n$  are integrable, we can utilize the above identity with  $Y_i = |X_i|$  to conclude that the product  $X_1 \cdots X_n$  is also integrable. By repeating the steps with  $X_i$  replacing  $Y_i$  for every  $i$  and  $-\infty$  replacing 0 in all lower limits, we obtain

$$E[X_1 \cdots X_n] = E[X_1] \cdots E[X_n] .$$

**Example.** The assumption of integrability was essential in establishing the above product relationship. Consider a counterexample in which  $X$  and  $Y \geq 0$  are independent with

$$f_X(t) = \frac{1}{2} I_{[-1,1]}(t) , \quad f_Y(t) = \frac{2}{\pi(1+t^2)} I_{[0,\infty)}(t) .$$

In this case  $E[X^+] = E[X^-] = 1/2$ , while  $EY = +\infty$ . We have

$$E[(XY)^+] = E[(XY)^-] = (1/2)(+\infty) = +\infty ,$$

and thus  $E[XY]$  does not exist. On the other hand,  $E[X]E[Y]$  equals  $0(+\infty)$  which is by convention equal to zero. The same error can be made by setting  $E[XY]$  equal to the iterated integral

$$\int_{-\infty}^{\infty} y f_Y(y) \left\{ \int_{-\infty}^{\infty} x f_X(x) dx \right\} dy ,$$

which is equal to zero.

## 17. Applications of expectation

Billingsley, Sec. 21, *Applications of expectation*.

### Moments and Correlation

For positive integer  $k$ , the  $k^{\text{th}}$  **moment** of a random variable  $X$  is the expectation of  $X^k$  (provided it exists).

In what follows we will assume random variables  $X$  and  $Y$  are **square-integrable**, i.e., they have finite second moments. This also implies that  $X$  and  $Y$  also have finite expectations (or first moments).

The **correlation** of  $X$  and  $Y$  is the expectation of the product  $XY$ . This exists and is finite by virtue of the fact that  $|XY| \leq X^2 + Y^2$  and the assumption that  $X$  and  $Y$  are square-integrable.

The **covariance** of  $X$  and  $Y$  is the correlation of  $X - EX$  and  $Y - EY$ , i.e., the correlation of the original random variables *centered* at their expectations:

$$\text{Cov}(X, Y) \stackrel{\text{def}}{=} E[(X - EX)(Y - EY)] .$$

We also have

$$\text{Cov}(X, Y) = E[XY - X(EY) - Y(EX) + (EX)(EY)] = E[XY] - E[X]E[Y] .$$

If we let  $X = Y$ , then we obtain the **variance** of  $X$ :

$$\text{Var}X \stackrel{\text{def}}{=} \text{Cov}(X, X) = E(X - EX)^2 = EX^2 - (EX)^2 .$$

Since the variance equals the expectation of the nonnegative random variable  $(X - EX)^2$ , it is always nonnegative; from the above we also deduce that the second moment of  $X$  is always greater than or equal to the square of its first moment.

We say that  $X$  and  $Y$  are **uncorrelated** if  $\text{Cov}(X, Y) = 0$ , or equivalently, if  $E[XY] = E[X]E[Y]$ . Thus two independent random variables are always uncorrelated; the converse is **not** true, as the following example shows.

**Example.** Let  $X$  be such that  $EX = 0$  and  $EX^3 \neq 0$ . Then for  $\alpha, \beta \in \mathbf{R}$ , the random variable

$$Y = \alpha X^2 + \beta X$$

is clearly not independent of  $X$ . Yet

$$\text{Cov}(X, Y) = E[XY] = \alpha EX^3 + \beta EX^2 ,$$

which can be zero for suitable  $\alpha$  and  $\beta$ .

Thus uncorrelatedness is a much weaker condition than independence. It should be emphasized that the product relationship by which uncorrelatedness is defined always involves **two** random variables. Thus when speaking of a collection of uncorrelated variables, we understand that every two variables in the collection satisfy the said relationship; we do **not** imply the validity of product relationships involving three or more variables at once. This is in clear contrast to the case of independence. Another salient difference between uncorrelatedness and independence is that the former is not preserved under measurable transformations of the variables involved, whereas the latter is.

Using the definition of covariance and the linearity of expectation, it is easy to show that for square integrable variables  $X_1, \dots, X_M$  and  $Y_1, \dots, Y_N$ ,

$$\text{Cov}\left(\sum_{i=1}^M a_i X_i + b_i, \sum_{j=1}^N c_j Y_j + d_j\right) = \sum_{i=1}^M \sum_{j=1}^N a_i c_j \text{Cov}(X_i, Y_j) .$$

In particular, covariance is not affected by additive constants.

## The Cauchy-Schwarz inequality

The Cauchy-Schwarz inequality states that for square-integrable random variables  $X$  and  $Y$ ,

$$(E[XY])^2 \leq E[X^2]E[Y^2] ,$$

with equality if and only if there exists a constant  $\lambda$  such that  $X + \lambda Y = 0$  with probability one.

To prove this, consider the nonnegative random variable  $Z_\lambda$  defined by

$$Z_\lambda = (X + \lambda Y)^2 .$$

Here  $\lambda$  is a (nonrandom) constant. We have

$$\begin{aligned} EZ_\lambda &= E[X^2 + 2\lambda XY + \lambda^2 Y^2] \\ &= E[X^2] + 2\lambda E[XY] + \lambda^2 E[Y^2] \geq 0 , \end{aligned}$$

where the last inequality follows from the fact that  $Z_\lambda \geq 0$ . Since the inequality holds for all  $\lambda$ , we have (by the elementary theory of quadratics) that

$$(2E[XY])^2 \leq 4E[X^2]E[Y^2] , \quad \text{or} \quad (E[XY])^2 \leq E[X^2]E[Y^2] .$$

This establishes the sought inequality. Equality will hold if and only if the quadratic has exactly one real root (of multiplicity two), i.e., there exists a unique  $\lambda$  such that

$$EZ_\lambda = 0 .$$

Since  $Z_\lambda \geq 0$ , the above statement is equivalent to

$$Z_\lambda = 0 , \quad \text{or} \quad X + \lambda Y = 0$$

with probability one.

**Remark.** It is clear that if  $Z_\lambda = 0$  with probability one, then  $EZ_\lambda = 0$ . To see why  $EZ_\lambda > 0$  if the inequality  $Z_\lambda > 0$  is true with positive probability, note that the event over which  $Z_\lambda > 0$  holds can be decomposed into events over which  $r^{-1} \leq Z < (r-1)^{-1}$  for  $r \in \mathbf{N}$ . If  $Z_\lambda > 0$  with positive probability, one of the above disjoint events has positive probability, and thus  $Z_\lambda$  is lower bounded by a simple random variable of positive expectation.

A corollary of the Cauchy-Schwarz inequality is obtained by centering the random variables at their expectations:

$$\left( \text{Cov}(X, Y) \right)^2 \leq \text{Var}X \cdot \text{Var}Y ,$$

with equality if and only if there exist constants  $\lambda$  and  $c$  such that  $X + \lambda Y = c$  with probability one.

## The Markov and Chebyshev inequalities

**Notation.** Where there is no ambiguity about the choice of probability space, we will use “Pr” as an abbreviation for “probability that.” For example, on a probability space  $(\Omega, \mathcal{F}, P)$ ,

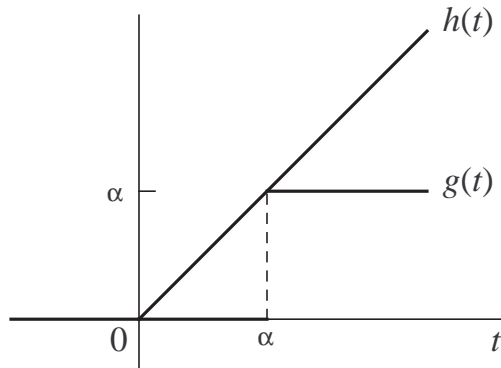
$$\Pr\{\text{infinitely many } A_i\text{'s occur}\} = P(\limsup_i A_i) ,$$

$$\Pr\{X \in H\} = P\{\omega : X(\omega) \in H\} = P_X(\omega) .$$

Consider a nonnegative random variable  $U$ . For a positive constant  $\alpha$ , the Markov inequality provides an upper bound to  $\Pr\{U \geq \alpha\}$  in terms of the expectation  $EU$ . To derive this inequality, we consider the functions

$$g(t) = \alpha I_{[\alpha, \infty)}(t) \quad \text{and} \quad h(t) = t I_{[0, \infty)}(t)$$

defined on the real line.



Since

$$g(t) \leq h(t) ,$$

it is also true that

$$\int g(t) dP_U(t) \leq \int h(t) dP_U(t) .$$

or equivalently,

$$\alpha P_U[\alpha, \infty) \leq EU .$$

Thus we have obtained the **Markov inequality**:

$$\Pr\{U \geq \alpha\} \leq \frac{EU}{\alpha} .$$

The Markov inequality can be applied to an arbitrary r.v.  $X$  in order to obtain an upper bound to  $\Pr\{|X| \geq \alpha\}$  in terms of the  $r^{\text{th}}$  moment of  $|X|$ . Indeed, if we set  $U = |X|$ , then

$$\Pr\{|X| \geq \alpha\} = \Pr\{|X|^r \geq \alpha^r\} \leq \frac{E|X|^r}{\alpha^r} .$$

For  $r = 2$ ,  $U = |X|$  and  $|X - EX|$ , we obtain two versions of the **Chebyshev inequality**:

$$\Pr\{|X| \geq \alpha\} \leq \frac{EX^2}{\alpha^2};$$

$$\Pr\{|X - EX| \geq \alpha\} \leq \frac{\text{Var}X}{\alpha^2}.$$

**Remark.** Except for trivial cases, the Markov and Chebyshev inequalities are also valid with strict inequality signs *on both sides* of the appropriate expressions. This can be seen by taking  $g = I_{(\alpha, \infty)}$  in the above.

As an application of the Chebyshev inequality, consider a sequence  $X_1, X_2, \dots$  of square-integrable uncorrelated random variables such that  $EX_i = \mu_i$ ,  $\text{Var}X_i = \sigma_i^2$ .

We are interested in the behavior of the *sample* (or *time*) average

$$\frac{X_1 + \dots + X_n}{n}$$

of the first  $n$  variables. Clearly

$$E \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n EX_i = \frac{1}{n} \sum_{i=1}^n \mu_i$$

and

$$\text{Var} \left( \frac{1}{n} \sum_{i=1}^n X_i \right) = \text{Cov} \left( \frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{j=1}^n X_j \right) = \sum_{i=1}^n \sum_{j=1}^n \frac{1}{n^2} \text{Cov}(X_i, X_j)$$

(by uncorrelatedness)  $= \frac{1}{n^2} \sum_{i=1}^n \text{Cov}(X_i, X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2.$

Thus by the Chebyshev inequality,

$$\Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n \mu_i \right| \geq \alpha \right\} \leq \frac{1}{n^2 \alpha^2} \sum_{i=1}^n \sigma_i^2.$$

In particular, if  $\mu_i = \mu$ ,  $\sigma_i^2 = \sigma^2$ , then

$$\Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \alpha \right\} \leq \frac{1}{n^2 \alpha^2} (n \sigma^2) = \frac{\sigma^2}{n \alpha^2}.$$

Thus regardless of the choice of  $\alpha \geq 0$ , we have

$$\lim_n \Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \alpha \right\} = 0.$$

As we shall see in the following section, the above statement can be read as “the sample average converges to the expectation (mean)  $\mu$  in probability.”

## 18. Convergence of sequences of random variables

Billingsley, Sec. 20, *Convergence in Probability*; Sec. 25, *Convergence in Distribution*, *Convergence in Probability*.

The notion of convergence of a sequence of random variables was introduced in Section 12. As we saw in that section, if  $X_1, X_2, \dots$  are random variables on the same probability space  $(\Omega, \mathcal{F}, P)$ , then

- (i) the set  $C$  of  $\omega$ 's for which  $X(\omega)$  converges to a finite or infinite limit is an event (i.e.,  $C \in \mathcal{F}$ );
- (ii) the mapping that carries every  $\omega$  in  $C$  to the corresponding limit defines an extended random variable on the restriction of  $(\Omega, \mathcal{F})$  on  $C$ .

In this section we will discuss different modes in which a sequence of random variables  $X_1, X_2, \dots$  may converge to a limiting random variable  $X$ . Some modes of convergence are defined in the framework of the previous paragraph, i.e. in regard to the limiting behavior of  $X_n(\omega)$  as  $\omega$  varies over a suitable event  $C$ . Other modes are defined in an *aggregate* sense, and thus do not entail the convergence of individual sequences  $X_1(\omega), X_2(\omega), \dots$

**Remark.** It is important to appreciate the difference between two expressions such as

$$X_1(\omega), X_2(\omega), \dots$$

and

$$X_1, X_2, \dots$$

For a given  $\omega$ , the former is a sequence of numbers; the latter is a sequence of functions on  $\Omega$ . Thus there is nothing equivocal about the statement “ $X_n(\omega)$  converges;” on the other hand, as we shall soon see, “ $X_n$  converges” could mean one of several things.

In what follows, all random variables are defined on the same probability space and are *finite-valued* (i.e., they are not extended random variables).

### Definitions

(a) *Pointwise convergence on  $\Omega$ .* The sequence  $(X_n)_{n \in \mathbf{N}}$  converges to  $X$  *pointwise on  $\Omega$*  if

$$(\forall \omega \in \Omega) \quad \lim_n X_n(\omega) = X(\omega) .$$

Thus in this case, the set  $C$  introduced above coincides with  $\Omega$ .

(a) *Almost sure convergence.* The sequence  $(X_n)_{n \in \mathbf{N}}$  converges to  $X$  *almost surely* if the event  $A$  over which

$$\lim_n X_n(\omega) = X(\omega)$$

has probability one. In this case, the set  $C$  introduced above contains  $A$  as a subset, and thus  $P(C) = 1$ . We write  $X_n \xrightarrow{a.s.} X$ .

It is instructive to derive an alternative representation of the event

$$A = \{\omega : \lim_n X_n(\omega) = X(\omega)\} .$$

To this end we recall the definition of convergence of a sequence  $(a_n)_{n \in \mathbf{N}}$  to a finite limit  $a$ :

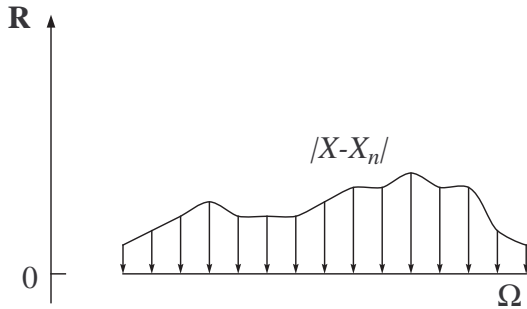
$$\lim_n a_n = a \iff (\forall r)(\exists m)(\forall n \geq m) |a - a_n| < r^{-1} .$$

Here  $m$ ,  $n$ , and  $q$  are taken to range over  $\mathbf{N}$ . We also note that in Section 12 we gave an alternative definition of convergence of real sequences in terms of iterated infima and suprema. The one given directly above is more elementary, and leads to the following representation for  $A$ :

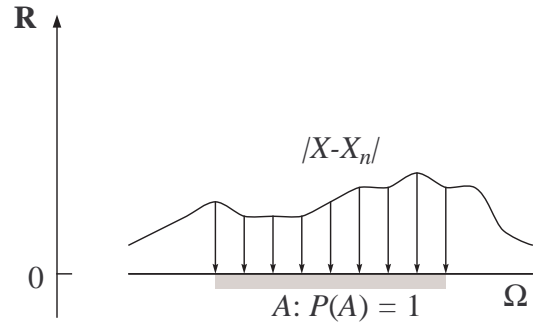
$$\begin{aligned} A &= \left\{ (\omega) : (\forall q)(\exists m)(\forall n \geq m) |X(\omega) - X_n(\omega)| < q^{-1} \right\} \\ &= \bigcap_q \bigcup_m \bigcap_{n \geq m} \{\omega : |X(\omega) - X_n(\omega)| < q^{-1}\} . \end{aligned}$$

Thus the event  $A$  over which  $X_n(\omega)$  converges to  $X(\omega)$  is the set of  $\omega$ 's with the property that for every  $q$ , the distance  $|X(\omega) - X_n(\omega)|$  is *eventually* smaller than  $1/q$ . In Section 7 we linked the qualifier “eventually” with the limit inferior of a sequence of events; this connection is again apparent in the last expression, as each set in the first intersection (over  $q$ ) is a limit inferior of a sequence of events.

Pointwise convergence on  $\Omega$  and almost sure convergence are illustrated in the figure below, where downward arrows denote convergence at the given abscissae.



Pointwise convergence on  $\Omega$



Almost sure convergence

(c) *Convergence in probability.* The sequence  $(X_n)_{n \in \mathbf{N}}$  converges to  $X$  in probability if for every  $\epsilon > 0$ ,

$$\lim_n \Pr\{|X - X_n| > \epsilon\} = 0 .$$

We write  $X_n \xrightarrow{P} X$ . It is easily seen that the  $>$  sign in the above expression can be replaced by  $\geq$ , and that  $\epsilon$  need only range over the reciprocals of positive integers. Thus for  $X_n \xrightarrow{P} X$

to be true, we require that for every  $q \in \mathbf{N}$ , the sequence of events  $B_{1,q}, B_{2,q}, \dots$  defined by

$$B_{n,q} = \{\omega : |X(\omega) - X_n(\omega)| \geq \epsilon\}$$

satisfy

$$\lim_n P(B_{n,q}) = 0 .$$

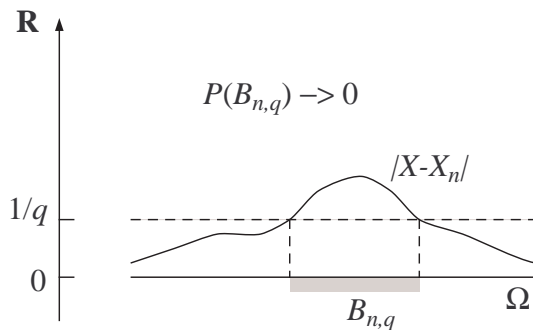
(d) *Convergence in  $r^{\text{th}}$  mean.* The sequence  $(X_n)_{n \in \mathbf{N}}$  converges to  $X$  in  $r^{\text{th}}$  mean if

$$\lim_n E[|X - X_n|^r] = 0 .$$

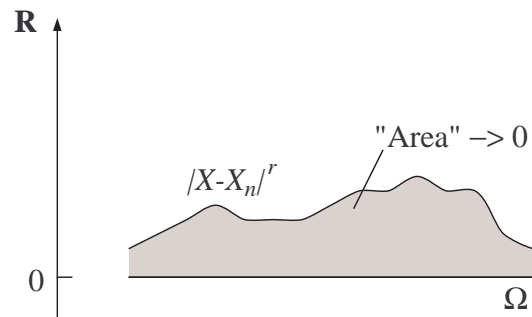
We write  $X_n \xrightarrow{L_r} X$ . In the special case  $r = 2$  we speak of convergence in *quadratic mean*, or *mean-square convergence*, which we also denote by

$$X \xrightarrow{q.m.} X \quad \text{and} \quad X \xrightarrow{m.s.} X .$$

Thus convergence in probability and in  $r^{\text{th}}$  mean differ from the first two modes of convergence discussed above in that their definition does not entail convergence of the sequence  $(X_n(\omega))_{n \in \mathbf{N}}$ . For convergence in probability, we require that certain sequences of *events* exhibit vanishing probabilities; for convergence in  $r^{\text{th}}$  mean, we require that a certain sequence of *integrals* exhibit vanishing values. This is illustrated in the figure below.



Convergence in probability



Convergence in  $r^{\text{th}}$  mean

(e) *Convergence in distribution.* The sequence  $(X_n)_{n \in \mathbf{N}}$  converges to  $X$  in *distribution* if at every continuity point  $x$  of  $F_X$  (i.e., where  $F_X\{x\} = 0$ ), the following is true:

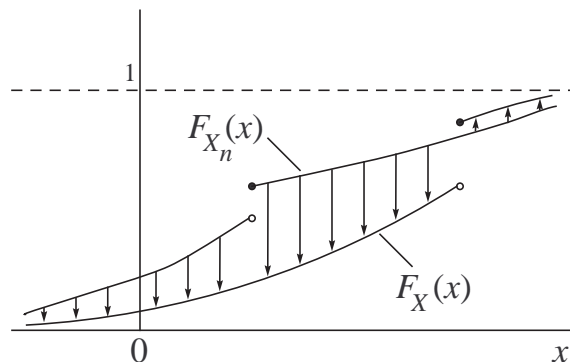
$$\lim_n F_{X_n}(x) = F_X(x) .$$

We write  $X_n \xrightarrow{d} X$ . As this definition only involves the (marginal) distributions  $F_X$  and (for all  $n$ )  $F_{X_n}$ , it is acceptable to identify convergence *in distribution* with convergence of distributions. Thus if we let  $F_n \equiv F_{X_n}$  and  $F \equiv F_X$ , we can also express  $X_n \xrightarrow{d} X$  as

$$F_{X_n} \xrightarrow{w} F_X \quad \text{or} \quad F_n \xrightarrow{w} F .$$

Here “w” stands for “weakly,” and is interpreted as convergence at all points of continuity.

The following figure illustrates convergence in distribution.



### Uniqueness of limits

In each of the five modes of convergence discussed above, the limiting random variable  $X$  is to a larger or smaller extent determined by the sequence  $(X_n)_{n \in \mathbf{N}}$ . More specifically, if both  $X$  and  $Y$  are limits (in the appropriate mode) of the sequence  $(X_n)_{n \in \mathbf{N}}$ , then

(a) under *pointwise* convergence,

$$(\forall \omega) X(\omega) = Y(\omega) , \quad \text{i.e.,} \quad X \equiv Y ;$$

(b) under *almost sure* convergence, convergence *in probability*, or convergence *in  $r^{\text{th}}$  mean*,

$$(\exists D \in \mathcal{F}) (\forall \omega \in D) \quad X(\omega) = Y(\omega) , \quad \text{i.e.,} \quad X = Y \text{ a.s. ;}$$

(c) under convergence *in distribution*,

$$F_X \equiv F_Y ,$$

and thus unlike other modes of convergence, the mappings  $X(\cdot)$  and  $Y(\cdot)$  here need not be related in any particular way.

The above facts are fairly easy to establish, and the proofs appear in the references for this section.

## Criteria for convergence

One is often interested in establishing convergence of  $(X_n)_{n \in \mathbf{N}}$  in one of the five modes discussed above without regard to the properties of the limiting random variable  $X$ . It is thus useful to have recourse to convergence criteria that do not explicitly involve  $X$ , especially in situations where a plausible candidate for  $X$  may be difficult to find. For the first four modes of convergence, such criteria exist and are expressed in terms of the behavior of the random variable

$$|X_m - X_n|$$

for large values of  $m$  and  $n$ .

The idea here stems from the concept of *mutual convergence* in real sequences: the sequence  $(a_n)_{n \in \mathbf{N}}$  converges mutually if

$$(\forall q \in \mathbf{N}) (\exists K) (\forall m, n \geq K) \quad |a_m - a_n| < q^{-1} .$$

We also write the above as

$$\lim_{m, n \rightarrow \infty} |a_m - a_n| = 0 ,$$

where the above limit is understood over *all* sequences of pairs  $(m, n)$  such that both  $m$  and  $n$  increase to infinity.

As it turns out, the sequence  $(a_n)_{n \in \mathbf{N}}$  converges to a finite limit if and only if it converges mutually. This equivalence can be readily exploited to yield the following criteria for pointwise and almost sure convergence:

(a)  $(X_n)_{n \in \mathbf{N}}$  converges pointwise (to some random variable) if and only if

$$(\forall \omega \in \Omega) \quad \lim_{m, n \rightarrow \infty} |X_m(\omega) - X_n(\omega)| = 0 ;$$

(b)  $(X_n)_{n \in \mathbf{N}}$  converges almost surely if and only if

$$P\left\{\omega : \lim_{m, n \rightarrow \infty} |X_m(\omega) - X_n(\omega)| = 0\right\} = 1 ;$$

For convergence in probability and  $r^{\text{th}}$  mean, the criteria are similar but rather more difficult to obtain:

(c)  $(X_n)_{n \in \mathbf{N}}$  converges in probability if and only if for every  $q \in \mathbf{N}$ ,

$$\lim_{m, n \rightarrow \infty} P\{\omega : |X_m(\omega) - X_n(\omega)| > q^{-1}\} = 0 ;$$

(d)  $(X_n)_{n \in \mathbf{N}}$  converges in  $r^{\text{th}}$  mean if and only if

$$\lim_{m, n \rightarrow \infty} E|X_m - X_n|^r = 0 .$$

Recall that  $X_n \xrightarrow{d} X$  entails convergence of the cdf of  $X_n$  to that of  $X$  at every continuity point of the latter cdf. Although it is tempting to think that a sufficient condition for convergence is the existence of

$$\lim_n F_{X_n}(x)$$

for all  $x$ , this is not so; the above limit should also coincide with a legitimate cdf at all points where the latter is continuous.

To see a simple example where the above limit exists but does not define a cdf, take

$$X_n \equiv n .$$

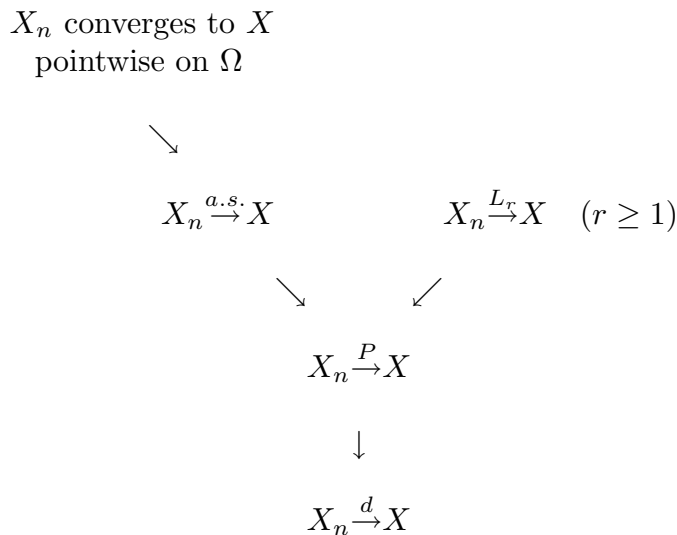
Then the cdf of  $X_n$  is the indicator function  $I_{[n, \infty)}$ , and for every  $x$ ,

$$\lim_n F_{X_n}(x) = 0 .$$

Clearly there exists no cdf  $F$  such that  $F = 0$  at all points of continuity of  $F$ , and thus  $(X_n)_{n \in \mathbf{N}}$  does not converge in distribution.

### Relationships between types of convergence

The following table summarizes the principal relationships between the five types of convergence defined in this section. Here, a *single* arrow denotes implication.



*Pointwise convergence on  $\Omega \Rightarrow$  Almost sure convergence:* Obvious by definition.

*Almost sure convergence  $\Rightarrow$  Convergence in probability:* Recall that the set of  $\omega$ 's for which  $X_n(\omega)$  converges to  $X(\omega)$  is given by

$$A = \bigcap_q \bigcup_m \bigcap_{n \geq m} \{\omega : |X(\omega) - X_n(\omega)| < q^{-1}\} ,$$

where  $q$ ,  $m$  and  $n$  range over the positive integers. If  $P(A) = 1$ , then for any given  $q$ ,

$$P\left(\bigcup_m \bigcap_{n \geq m} \{\omega : |X(\omega) - X_n(\omega)| < q^{-1}\}\right) = 1.$$

As the outer union is over an increasing sequence of events, we have

$$\lim_m P\left(\bigcap_{n \geq m} \{\omega : |X(\omega) - X_n(\omega)| < q^{-1}\}\right) = 1,$$

and thus also

$$\lim_m P\{\omega : |X(\omega) - X_m(\omega)| < q^{-1}\} = 1.$$

Upon complementation we obtain

$$\lim_m P\{\omega : |X(\omega) - X_m(\omega)| \geq q^{-1}\} = 0,$$

which proves that  $X_n \xrightarrow{P} X$ .

*Convergence in  $r^{\text{th}}$  mean  $\Rightarrow$  Convergence in probability:* By the Markov inequality, for every  $\epsilon > 0$  we have

$$\Pr\{|X - X_n| \geq \epsilon\} \leq \frac{E|X - X_n|^r}{\epsilon^r}.$$

If  $X_n \xrightarrow{L_r} X$ , then the r.h.s. approaches 0 as  $n$  tends to infinity, and thus

$$\lim_n \Pr\{|X - X_n| \geq \epsilon\} = 0.$$

*Convergence in probability  $\Rightarrow$  Convergence in distribution:* See Billingsley.

To see that no equivalences or other implications exist between the five modes of convergence, we consider some counterexamples.

(a) *Convergence in distribution  $\not\Rightarrow$  Convergence in probability:* Let  $\Omega = \{0, 1\}$  and  $P\{0\} = P\{1\} = 1/2$ . Also let

$$X_n(\omega) = \begin{cases} \omega, & \text{if } n \text{ is odd;} \\ 1 - \omega, & \text{if } n \text{ is even.} \end{cases}$$

The above sequence is identically distributed with

$$\Pr\{X_n = 0\} = \Pr\{X_n = 1\} = 1/2,$$

and thus trivially  $(X_n)_{n \in \mathbf{N}}$  converges in distribution. By the criterion of the previous subsection,  $(X_n)_{n \in \mathbf{N}}$  does not converge in probability, since for all  $n$ ,

$$\Pr\{|X_n - X_{n+1}| = 1\} = 1.$$

(b) *Convergence in Probability  $\not\Rightarrow$  Almost sure convergence or convergence in  $r^{\text{th}}$  mean:* Consider the probability space  $(\Omega, \mathcal{F}, P)$  where  $\Omega = (0, 1]$ ,  $\mathcal{F} = \mathcal{B}(\Omega)$ , and  $P$  is the Lebesgue measure. We construct a sequence of events  $(G_n)_{n \in \mathbf{N}}$  ranging over the dyadic intervals:

$$\begin{aligned} G_1 &= (0, 1] ; \\ G_2 &= (0, 1/2] , \quad G_3 = (1/2, 1] ; \\ G_4 &= (0, 1/4] , \quad G_5 = (1/4, 1/2] , \quad G_6 = (1/2, 3/4] , \quad G_7 = (3/4, 1] ; \\ &\text{etc.} \end{aligned}$$

If  $G_n = (a_n, b_n]$ , we define the simple random variable  $X_n$  by

$$X_n(\omega) = (b_n - a_n)^{-1} I_{G_n}(\omega) = P^{-1}(G_n) I_{G_n}(\omega) .$$

The sequence  $(X_n)_{n \in \mathbf{N}}$  converges to 0 in probability, since for every  $\epsilon > 0$ ,

$$P\{\omega : |X(\omega) - X_n(\omega)| > \epsilon\} \leq P(G_n) ,$$

and  $P(G_n)$  converges to zero by construction of the  $G_n$ 's. To see that it does not converge almost surely, first note that every  $\omega$  lies in infinitely many  $G_n$ 's and thus

$$X_n(\omega) = (b_n - a_n)^{-1} \geq 1$$

for infinitely many values of  $n$ . On the other hand, every  $\omega$  also lies outside infinitely many  $G_n$ 's and thus  $X_n(\omega) = 0$  infinitely often. Therefore for every  $\omega$ , the sequence  $(X_n(\omega))_{n \in \mathbf{N}}$  fails to converge, and certainly  $(X_n)_{n \in \mathbf{N}}$  does not converge almost surely.

To see that  $(X_n)_{n \in \mathbf{N}}$  does not converge in  $r^{\text{th}}$  mean (where  $r \geq 1$ ), note that for every  $K$ , there exist  $m, n > K$  such that  $G_m \cap G_n = \emptyset$ . Then

$$E|X_m - X_n|^r = E|X_m|^r + E|X_n|^r \geq 2 ,$$

whence we conclude that

$$\lim_{m, n \rightarrow \infty} E|X_m - X_n|^r \neq 0$$

and  $(X_n)_{n \in \mathbf{N}}$  does not converge in  $r^{\text{th}}$  mean for  $r \geq 1$ .

(c) *Convergence in  $r^{\text{th}}$  mean  $\not\Rightarrow$  almost sure convergence:* We modify the above example by letting  $X_n = I_{G_n}$ . Then

$$\lim_n E|X_n|^r = \lim_n P(G_n) = 0$$

and  $X_n \xrightarrow{L^r} 0$ ; yet  $(X_n)_{n \in \mathbf{N}}$  again does not converge in probability.

(d) *Almost sure convergence  $\not\Rightarrow$  convergence in  $r^{\text{th}}$  mean:* On the same probability space as above we let  $G_n = (0, 1/n]$ ,  $X_n = n I_{G_n}$ . Then for every  $\omega$  in  $(0, 1]$  we have

$X_n(\omega) = 0$  for  $n > 1/\omega$ , and thus  $(X_n)_{n \in \mathbf{N}}$  converges to 0 pointwise on  $\Omega$ . On the other hand, for  $n = 2m$  we have

$$E|X_n - X_m|^r = m^{r-1},$$

whence we conclude that for  $r \geq 1$ ,

$$\lim_{m,n \rightarrow \infty} E|X_m - X_n|^r \neq 0,$$

and  $(X_n)_{n \in \mathbf{N}}$  does not converge in  $r^{\text{th}}$  mean for  $r \geq 1$ .

(e) *Almost sure convergence*  $\not\Rightarrow$  *pointwise convergence on  $\Omega$* : left as an easy exercise.

**Remark.** As it turns out, the following weaker implications are also true (see Billingsley):

(i) if  $X_n \xrightarrow{P} X$ , then a subsequence of  $(X_n)_{n \in \mathbf{N}}$  converges to  $X$  almost surely;

(ii) if  $X_n \xrightarrow{a.s.} X$  and there exists an integrable random variable  $Y \geq 0$  such that  $(\forall n) Y \geq |X_n|^r$ , then  $X_n \xrightarrow{L_r} X$ .

Thus in (b) above, the subsequence  $(X_{n(k)})_{k \in \mathbf{N}}$ , where  $n(k) = 2^k$ , converges to 0 almost surely; whereas the condition on the existence of an auxiliary r.v.  $Y$  is violated in (d).

## 19. Convergence and expectation

Billingsley, Sec. 5, *Inequalities*; Sec. 16, *Integration to the Limit*; Sec. 21, *Inequalities*.

A question often asked is the following: if a sequence  $(X_n)_{n \in \mathbf{N}}$  converges (in some sense) to a random variable  $X$ , is it true that the expectations  $EX_n$  converge to  $EX$ ?

As it turns out, this is always true if convergence to  $X$  is in the  $r$ -th mean, where  $r \geq 1$ . For the other standard modes of convergence, the above statement is not always true. To see this, consider the example given under (d) in the previous section. We had  $X_n \rightarrow X = 0$  pointwise on  $\Omega$  (and hence also almost surely, in probability and in distribution), yet  $EX_n = 1 \neq 0$ .

To show that  $X_n \xrightarrow{L_r} X$  for  $r \geq 1$  implies  $EX_n \rightarrow EX$ , we invoke the following variant of the powerful Hölder inequality (see Billingsley, Section 5, *Inequalities*):

$$\left(E[|Y|^s]\right)^{1/s} \leq \left(E[|Y|^r]\right)^{1/r}, \quad 0 < s \leq r.$$

Setting  $Y = X - X_n$  in the above, we obtain the general fact

$$X_n \xrightarrow{L_r} X \implies X_n \xrightarrow{L_s} X, \text{ for } s \leq r.$$

The desired conclusion is reached by setting  $s = 1$ , and using the simple inequality

$$|EX - EX_n| \leq E|X - X_n|. \quad (1)$$

**(Remark.** It is actually true that  $X_n \xrightarrow{L_r} X$  actually implies  $E|X_n|^s \rightarrow E|X|^s$  for all  $1 \leq s \leq r$ . To prove this, instead of (1) use

$$\left| E[|X|^s] - E[|X_n|^s] \right| \leq E[|X - X_n|^s] ,$$

which follows from Minkowski's inequality (ibid.):

$$E[|Y + Z|^s] \leq E[|Y|^s] + E[|Z|^s] , \quad s \geq 1.)$$

In regard to other modes of convergence, we give sufficient conditions for the interchange of limits and expectation in the case of almost sure convergence. These conditions are contained in the following important theorems, which are discussed in Billingsley, Section 16.

**Monotone convergence theorem.** If  $(X_n)_{n \in \mathbf{N}}$  is a nondecreasing sequence of r.v.'s that converges almost surely to  $X$ , and is bounded from below by an integrable r.v., then  $EX_n \rightarrow EX$ . In other words,

$$X_n \xrightarrow{a.s.} X, \quad (\forall n) Y \leq X_n \leq X_{n+1}, \quad E|Y| < \infty \quad \implies \quad EX_n \rightarrow EX .$$

**(Remark.** Since  $Y \leq X$ ,  $EX \leq +\infty$ . The theorem is true even if  $EX = +\infty$ .)

**Dominated convergence theorem.** If  $(X_n)_{n \in \mathbf{N}}$  is a sequence of r.v.'s that converges almost surely to  $X$  and is absolutely bounded by an integrable r.v., then  $EX_n \rightarrow EX$ . In other words,

$$X_n \xrightarrow{a.s.} X, \quad (\forall n) |X_n| \leq Y, \quad E|Y| < \infty \quad \implies \quad EX_n \rightarrow EX .$$

As an application of the monotone convergence theorem, consider a sequence of non-negative random variables  $Y_n$ , and let  $X_n = Y_1 + \dots + Y_n$ . Then the  $X_n$ 's form a nondecreasing sequence that converges pointwise to  $\sum_{i=1}^{\infty} Y_i$  and is bounded from below by 0. By the monotone convergence theorem,

$$\lim_n EX_n = E\left[\sum_{i=1}^{\infty} Y_i\right] , \quad \text{or equivalently,} \quad E\left[\sum_{i=1}^{\infty} Y_i\right] = \sum_{i=1}^{\infty} EY_i .$$

Note that by the remark following the statement of the theorem, the above holds even if the sum of all expectations is infinite in value.

As an application of the dominated convergence theorem, consider the following example. Let  $Y \geq 0$  be integrable,  $X \geq 0$  be arbitrary, and define

$$Z_n = \frac{XY}{n} \wedge \alpha Y ,$$

where  $\alpha > 0$ . Note that  $(Z_n)_{n \in \mathbf{N}}$  is a nonincreasing sequence such that

$$(\forall \omega) \quad 0 \leq \lim_n Z_n(\omega) \leq \lim_n \frac{X(\omega)Y(\omega)}{n} = 0 .$$

Thus  $Z_n \rightarrow Z = 0$  pointwise on  $\Omega$ . Since we also have  $0 \leq Z_n \leq \alpha Y$  and  $Y$  is integrable, we can apply the dominated convergence theorem to conclude that

$$\lim_n EZ_n = EZ = 0 .$$

Note that the above conclusion does not necessarily hold if  $Y$  is nonintegrable. For example, if  $EY = +\infty$  and  $X$  is a constant r.v., then  $EZ_n = +\infty$  for all  $n$ , and thus  $EZ_n \not\rightarrow 0$ .

It is instructive to verify that the hypotheses of the above two theorems do not hold in the example given under (d) in the previous section. Indeed, the sequence  $(X_n)_{n \in \mathbf{N}}$  is not monotone nondecreasing, and hence the monotone convergence theorem does not apply here. Furthermore, any dominating r.v.  $Y$  will have to dominate  $\sup_n X_n$ ; the latter is a simple r.v. whose expectation is

$$\sum_{n=1}^{\infty} n \left( \frac{1}{n} - \frac{1}{n+1} \right) = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty .$$

Hence the dominated convergence theorem does not apply here either.

## 20. Laws of large numbers

Billingsley, Sec. 6; Sec. 21, *Moment Generating Functions*; Sec. 22, *Kolmogorov's Zero-One Law*, *Kolmogorov's Inequality*, *The Strong Law of Large Numbers*.

Consider a sequence  $(X_n)_{n \in \mathbf{N}}$  of random variables in a probability experiment  $(\Omega, \mathcal{F}, P)$ , and suppose that the marginal distributions  $(F_{X_n})_{n \in \mathbf{N}}$  have some common numerical attribute whose value is unknown to us. In particular, we will assume that this attribute coincides with the mean  $\mu$  of each  $F_{X_n}$  (by a simple change of variable, we can see that this restriction also covers cases in which the common attribute is expressible as an expectation of a function of  $X_n$ ).

Suppose now that we wish to estimate the value of  $\mu$  on the basis of the observed sequence

$$X_1(\omega), X_2(\omega), \dots ;$$

here  $\omega$  is the actual outcome of the probability experiment. The *weak* and *strong laws of large numbers* ensure that such inference is possible (with reasonable accuracy) provided the dependencies between the  $X_n$ 's are suitably restricted: the weak law is valid for uncorrelated  $X_n$ 's, while the strong law is valid for independent  $X_n$ 's. Since independence is a more restrictive condition than absence of correlation (for square-integrable random variables), one expects the strong law to be more powerful than the weak law. This is indeed the case, as the weak law states that the sample average

$$\frac{X_1 + \dots + X_n}{n}$$

converges to the constant  $\mu$  in probability, while the strong law asserts that this convergence takes place almost surely.

### Weak law of large numbers

**Theorem.** Let  $(X_n)_{n \in \mathbf{N}}$  be a sequence of uncorrelated random variables with common mean  $EX_i = \mu$ . If the variables also have common variance, or more generally, if

$$\lim_n \frac{1}{n^2} \sum_{i=1}^n \text{Var} X_i = 0 ,$$

then the sample average

$$\frac{X_1 + \cdots + X_n}{n}$$

converges to the mean  $\mu$  in probability.

**Proof.** As was shown in Sec. 17, for every  $\alpha > 0$  the Chebyshev inequality gives

$$\Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \alpha \right\} \leq \frac{1}{n^2 \alpha^2} \sum_{i=1}^n \text{Var} X_i .$$

Thus under the stated assumptions on the variances, we have

$$\lim_n \Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| \geq \alpha \right\} = 0 ,$$

which proves that  $X_n \xrightarrow{P} \mu$ .

**Remark.** Note that as it appears above, the right-hand side in the Chebyshev inequality is just the second moment of the difference between the  $n$ -sample average and the mean  $\mu$ . Thus the variance constraint is equivalent to the statement that  $X_n \xrightarrow{q.m.} \mu$ , which also implies  $X_n \xrightarrow{P} \mu$ .

**Example.** Let  $\Omega = (0, 2\pi]$ ,  $\mathcal{F} = \mathcal{B}(\Omega)$ ,  $P =$  uniform measure, and define  $X_n$  by

$$X_n(\omega) = a_n \sin n\omega , \quad (a_n \in \mathbf{R}).$$

Then

$$EX_n = \int X_n(\omega) dP(\omega) = \frac{1}{2\pi} \int_0^{2\pi} \sin n\omega d\omega = 0 ,$$

and

$$\text{Cov}(X_m, X_n) = \frac{a_m a_n}{2\pi} \int_0^{2\pi} \sin m\omega \sin n\omega d\omega = \begin{cases} a_n^2/4, & \text{if } n = m; \\ 0 & \text{otherwise.} \end{cases}$$

Thus  $(X_n)_{n \in \mathbf{N}}$  is an uncorrelated sequence. Assuming that  $\lim_n (1/n^2) \sum_{i=1}^n a_i^2 = 0$ , we can apply the weak law of large numbers to conclude that for every  $\epsilon > 0$ ,

$$\lim_n P \left\{ \omega : \frac{1}{n} \left| \sum_{i=1}^n a_i \sin n\omega \right| \geq \epsilon \right\} = 0 .$$

Note that the set in the above expression is a union of intervals, since the function

$$h(\omega) = a_0 + \sum_{i=1}^n a_n \sin n\omega$$

has only finitely many roots in the interval  $(0, 2\pi]$ . Thus the law of large numbers implies that the total length of those subintervals of  $(0, 2\pi]$  over which

$$\frac{1}{n} \left| \sum_{i=1}^n a_n \sin n\omega \right| \geq \epsilon$$

approaches zero as  $n$  tends to infinity.

### A version of the strong law of large numbers

The *moment generating function*  $M_X(s)$  of a random variable  $X$  is defined for real  $s$  by

$$M_X(s) \stackrel{\text{def}}{=} E[e^{sX}] = \int e^{st} dP_X(t) .$$

It is easily seen that for absolutely continuous  $P_X$ , the above coincides with the bilateral Laplace transform of the density  $f_X$ .

We always have  $M_X(0) = 1$ ; if  $X$  is almost surely bounded, then it is also true that  $M_X(s) < \infty$  for every  $s$ . For unbounded  $X$ , it is possible to obtain  $M_X(s) = +\infty$  for some or all values of  $s$  other than 0. However, for many distributions encountered in practice (including the gaussian, gamma, geometric, Laplace and Poisson distributions)  $M_X(s)$  is finite for  $s$  ranging over some interval containing the origin. For such distributions, we can prove the following version of the strong law of large numbers.

**Theorem.** Let  $(X_n)_{n \in \mathbf{N}}$  be an i.i.d. sequence whose marginal moment generating function  $M_X(\cdot)$  is finite over a nontrivial interval containing the origin. Then

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{a.s.} EX_1 = \mu .$$

**Proof.** Let  $M_X(s)$  be finite for  $s \in [-s_0, s_0]$ , and note that since

$$s_0|X_1| < e^{s_0 X_1} + e^{-s_0 X_1} ,$$

the mean  $\mu$  is also finite. Thus we can assume  $\mu = 0$  without loss of generality.

We will compute an upper bound to the probability that the sample average of the first  $n$  terms is greater than or equal to some  $\alpha > 0$ . For this purpose we use the Chernoff bound:

$$(\forall s > 0) \quad \Pr\{Y \geq \alpha\} \leq E \left[ e^{s(Y-\alpha)} \right] = e^{-s\alpha} M_Y(s) .$$

Thus we obtain

$$\begin{aligned} \Pr\left\{\frac{X_1 + \cdots + X_n}{n} \geq \alpha\right\} &= \Pr\{X_1 + \cdots + X_n > n\alpha\} \\ &\leq E\left[e^{s(X_1 + \cdots + X_n - n\alpha)}\right] \\ &= E\left[\prod_{i=1}^n e^{s(X_i - \alpha)}\right] = \prod_{i=1}^n E\left[e^{s(X_i - \alpha)}\right] = \left(e^{-s\alpha} M_X(s)\right)^n \end{aligned}$$

where the last two equalities follow from the i.i.d. assumption.

The next step is to show that there exists a point  $s$  in  $(0, s_0)$  such that  $e^{-s\alpha} M_X(s) < 1$ . We know that  $e^{-0\cdot\alpha} M_X(0) = 1$ , hence it suffices to show that the derivative of  $e^{-s\alpha} M_X(s)$  takes a negative value at the origin. We have

$$\left.\frac{d}{ds}\left(e^{-s\alpha} M_X(s)\right)\right|_{s=0} = -\alpha M_X(0) + M'_X(0) = -\alpha + M'_X(0).$$

To compute  $M'_X(0)$ , we write

$$M'_X(s) = \frac{d}{ds} E[e^{sX}] = E\left[\frac{d}{ds} e^{sX}\right] = E[Xe^{sX}],$$

where the interchange of derivative and Lebesgue integral can be justified using the dominated convergence theorem (see also the discussion of moment generating functions in Billingsley). We thus obtain

$$\left.\frac{d}{ds}\left(e^{-s\alpha} M_X(s)\right)\right|_{s=0} = -\alpha + E[X] = -\alpha < 0$$

as sought.

Hence for every  $\alpha > 0$  there exists a constant  $\rho_+ < 1$  such that

$$\Pr\left\{\frac{X_1 + \cdots + X_n}{n} \geq \alpha\right\} \leq \rho_+^n.$$

Using the same argument with  $-X_i$  replacing  $X_i$  and  $-s$  replacing  $s$ , we infer that there exists a constant  $\rho_- < 1$  such that

$$\Pr\left\{\frac{X_1 + \cdots + X_n}{n} \leq -\alpha\right\} \leq \rho_-^n.$$

Thus if  $\rho = \max(\rho_+, \rho_-)$ , we have

$$\Pr\left\{\left|\frac{X_1 + \cdots + X_n}{n}\right| \geq \alpha\right\} \leq 2\rho^n,$$

and since  $\rho < 1$ ,

$$\sum_{n=1}^{\infty} \Pr \left\{ \left| \frac{X_1 + \cdots + X_n}{n} \right| \geq \alpha \right\} \leq \sum_{n=1}^{\infty} 2\rho^n = \frac{2\rho}{\rho - 1} < \infty. \quad (1)$$

Now the event over which the sample average converges to zero is given by

$$A = \bigcap_{q \in \mathbf{N}} \bigcup_m \bigcap_{n \geq m} \left\{ \omega : \left| \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} \right| < q^{-1} \right\},$$

and thus

$$\begin{aligned} A^c &= \bigcup_{q \in \mathbf{N}} \bigcap_m \bigcup_{n \geq m} \left\{ \omega : \left| \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} \right| \geq q^{-1} \right\} \\ &= \bigcup_{q \in \mathbf{N}} \limsup_n \left\{ \omega : \left| \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} \right| \geq q^{-1} \right\}. \end{aligned}$$

By the first Borel-Cantelli lemma (Billingsley, Sec. 4),

$$\sum_{n=1}^{\infty} P(B_n) < \infty \implies P(\limsup_n B_n) = 0,$$

and hence by virtue of (1), we have  $P(A^c) = 0$ . Thus the sample average converges to zero almost surely.

### Kolmogorov's Strong Law of Large Numbers

The general version of the strong law of large numbers is due to Kolmogorov. It is appreciably stronger than the statement of the previous theorem, in that it posits convergence of sample averages under constraints on the first two moments of the independent variables. For a proof, see Billingsley.

**Theorem.** Let  $(X_n)_{n \in \mathbf{N}}$  be an independent sequence of random variables with common mean  $EX_n = \mu$ . If either

- (i) the  $X_n$ 's are identically distributed; or
- (ii) the  $X_n$ 's are square-integrable with

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

then the sample average

$$\frac{X_1 + \cdots + X_n}{n}$$

converges to  $\mu$  almost surely.

### Remarks.

1. Note that the i.i.d. assumption (case (i)) above) does not exclude the possibility  $\mu = \pm\infty$ , in which case the sample average converges almost surely to a constant *extended* random variable.

2. For most practical purposes, when considering sequences of independent (not just uncorrelated) square-integrable random variables, Kolmogorov's strong law of large numbers subsumes the weak law as stated earlier in this section. This is because almost sure convergence always implies convergence in probability. It is worth noting, however, that the variance constraint employed in the statement of the weak law, namely

$$\lim_n \frac{1}{n^2} \sum_{i=1}^n \text{Var} X_i = 0 ,$$

is somewhat more general than the condition  $\sum_{n=1}^{\infty} \text{Var} X_n/n^2 < \infty$  in Kolmogorov's strong law of large numbers. Thus there are instances of independent sequences to which the weak law applies, but the strong law does not.

### Examples

(a) Consider again the sequence  $(X_n)_{n \in \mathbf{N}}$  of random variables of Section 2. For a random point  $\omega$  drawn uniformly from the unit interval  $(0, 1]$ , we defined

$$X_k(\omega) = k^{\text{th}} \text{ digit in the binary expansion of } \omega .$$

The  $X_k$ 's were easily shown to be i.i.d. with  $EX_k = 1/2$ . By the strong law of large numbers,

$$P \left\{ \omega : \lim_{n \rightarrow \infty} \frac{X_1(\omega) + \cdots + X_n(\omega)}{n} = \frac{1}{2} \right\} = 1 .$$

As a consequence, the set  $A_{1/2}$  of points on the unit interval whose binary expansions have running averages that converge to  $1/2$  has measure 1 under the Lebesgue measure; thus  $A_{1/2}$  is in a sense *as large* as the unit interval itself. Yet in a different sense,  $A_{1/2}$  is *much smaller* than the unit interval:  $A_{1/2}$  does not overlap with any of the equally populous sets  $D$  (consisting of all points whose asymptotic expansions have divergent time averages) or  $A_p$  (defined as  $A_{1/2}$  with  $p \neq 1/2$  replacing  $1/2$ ). We have thus exhibited a Borel subset of the unit interval whose Lebesgue measure equals one and whose complement is uncountable.

**Question.** What is the Lebesgue measure of all points in the unit interval whose *ternary* expansions have running averages that converge to  $1/3$ ?

(b) Note that if  $(X_n)_{n \in \mathbf{N}}$  is an i.i.d. sequence, then so is  $(g(X_n))_{n \in \mathbf{N}}$ , where  $g : \mathbf{R} \mapsto \mathbf{R}$  is Borel measurable. Thus if  $E[g(X_1)]$  exists,

$$\frac{g(X_1) + \cdots + g(X_n)}{n} \xrightarrow{a.s.} E[g(X_1)] ,$$

and a simple variable transformation can be used to estimate  $E[g(X_1)]$  from the observations. Thus for example, we can use

$$\frac{X_1^2 + \cdots + X_n^2}{n}$$

to estimate the second moment, which always exists.

The assumption that the  $X_n$ 's are identically distributed is clearly crucial here. Also note that we **cannot** give a similar argument for uncorrelated random variables based on the weak law, even under the assumption that the marginals are identical; this is so because  $g(Y_i)$  and  $g(Y_j)$  can be correlated even if  $Y_i$  and  $Y_j$  are not.

Finally, consider the transformation  $g = I_H$ , where  $H$  is a linear Borel set. As before, we have

$$\frac{I_H(X_1) + \cdots + I_H(X_n)}{n} \xrightarrow{a.s.} E[I_H(X_1)] ,$$

or equivalently,

$$\frac{\text{number of } k\text{'s (where } 1 \leq k \leq n) \text{ for which } X_k \in H}{n} \xrightarrow{a.s.} P_{X_1}(H) .$$

We can interpret the above result as follows.  $(X_n)_{n \in \mathbf{N}}$  defines a sequence of subexperiments of  $(\Omega, \mathcal{F}, P)$  that serve as mathematical models for the *independent repetitions* of some physical experiment  $X$ . The last relationship effectively states that the *relative frequency* of the event  $\{X \in H\}$  amongst the first  $n$  independent repetitions of  $X$  converges almost surely to the probability of that event. This corroborates the so-called *frequentist* view of probability, according to which the probability of an uncertain event in a random experiment is an objective entity which is revealed to an (infinitely patient) observer by the limit of the relative frequency of that event in a sequence of independent repetitions of the experiment.

## 21. Characteristic functions

Billingsley, Sec. 26.

The **characteristic function**  $\phi_X(u)$  of a random variable  $X$  is defined for real  $u$  by

$$\phi_X(u) \stackrel{\text{def}}{=} E[e^{iuX}] = E[\cos uX] + iE[\sin uX] .$$

If  $X$  has density  $f_X$ , then

$$\phi_X(u) = \int_{-\infty}^{\infty} e^{iut} f_X(t) dt ,$$

and thus  $\phi_X$  is the Fourier transform of  $f_X$  (the moment generating function was seen to coincide with the bilateral Laplace transform of  $f_X$ ).

The function  $\phi_X$  is continuous on the real line, and is such that

$$\phi_X(0) = 1 , \quad (\forall u) \quad |\phi_X(u)| \leq 1 .$$

When  $X$  has a symmetric distribution  $P_X$  (e.g., when  $f_X$  is an even function), the characteristic function  $\phi_X$  is real-valued and symmetric.

Characteristic functions are discussed at length in Billingsley. We briefly note the following essential properties.

(a)  $P_X$  completely specifies  $\phi_X$  and vice versa. The forward assertion is true by definition of  $\phi_X$ . The converse can be also shown to be true: the increment  $F_X(b) - F_X(a)$  equals the quantity

$$\lim_{M \rightarrow \infty} \frac{1}{2\pi} \int_{-M}^M \frac{e^{-iua} - e^{-iub}}{iu} \phi(u) du$$

whenever  $P_X\{a\} = P_X\{b\} = 0$ , and thus it is possible to deduce the values of  $F_X$  at all points of continuity. In the special case in which the modulus  $|\phi_X|$  is integrable, the distribution  $P_X$  has a continuous density  $f_X$  which can be recovered from  $\phi_X$  using the inverse Fourier transform:

$$f_X(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iut} \phi_X(u) du .$$

(b) *Convergence of characteristic functions.* Recall the definition of convergence in distribution:  $X_n \xrightarrow{d} X$  if  $F_{X_n}$  converges to  $F_X$  at every point of continuity of  $F_X$ . As it turns out, convergence in distribution can be also defined in terms of characteristic functions:  $X_n \xrightarrow{d} X$  if and only if  $\phi_{X_n}(u)$  converges to  $\phi_X(u)$  at every point  $u$ .

It is often advantageous to study convergence in distribution in terms of characteristic functions. If a sequence  $(\phi_n)_{n \in \mathbf{N}}$  of characteristic functions has the property that  $\phi_n(u)$  converges at every  $u$  to limit  $\phi(u)$  such that  $\phi(0) = 1$ , then the limiting function  $\phi$  is also a characteristic function and convergence in distribution is ensured. This gives us a convenient criterion for this mode of convergence that was absent in the original formulation in terms of cdf's (see Section 18).

(c) *Independence and Characteristic Functions.* Suppose  $X_1, \dots, X_n$  are independent random variables. Then the characteristic function of the sum

$$Y = X_1 + \dots + X_n$$

is given by

$$\begin{aligned} \phi_Y(u) &= E[e^{i(uX_1 + \dots + uX_n)}] \\ \text{(by independence)} \quad &= E[e^{iuX_1}] \dots E[e^{iuX_n}] = \phi_{X_1}(u) \dots \phi_{X_n}(u) . \end{aligned}$$

An analogous property was seen to be true for the moment generating function. If each  $X_k$  has absolutely continuous distribution, the density of  $Y$  is given by the convolution of the densities of the  $X_k$ 's. This is consistent with the well-known fact of Fourier analysis that convolution in the time ( $t$ ) domain is equivalent to multiplication in the frequency ( $u$ ) domain.

## Characteristic function of a random vector

**Convention.** In all algebraic manipulations that follow, boldface symbols such as  $\mathbf{a}$  and  $\mathbf{X}$  will denote **column** vectors. Thus row vectors will be denoted by  $\mathbf{a}^T$  and  $\mathbf{X}^T$ .

The transposition symbol will be omitted where no confusion is likely to arise, e.g., when writing  $\mathbf{X} = (X_1, \dots, X_n)$ .

The characteristic function of an  $n$ -dimensional random vector  $\mathbf{X} = (X_1, \dots, X_n)$  is the function  $\phi_{\mathbf{X}} : \mathbf{R}^n \mapsto \mathbf{R}$  defined by

$$\phi_{\mathbf{X}}(\mathbf{u}) \stackrel{\text{def}}{=} E \left[ e^{i\mathbf{u}^T \mathbf{X}} \right] ,$$

or equivalently,

$$\phi_{\mathbf{X}}(u_1, \dots, u_n) = E[\exp i(u_1 X_1 + \dots + u_n X_n)] .$$

If  $\mathbf{X}$  has absolutely continuous distribution,

$$\phi_{\mathbf{X}}(u_1, \dots, u_n) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{i(u_1 x_1 + \dots + u_n x_n)} f_{\mathbf{X}}(x_1, \dots, x_n) dx_1 \dots dx_n ,$$

and thus  $\phi_{\mathbf{X}}$  coincides with the  $n$ -dimensional Fourier transform of the density  $f_{\mathbf{X}}$ .

The general properties of  $n$ -variate characteristic functions are similar to those of univariate ones (see (a)–(c) in the previous subsection). A simple property that will prove quite useful is the following: if  $A$  is an  $m \times n$  matrix and  $\mathbf{b}$  is an  $m$ -dimensional (column) vector, then the characteristic function of  $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$  evaluated at  $\mathbf{u} \in \mathbf{R}^m$  is given by

$$\phi_{\mathbf{Y}}(\mathbf{u}) = E \left[ e^{i\mathbf{u}^T (A\mathbf{X} + \mathbf{b})} \right] = e^{i\mathbf{b}^T \mathbf{u}} E \left[ e^{i(A^T \mathbf{u})^T \mathbf{X}} \right] = e^{i\mathbf{b}^T \mathbf{u}} \phi_{\mathbf{X}}(A^T \mathbf{u}) .$$

Also, if the components of  $\mathbf{X}$  are *independent*,

$$\phi_{\mathbf{X}}(u_1, \dots, u_n) = \phi_{X_1}(u_1) \dots \phi_{X_n}(u_n) .$$

## Expectation of vectors and the covariance matrix

If

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \quad \text{and} \quad \Theta = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{12} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \dots & X_{mn} \end{pmatrix} ,$$

we define

$$E\mathbf{X} = \begin{pmatrix} EX_1 \\ \vdots \\ EX_n \end{pmatrix} \quad \text{and} \quad E\Theta = \begin{pmatrix} EX_{11} & EX_{12} & \dots & EX_{1n} \\ EX_{12} & EX_{22} & \dots & EX_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ EX_{m1} & EX_{m2} & \dots & EX_{mn} \end{pmatrix} .$$

The **covariance matrix** of  $\mathbf{X}$  is the  $n \times n$  matrix  $C_{\mathbf{X}}$  defined by

$$C_{\mathbf{X}} \stackrel{\text{def}}{=} E[(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})^T] .$$

Thus the  $(i, j)^{\text{th}}$  entry of  $C_{\mathbf{X}}$  is simply given by

$$(C_{\mathbf{X}})_{ij} = E[(X_i - EX_i)(X_j - EX_j)] = \text{Cov}(X_i, X_j) ,$$

and  $C_{\mathbf{X}}$  is symmetric.

It is easy to verify that expectation is linear, in that

$$E[A\mathbf{X}] = AEX, \quad E[\mathbf{X}A] = E[\mathbf{X}]A$$

and

$$E[B\Theta] = BE[\Theta], \quad E[\Theta B] = E[\Theta]B$$

for constant matrices  $A$  and  $B$  of appropriate size. We can use linearity to evaluate the covariance matrix of  $A\mathbf{X}$  as follows:

$$\begin{aligned} C_{A\mathbf{X}} &= E[(A\mathbf{X} - E(A\mathbf{X}))(A\mathbf{X} - E(A\mathbf{X}))^T] \\ &= E[A(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})^T A^T] \\ &= E[A(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})^T] A^T \\ &= AE[(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})^T] A^T \\ &= AC_{\mathbf{X}}A^T . \end{aligned}$$

## 22. Gaussian random variables and vectors

Billingsley, Sec. 29, *Normal Distributions in  $R^k$* . s

### Gaussian variables

**Definition.** A random variable  $X$  is Gaussian if either of the following is true:

- (a)  $X$  is constant with probability 1;
- (b)  $X$  has density

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

for  $\sigma^2 \geq 0$ .

In case (a), we have  $X = EX$  with probability 1, and  $\text{Var}X = 0$ . In case (b), we have  $EX = \mu$ ,  $\text{Var}X = \sigma^2$ . In both cases we use the notation

$$X \sim \mathcal{N}(EX, \text{Var}X) ;$$

here  $\mathcal{N}$  stands for “normal,” which synonymous with “Gaussian.” Thus the univariate Gaussian distribution is parametrically specified through its mean and its variance.

The characteristic function of a  $\mathcal{N}(0, 1)$  distribution is given by

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{iux - (x^2/2)} dx = \frac{e^{-u^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-iu)^2/2} dx .$$

The last integral in the above expression is equal to that of the analytic function  $\exp(-z^2/2)$  along a path parallel to the real axis, and reduces by a standard argument to the integral of the same function over the real axis itself. Thus

$$\phi(u) = \frac{e^{-u^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx = e^{-u^2/2} .$$

If  $X \sim \mathcal{N}(\mu, \sigma^2)$  and  $\sigma^2 \geq 0$ , we can write

$$X = \sigma Y + \mu$$

for appropriate  $Y \sim \mathcal{N}(0, 1)$ , and thus

$$\phi_X(u) = e^{i\mu u} \phi_Y(\sigma u) = e^{i\mu u - \frac{1}{2}\sigma^2 u^2} .$$

If  $X \sim \mathcal{N}(\mu, 0)$ , then  $X = \mu$  with probability 1, and

$$\phi_X(u) = e^{i\mu u} .$$

Thus the identity

$$\phi_X(u) = \exp \left\{ i(EX)u - \frac{1}{2}(\text{Var}X)u^2 \right\}$$

holds for all Gaussian  $X$ .

## Gaussian vectors

**Definition.** An  $n$ -dimensional random vector  $\mathbf{X}$  is Gaussian if it can be expressed in the form

$$\mathbf{X} = A\mathbf{Y} + \mathbf{b} ,$$

where  $A$  is a nonrandom  $n \times n$  matrix,  $\mathbf{b}$  is a nonrandom  $n$ -dimensional vector, and  $Y$  is an  $n$ -dimensional random vector of *independent*  $\mathcal{N}(0, 0)$  or  $\mathcal{N}(0, 1)$  components.

Two alternative definitions of a Gaussian random vector are suggested by the following theorem.

### Theorem.

(i) A  $n$ -dimensional random vector  $\mathbf{X}$  is Gaussian if and only if for all  $\mathbf{a} \in \mathbf{R}^n$ ,  $\mathbf{a}^T \mathbf{X}$  is a Gaussian random variable.

(ii) A  $n$ -dimensional random vector  $\mathbf{X}$  is Gaussian if and only if there exists a vector  $\mathbf{m} \in \mathbf{R}^n$  and an  $n \times n$  nonnegative-definite symmetric matrix  $Q$  such that

$$\phi_{\mathbf{X}}(\mathbf{u}) = \exp \left\{ i\mathbf{m}^T \mathbf{u} - \frac{1}{2} \mathbf{u}^T Q \mathbf{u} \right\} .$$

**Proof.** We will establish each of the following implications:

1.  $\mathbf{X}$  is Gaussian  $\implies \mathbf{X}$  satisfies condition of statement (i)
2.  $\mathbf{X}$  satisfies condition of statement (i)  $\implies \mathbf{X}$  satisfies condition of statement (ii)
3.  $\mathbf{X}$  satisfies condition of statement (ii)  $\implies \mathbf{X}$  is Gaussian

1. Let  $\mathbf{X} = A\mathbf{Y} + \mathbf{b}$ , where  $\mathbf{Y}$  is an  $n$ -dimensional random vector with independent  $\mathcal{N}(0, 0)$  or  $\mathcal{N}(0, 1)$  components; for concreteness, let  $Y_k \sim \mathcal{N}(0, \sigma_k^2)$ , so that  $(C_{\mathbf{Y}})_{kk} = \sigma_k^2$ .

By virtue of independence,  $\phi_{\mathbf{Y}}$  is given by

$$\begin{aligned} \phi_{\mathbf{Y}}(u_1, \dots, u_n) &= \prod_{k=1}^n \phi_{Y_k}(u_k) \\ &= \prod_{k=1}^n e^{-\sigma_k^2 u_k^2 / 2} \\ &= \exp \left\{ -\frac{1}{2} \mathbf{u}^T C_{\mathbf{Y}} \mathbf{u} \right\} . \end{aligned}$$

Since  $\mathbf{a}^T \mathbf{X} = \mathbf{a}^T A \mathbf{Y} + \mathbf{a}^T \mathbf{b}$ , we have

$$\begin{aligned} \phi_{\mathbf{a}^T \mathbf{X}}(u) &= \exp\{i(\mathbf{a}^T \mathbf{b})u\} \phi_{\mathbf{Y}}(A^T \mathbf{a}u) \\ &= \exp \left\{ i(\mathbf{a}^T \mathbf{b})u - \frac{1}{2} (\mathbf{a}^T A) C_{\mathbf{Y}} (\mathbf{a}^T A)^T u^2 \right\} . \end{aligned}$$

Since  $E[\mathbf{a}^T \mathbf{X}] = \mathbf{a}^T \mathbf{b}$  and

$$\text{Var}(\mathbf{a}^T \mathbf{X}) = C_{\mathbf{a}^T \mathbf{X}} = C_{\mathbf{a}^T A \mathbf{Y}} = (\mathbf{a}^T A) C_{\mathbf{Y}} (\mathbf{a}^T A)^T ,$$

we conclude that  $\mathbf{a}^T \mathbf{X}$  is a Gaussian random variable.

2. Assume that  $\mathbf{a}^T \mathbf{X}$  is Gaussian for every choice of  $\mathbf{a} \in \mathbf{R}^k$ . Thus for  $\mathbf{u} \in \mathbf{R}^k$  we can write

$$\begin{aligned} \phi_{\mathbf{X}}(\mathbf{u}) &= E \left[ e^{i\mathbf{u}^T \mathbf{X}} \right] \\ &= \phi_{\mathbf{u}^T \mathbf{X}}(1) = \exp \left\{ iE[\mathbf{u}^T \mathbf{X}] - \frac{1}{2} \text{Var}(\mathbf{u}^T \mathbf{X}) \right\} . \end{aligned}$$

Now  $E[\mathbf{u}^T \mathbf{X}] = \mathbf{u}^T E[\mathbf{X}]$  and  $\text{Var}(\mathbf{u}^T \mathbf{X}) = \mathbf{u}^T C_{\mathbf{X}} \mathbf{u}$ , whence we conclude that

$$\phi_{\mathbf{X}}(\mathbf{u}) = \exp \left\{ i(E\mathbf{X})^T \mathbf{u} - \frac{1}{2} \mathbf{u}^T C_{\mathbf{X}} \mathbf{u} \right\} .$$

$C_{\mathbf{X}}$  is always symmetric; for nonnegative-definiteness we need  $\mathbf{u}^T C_{\mathbf{X}} \mathbf{u} \geq 0$  for all  $\mathbf{u}$ , which is true since  $\mathbf{u}^T C_{\mathbf{X}} \mathbf{u} = \text{Var}(\mathbf{u}^T \mathbf{X})$ .

3. Let the  $n$ -dimensional random vector  $\mathbf{X}$  have characteristic function

$$\phi_{\mathbf{X}}(\mathbf{u}) = \exp \left\{ i\mathbf{m}^T \mathbf{u} - \frac{1}{2} \mathbf{u}^T Q \mathbf{u} \right\},$$

where  $Q$  is a  $n \times n$  nonnegative-definite symmetric matrix. From elementary linear algebra, we can express any  $n \times n$  symmetric matrix  $Q$  as

$$Q = B\Lambda B^T,$$

where  $\Lambda$  is an  $n \times n$  diagonal matrix whose diagonal entries coincide with the eigenvalues of  $Q$ , and  $B$  is an  $n \times n$  matrix whose columns coincide with the column eigenvectors of  $Q$ .  $B$  is *orthogonal*, i.e.,  $B^T B = I$ .

The assumption that  $Q$  is nonnegative-definite is the same as requiring that all eigenvalues of  $Q$  be nonnegative. If all eigenvalues are positive (i.e., if  $Q$  is positive-definite), then both  $Q^{-1}$  and  $\Lambda^{-1}$  exist. If, however, one or more eigenvalues are zero, then neither  $Q^{-1}$  nor  $\Lambda^{-1}$  exist. We surmount such difficulties by defining the matrix  $\Delta$  as follows:

$$\Delta_{kl} = \begin{cases} (\Lambda_{kl})^{-1/2}, & \text{if } \Lambda_{kl} > 0, \\ 1, & \text{otherwise.} \end{cases}$$

$\Delta$  is clearly invertible, and

$$J = \Delta \Lambda \Delta^T$$

is a diagonal matrix such that  $J_{kl} = 1$  if  $\Lambda_{kl} > 0$ ,  $J_{kl} = 0$  otherwise.

Consider now the transformation

$$\mathbf{Y} = \Delta B^T (\mathbf{X} - \mathbf{m}).$$

We have

$$\begin{aligned} \phi_{\mathbf{Y}}(\mathbf{u}) &= \phi_{\mathbf{X}-\mathbf{m}}(B\Delta^T \mathbf{u}) \\ &= \exp \left\{ -\frac{1}{2} \mathbf{u}^T (\Delta B^T) Q (\Delta B^T)^T \mathbf{u} \right\} \\ &= \exp \left\{ -\frac{1}{2} \mathbf{u}^T (\Delta B^T) B \Lambda B^T (\Delta B^T)^T \mathbf{u} \right\} \\ &= \exp \left\{ -\frac{1}{2} \mathbf{u}^T J \mathbf{u} \right\}. \end{aligned}$$

Thus  $\mathbf{Y}$  is a vector of independent  $\mathcal{N}(0, 0)$  or  $\mathcal{N}(0, 1)$  Gaussian components. It is easily verified that

$$\mathbf{X} = B\Delta^{-1} \mathbf{Y} + \mathbf{m}.$$

### Further properties of Gaussian vectors

(a) If  $\mathbf{X}$  is a Gaussian vector in  $\mathbf{R}^n$  and  $A : \mathbf{R}^n \mapsto \mathbf{R}^m$  is a linear transformation, then  $A\mathbf{X}$  is also a Gaussian vector in  $\mathbf{R}^m$ . Thus the Gaussian property is preserved under

all linear transformations. This fact follows easily from the characterization of Gaussian vectors in terms of their characteristic function; the special case  $m = 1$  was also treated in the theorem of the previous subsection.

**Remark.** Transformations of the form  $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$  are known as *affine*. Clearly the Gaussian property is also preserved under affine transformations.

(b) The distribution of a Gaussian vector  $\mathbf{X}$  is fully specified by its expectation  $E\mathbf{X}$  and covariance function  $C_{\mathbf{X}}$ ; no other parameters are needed.

(c) If the components of a Gaussian vector  $\mathbf{X}$  are uncorrelated, then they are also independent. This is so because for a diagonal covariance matrix  $C_{\mathbf{X}}$ , the characteristic function  $\phi_{\mathbf{X}}$  can be decomposed into a product of univariate Gaussian characteristic functions. Thus in the Gaussian case, the two properties of independence and absence of correlation are equivalent. In the general *non*-Gaussian case, we have seen that uncorrelatedness does *not* imply independence.

### Whitening of random vectors

Given *any* random vector  $\mathbf{X}$  in  $\mathbf{R}$  (not necessarily Gaussian), we can always find a zero-mean vector  $\mathbf{Y}$  in  $\mathbf{R}$  such that

$$\mathbf{X} = A\mathbf{Y} + \mathbf{m}$$

and the components of  $\mathbf{Y}$  are uncorrelated. We have essentially shown this in step (3) of the proof given earlier:  $C_{\mathbf{X}}$  is a nonnegative-definite symmetric matrix, and can thus be represented as

$$C_{\mathbf{X}} = B\Lambda B^T$$

for  $B$  orthogonal and  $\Lambda$  diagonal with nonnegative entries. Defining  $\Delta$  as before, we have

$$\mathbf{Y} = \Delta B^T(\mathbf{X} - \mathbf{m}), \quad \mathbf{X} = B\Delta^{-1}\mathbf{Y} + \mathbf{m}.$$

Hence  $E\mathbf{Y} = 0$ , and

$$C_{\mathbf{Y}} = \Delta B^T C_{\mathbf{X}} B \Delta^T = \Delta \Lambda \Delta^T = J,$$

where  $J$  is a diagonal matrix such that  $J_{kl} = 1$  if  $\Lambda_{kl} > 0$ ,  $J_{kl} = 0$  otherwise. Note that if  $\mathbf{X}$  is Gaussian, then  $\mathbf{Y}$  is also Gaussian with uncorrelated, hence *independent*, components. In the non-Gaussian case, the components of  $\mathbf{Y}$  need not be independent.

The choice of transformation  $A$  is not unique; the above construction in terms of  $B$ ,  $\Lambda$  and  $\Delta$  is just one standard method of *whitening*  $\mathbf{X}$ , i.e., expressing  $\mathbf{X}$  as an affine transformation of a zero-mean vector  $\mathbf{Y}$  with uncorrelated components. In the special case in which  $C_{\mathbf{X}}$  is positive definite (hence invertible), any matrix  $A$  such that

$$C_{\mathbf{X}} = AA^T$$

can yield a whitening transformation. Indeed,  $A$  is then also invertible, and if

$$\mathbf{Y} = A^{-1}(\mathbf{X} - \mathbf{m}),$$

we have

$$\mathbf{X} = A\mathbf{Y} + \mathbf{m}, \quad C_{\mathbf{Y}} = A^{-1}AA^T(A^{-1})^T = I.$$

Different choices of  $A$  include  $B\Lambda^{1/2}$  (coincides with  $B\Delta^{-1}$ ),  $B\Lambda^{1/2}B^T$  (symmetric), as well as the upper- and lower-diagonal forms that are obtained via Gram-Schmidt orthogonalization.

### The density of a Gaussian vector.

Let  $\mathbf{X} \in \mathbf{R}^n$  be a Gaussian random vector on  $(\Omega, \mathcal{F}, P)$  such that  $E\mathbf{X} = \mathbf{m}$ ,  $C_{\mathbf{X}} = Q$ . We seek an expression for the density of  $\mathbf{X}$  in  $\mathbf{R}^n$ .

(a)  $Q$  invertible. As in the previous subsection, we write

$$C_{\mathbf{X}} = AA^T$$

and consider the transformation  $\mathbf{Y} = A^{-1}(\mathbf{X} - \mathbf{m})$ . From the foregoing discussion we know that  $C_{\mathbf{Y}} = I$ , and hence the components of  $\mathbf{Y}$  are independent  $\mathcal{N}(0, 1)$  Gaussian variables. The density of  $\mathbf{Y}$  is then given by

$$f_{\mathbf{Y}}(\mathbf{y}) = \prod_{k=1}^n \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} = (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \mathbf{y}^T \mathbf{y} \right\}.$$

To derive the density of  $\mathbf{X}$ , we note that the probability of any  $n$ -dimensional rectangle  $H$  under  $P_{\mathbf{X}}$  is

$$\begin{aligned} P_{\mathbf{X}}(H) &= P_{\mathbf{Y}}(A^{-1}(H - \mathbf{m})) \\ &= \int \dots \int_{A^{-1}(H - \mathbf{m})} f_{\mathbf{Y}}(\mathbf{y}) dy_1 \dots dy_n. \end{aligned}$$

To evaluate the above integral, we use the invertible mapping

$$\mathbf{y} = A^{-1}(\mathbf{x} - \mathbf{m})$$

and invoke the rule for change of variables via Jacobians:

$$dy_1 \dots dy_n = |\det S| dx_1 \dots dx_n,$$

where the matrix-valued function  $S$  of  $\mathbf{x}$  is such that

$$S_{ij} = \frac{\partial y_i}{\partial x_j}.$$

In this case we simply have  $S = A^{-1}$ . We can therefore write

$$\begin{aligned} P_{\mathbf{X}}(H) &= \int \dots \int_H f_{\mathbf{Y}}(A^{-1}(\mathbf{x} - \mathbf{m})) |\det A^{-1}| dx_1 \dots dx_n \\ &= \int \dots \int_H (2\pi)^{-n/2} |\det A^{-1}| \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^T (A^{-1})^T A^{-1} (\mathbf{x} - \mathbf{m}) \right\} dx_1 \dots dx_n. \end{aligned}$$

Recalling that  $Q = AA^T$ , we obtain the final expression for the density of  $\mathbf{X}$ :

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-n/2}(\det Q)^{-1/2} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \mathbf{m})^T Q^{-1}(\mathbf{x} - \mathbf{m}) \right\} .$$

(b)  $Q$  non-invertible. In this case  $Q$  has at least one zero eigenvalue. If  $\mathbf{b}$  is an eigenvector corresponding to a zero eigenvalue, then

$$Q\mathbf{b} = \mathbf{0} ,$$

which implies that

$$\text{Var}(\mathbf{b}^T \mathbf{X}) = \mathbf{b}^T Q \mathbf{b} = 0 .$$

Thus the random variable  $\mathbf{b}^T \mathbf{X}$  is constant (equal to its expectation) with probability one, and the random vector  $\mathbf{X}$  lies on the *hyperplane*

$$\{\mathbf{x} : \mathbf{b}^T \mathbf{x} = \mathbf{b}^T \mathbf{m}\}$$

almost surely. Since hyperplanes are of dimension  $n - 1$ ,  $P_{\mathbf{X}}$  is a singular distribution on  $\mathbf{R}^n$  and  $f_{\mathbf{X}}$  does not exist.

In general, if exactly  $k$  eigenvalues of  $Q$  are equal to zero, then  $\mathbf{X}$  lies in a  $(n - k)$ -dimensional set which is the intersection of  $k$  orthogonal hyperplanes. This means that the randomness of  $\mathbf{X}$  is effectively “limited” to  $n - k$  components, from which the remaining  $k$  components via affine combinations. This is consistent with our earlier discussion on whitening: the random vector

$$\mathbf{Y} = \Delta B^T(\mathbf{X} - \mathbf{m})$$

will be such that

$$C_{\mathbf{Y}} = J ,$$

and will thus have only  $n - k$  truly random (and independent) components; the remaining components will be zero with probability one. Thus in writing

$$\mathbf{X} = A\mathbf{Y} + \mathbf{m} ,$$

we are effectively constructing  $\mathbf{X}$  using  $n - k$  random variables only.