

# MA8109, autumn 2007

## Measure and Integration Theory: the Basis of Probability Theory

Harald E. Krogstad, IMF.

Version: 9 September 2007

### 1 Introduction

*Measure and integration theory* was developed by Henri Leon Lebesgue and Emile Borel around 100 years ago. At that time, the theoretical founding of probability was rather shaky and almost 30 years passed until Andrei N. Kolmogorov put the probability theory on a sound foundation by merging probability and measure theory (of course, Kolmogorov was not the only one that worked on this, but his name is in the forefront). Since that time, advanced probability has always used the measure and integration theory language introduced by Kolmogorov (and others!).

For a student coming from introductory courses in probability to an advanced text/lecture using the measure and integration language this is a quite big (and difficult) step. Measure and integration theory is by itself an advanced mathematical topic which needs at least a one semester course. Moreover, the background for measure theory should preferably also include a course in linear analysis.

This informal note gives a survey of some of the key points, aiming to expand the even more condensed exposition in the first chapters of Øksendal's book.

**WARNING:** *It is important to note that the text does not always give appropriate references to the literature, and if this is the first time you see this material, the note is certainly not easy to read!*

Let us start by considering a uniformly distributed random variable  $\omega$  on the unit interval  $\Omega = [0, 1]$ . We know that the probability that  $\omega \in A = [a, b] \subset \Omega$  is

$$\Pr(\omega \in A) = \Pr(a \leq \omega \leq b) = b - a. \quad (1)$$

If  $A$  is a more complicated set, say a union of disjoint intervals,

$$A = \bigcup_{n=1}^N A_n, \quad (2)$$

where  $A_n = [a_n, b_n]$ , then

$$\Pr(\omega \in A) = \sum_{n=1}^N (b_n - a_n) = \sum_{n=1}^N \Pr(\omega \in A_n). \quad (3)$$

It turns out that it is possible to attach such a probability to almost all thinkable sets in  $\Omega$ , and this is a simple, but important example of what the mathematicians call a *measure*, or more specific, a *probability measure*.

If we have a function of  $\omega$ , say  $f : \Omega \rightarrow \mathbb{R}$ , the expectation of  $f(\omega)$  is equal to

$$E(f) = \int_{\Omega} f(\omega) d\omega. \quad (4)$$

Such integrals are of a more general kind than the Riemann integral we know from calculus, and the most famous construction of a more general integral is the so-called *Lebesgue integral*.

Below we shall consider in somewhat more detail how we can come up with "thinkable" sets and how it is possible to have a measure attached to them. This requires some set theory, which we turn to next.

## 2 Open sets

This section is an intermezzo before we introduce the more central concept of a  $\sigma$ -algebra.

Let  $A$  be a set on the real line  $\mathbb{R}$ . A point  $x$  is an *interior point* in  $A$  if there exists an open interval  $(a, b)$  such that

$$x \in (a, b) \subset A. \quad (5)$$

*A set on the real line is open if it contains only interior points.*

Hence, open intervals are open (!). Moreover, the empty set,  $\emptyset$ , is open by *definition*.

It may be proved that  $U \subset \mathbb{R}$  is open if and only if it can be written

$$U = \bigcup_{n=1}^N I_n, \quad N \leq \infty, \quad (6)$$

where  $\{I_n\}$  are disjoint (i.e.  $I_n \cap I_m = \emptyset$  when  $n \neq m$ ), open intervals. Thus, the open sets in  $\mathbb{R}$  consist of a disjoint union of open intervals.

If we return to the interval  $\Omega = [0, 1]$ , we say that a set " $U$  is open in  $\Omega$ " if there is an open set  $V \subset \mathbb{R}$  such that

$$U = V \cap \Omega. \quad (7)$$

Note therefore that  $[0, 1/2)$  is open in  $\Omega$ , but not as a set in  $\mathbb{R}$ . Thus,  $\{0\}$  is an interior point of  $[0, 1/2)$  if we fix our attention to  $[0, 1]$ , but *not* if we consider  $[0, 1/2)$  as a subset of  $\mathbb{R}$ !

It has turned out to be possible to introduce open sets in many different contexts, and we may even *define* what are the open sets from the very beginning. This has turned out to be quite useful in mathematics.

A set  $X$  with a family (collection)  $\mathcal{T}$  of open sets is called a *topological space* if  $\mathcal{T}$  satisfies

- (i)  $\emptyset \in \mathcal{T}$  and  $X \in \mathcal{T}$
- (ii)  $U, V \in \mathcal{T} \implies U \cap V \in \mathcal{T}$
- (iii)  $U_i \in \mathcal{T}$  for  $i \in I \implies \bigcup_{i \in I} U_i \in \mathcal{T}$

If this is the first time you see this, it looks cryptic, so note the following:

- In this case we have already decided what we mean by the open sets (which may be quite different from the open sets in  $\mathbb{R}$ !)
- *Space* here is used in a different meaning than in "vector space" or "linear space".
- In (iii), the index set  $I$  is arbitrary.
- On  $\mathbb{R}$  or  $\Omega$ , the collections of open sets make those topological spaces.

The reason topological spaces are useful is that they enable us to introduce convergence and define continuous functions.

Let  $X$  and  $Y$  be topological spaces and let  $f$  be a function from  $X$  to  $Y$ ,  $f : X \rightarrow Y$ . Then

$$f \text{ is } \textit{continuous} \iff f^{-1}(U) \text{ is open in } X \text{ for all open } U \text{ in } Y. \quad (8)$$

(The symbol  $\iff$  means "is defined by"). The notation  $f^{-1}(U)$  means

$$f^{-1}(U) = \{\omega ; f(\omega) \in U\}, \quad (9)$$

and *not* the inverse function (which requires that  $f$  is 1-1).

This definition of a continuous function looks strange, but reduces to the old  $\varepsilon$ - $\delta$ -definition when  $X = Y = \mathbb{R}$ .

The *complement* of a set  $U \subset X$  is all elements of  $X$  which are not in  $U$ . It is written  $U^C = X \setminus U$ .

The complements of the open sets are the *closed sets*. Note:

- *Some sets are neither open nor closed* (like the half-open intervals in  $\mathbb{R}$ ).
- *The whole space  $X$  is both open and closed!*

### 3 Sigma algebras and measures

A  $\sigma$ -algebra of sets,  $\mathcal{F}$ , is a collection of sets in  $X$  which satisfies

- (i)  $\emptyset \in \mathcal{F}$ ,
- (ii)  $A \in \mathcal{F} \implies A^C \in \mathcal{F}$ ,
- (iii)  $A_1, A_2, \dots, \in \mathcal{F} \implies A = \bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ .

The simplest  $\sigma$ -algebra we can have in  $X$  is  $\mathcal{F} = \{\emptyset, X\}$ . The largest possible  $\sigma$ -algebra is denoted  $2^X$  and is called the *power set*, that is, all possible subsets of  $X$ . We shall return to other examples later.

The collection  $\mathcal{T}$  of open sets considered in the previous section is *not* a  $\sigma$ -algebra (because of (ii)). However, we can have  $\sigma$ -algebras containing the open sets, and the *smallest*  $\sigma$ -algebra containing the open sets is called the *Borel  $\sigma$ -algebra*,  $\mathcal{B}$ . The formal definition of  $\mathcal{B}$  is as the intersection of all  $\sigma$ -algebras that contain the open sets,

$$\mathcal{B} = \cap \{ \mathcal{F}; \mathcal{F} \text{ is a } \sigma\text{-algebra containing the open sets} \}. \quad (10)$$

The sets in  $\mathcal{B}$  are called the *Borel sets*.  $\mathcal{B}$  contains (among others)

- all open sets
- all closed sets (because of (ii)!)
- *countable* unions (and intersections) of closed and open sets

*Countable* is an important concept in measure theory: Something is countable if it can be indexed by all or a finite set of the natural numbers,  $\{1, 2, 3, \dots\} = \mathbb{N}$ .

It is customary to say that  $\mathcal{B}$  is *closed under all countable set operations*.

Below we shall always use  $\mathcal{B}$  for the Borel sets when the underlying space and the open sets are obvious.

The countability is also the origin of the somewhat strange name  $\sigma$ -algebra. *Sigma* refers to the sum-sign, and means that we may write  $\sum_{n=1}^{\infty}$  or  $\cup_{n=1}^{\infty}$  etc. That is, we only have a *countable* number of elements in the sum, union, etc.

In the first section,  $X = \Omega = [0, 1]$ , and we introduced the probability  $P(A)$  for some sets in  $\Omega$ . This may be generalized to the Borel sets in  $\Omega$  by defining

$$P(A) = \inf_{A \subset U, U \text{ open}} P(U), \quad (11)$$

since the open sets were so simple in this case. This so-called *Lebesgue measure* has the (almost) obvious properties

- (i)  $0 \leq P(A) \leq 1$ ,
- (ii)  $P(\Omega) = 1$ ,  $P(\emptyset) = 0$ ,
- (iii) If  $A_1, A_2, \dots$  is a countable number of disjoint sets

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n). \quad (12)$$

The Lebesgue measure is defined similarly for the Borel sets on  $\mathbb{R}$  as well, but then  $P(A)$  may take any value from 0 to  $\infty$  (This definition of the Lebesgue measure is not quite the full story, but it is OK for us).

The Lebesgue measure is a prototype of measures. We shall actually need to introduce measures on more general  $\sigma$ -algebras than the Borel sets, but we shall not go into details how such measures may be defined (sometimes this is far from trivial). The measure has a number of useful properties, and we shall often need the following two:

- (1) Assume that  $A_1 \subset A_2 \subset \dots$  and that  $A = \cup_{n=1}^{\infty} A_n$ . Then  $P(A) = \lim_{n \rightarrow \infty} P(A_n)$ .
- (2) Assume that  $A_1 \supset A_2 \supset \dots$  and that  $A = \cap_{n=1}^{\infty} A_n$ . If  $P(A_1) < \infty$ , then  $P(A) = \lim_{n \rightarrow \infty} P(A_n)$ .

(If you wonder about the extra condition  $P(A_1) < \infty$  in (2): It is of course satisfied for probability measures. However, consider the following sets in  $\mathbb{R}$ :  $A_n = [n, \infty)$ ,  $n = 1, 2, \dots$ . What happens to  $A$  in this case?)

## 4 The Lebesgue integral

Let  $\Omega$  be any set and  $\mathcal{F}$  a  $\sigma$ -algebra in  $\Omega$ . Assume further that we have a measure  $m$  defined for the sets in the  $\sigma$ -algebra and satisfying

- (i)  $0 \leq m(A) \leq \infty$ ,
- (ii)  $m(\emptyset) = 0$ ,
- (iii) If  $A_1, A_2, \dots$  is a countable number of disjoint sets

$$m\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} m(A_n) \tag{13}$$

The triplet  $\{\Omega, \mathcal{F}, m\}$  is called a *measure space*.

Both  $[0, 1]$  and  $\mathbb{R}$  together with the Borel sets and the Lebesgue measure are measure spaces.

Let  $\{\Omega, \mathcal{F}, m\}$  be a measure space and  $f$  a function from  $\Omega$  to  $\mathbb{R}$ ,

$$f : \Omega \rightarrow \mathbb{R}. \tag{14}$$

Let as before  $\mathcal{B}$  be the Borel sets in  $\mathbb{R}$ . The function  $f$  is said to be  $\mathcal{F}$ -*measurable* if

$$f^{-1}(B) \in \mathcal{F} \text{ for all } B \in \mathcal{B} \subset \mathbb{R}. \tag{15}$$

Let  $A \in \mathcal{F}$ . The *indicator function*  $\chi_A : \Omega \rightarrow \mathbb{R}$  is defined as

$$\chi_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases} \tag{16}$$

This function is  $\mathcal{F}$ -measurable: Show that the only possibilities for  $\chi_A^{-1}(B)$  are  $\{\emptyset, \Omega, A, A^C\}$ , and all those are members of  $\mathcal{F}$ .

Sums, products and pointwise limits of measurable functions are also measurable. These are technical facts we shall not prove here.

It is also important to know that in order to check whether a function is measurable, it is actually enough to check that (15) holds for the open sets  $\mathcal{T} \subset \mathcal{B}$  (This is sometimes used as the definition of measurability, e.g. in Øksendal).

A *simple function* is a *finite* sum of indicator functions,

$$s = \sum_{i=1}^N a_i \chi_{A_i}, \quad N < \infty, \quad (17)$$

where the sets  $\{A_n\}$  are disjoint. The integral of a positive simple function is defined in a natural way as

$$\int_{\Omega} s dm = \sum_{i=1}^N a_i m(A_i). \quad (18)$$

Note that here we always replace  $0 \cdot \infty$  by 0.

Let  $s_1, s_2, \dots$  be a non-decreasing ( $s_n(x) \leq s_{n+1}(x)$ ) sequence of non-negative simple functions and let

$$f(x) = \lim_{n \rightarrow \infty} s_n(x) \quad (19)$$

(This limit exists for all  $x$  if we also include  $+\infty$ ).

The *Lebesgue integral* of  $f$  is defined

$$\int_{\Omega} f dm \triangleq \lim_{n \rightarrow \infty} \int_{\Omega} s_n dm \quad (20)$$

Note:

- $f$  is measurable
- the limit  $\lim_{n \rightarrow \infty} \int_{\Omega} s_n dm$  is either finite, or  $+\infty$ . Thus,  $\int_{\Omega} f dm = \infty$  is allowed.
- Any increasing sequence of simple functions converging pointwise to  $f$  leads to the same integral value (not so obvious, but true).

The Lebesgue integral exists for all non-negative measurable functions, but it may be  $+\infty$ .

Let now  $f$  be any measurable, real valued function and write

$$\begin{aligned} f(x) &= f^+(x) - f^-(x), \\ f^+(x) &= \max(0, f(x)), \\ f^-(x) &= -\min(0, f(x)). \end{aligned} \quad (21)$$

If  $\int f^+ dm$  and  $\int f^- dm < \infty$ , then  $f$  is called *Lebesgue-integrable* and

$$\int_{\Omega} f dm \triangleq \int_{\Omega} f^+ dm - \int_{\Omega} f^- dm. \quad (22)$$

Note that

$$\int_{\Omega} |f| dm = \int_{\Omega} f^+ dm + \int_{\Omega} f^- dm < \infty. \quad (23)$$

Similar definitions are also used for functions with values in  $\mathbb{C}$  and  $\mathbb{R}^n$ .

The functions' values on sets of measure 0 are of no importance in the Lebesgue theory since the contribution to the integral from these values would in any case be 0. It is customary to write that  $f$  and  $g$  are equal, apart from on a set of measure 0, as follows:

$$f = g \text{ a.e.} \iff f(x) = g(x) \text{ for all } x \in \Omega, \text{ apart from } x \in A, m(A) = 0.$$

a.e. = *almost everywhere*.

There are two BIG theorems about the Lebesgue integral:

**The Monotone Convergence Theorem:** Let  $f_1, f_2, f_3, \dots$ , be a non-decreasing ( $f_n(x) \leq f_{n+1}(x)$ ) sequence of non-negative functions. Then

$$\int \left( \lim_{n \rightarrow \infty} f_n \right) dm = \lim_{n \rightarrow \infty} \int f_n dm. \quad (24)$$

The proof is tricky and not included here (It is given in the note about the Lebesgue integral).

**The Dominated Convergence Theorem:** Let  $f_1, f_2, f_3, \dots$ , be a sequence of functions converging to a function  $f$ ,

$$f(x) = \lim_{n \rightarrow \infty} f_n(x) \text{ a.e.} \quad (25)$$

Assume that there exists a positive function  $h$  such that  $|f_n(x)| \leq h(x)$  for all  $n$  and that

$$\int h dm < \infty. \quad (26)$$

Then

$$\int \left( \lim_{n \rightarrow \infty} f_n \right) dm = \int f dm = \lim_{n \rightarrow \infty} \int f_n dm. \quad (27)$$

*Proof:* See the note about the Lebesgue integral.

The  $L^p$  spaces are important collections of Lebesgue integrable functions:

$$L^p(\Omega) = \left\{ f ; f \text{ is L.-integrable and } \int |f|^p dm < \infty \right\}, \quad 1 \leq p < \infty, \quad (28)$$

$$L^\infty(\Omega) = \{ f ; f \text{ is L.-integrable and } m\{x, |f(x)| > a\} = 0 \text{ for some } a < \infty \}. \quad (29)$$

The  $L^p$ -spaces are *Banach spaces* with norms defined as

$$\begin{aligned} \|f\|_p &= \left( \int |f|^p dm \right)^{1/p}, \quad 1 \leq p < \infty, \\ \|f\|_\infty &= \inf_a (m\{x, |f(x)| > a\} = 0). \end{aligned} \quad (30)$$

"Banach space" simply means that the  $L^p$ -spaces are linear vector spaces with the norm as the distance function. Moreover, limits of functions in  $L^p$  are also members of  $L^p$  (see note on the Lebesgue integral).

The simple functions are *dense* in  $L^p$  when  $1 \leq p < \infty$ . This means: If  $f \in L^p$  and  $\varepsilon > 0$ , then there is a simple function  $s$  so that

$$\|f - s\|_p < \varepsilon. \quad (31)$$

The most interesting spaces for us are  $p = 1$  and  $p = 2$ . The latter is a *Hilbert space*.

The monotone and dominated theorems are not the only theorems that could be mentioned here. Many times we will need to interchange the order of multiple integrals, and the big theorem about this is the *Fubini-Tonelli Theorem*. This theorem is actually two-in-one, where the two parts correspond to the two theorems above.

**Tonelli's Theorem:** *If all functions occurring below are measurable and non-negative,*

$$\iint_{X \times Y} f(x, y) dA = \int_X \left( \int_Y f(x, y) dy \right) dx = \int_Y \left( \int_X f(x, y) dx \right) dy. \quad (32)$$

(The values may all be equal to  $+\infty$ ).

**Fubini's Theorem:** *The conclusion in Tonelli's theorem holds for arbitrary measurable functions if one of the integrals, and hence all of them, has a finite value when  $f$  is replaced by  $|f|$ .*

The first integral in Eqn. 32 is a *double integral* defined on the so-called *product space* of  $X$  and  $Y$ . The others are called *iterated integrals* (one-dimensional integrals taken in sequence).

## 5 Random variables

We have now enough background to introduce some abstract probability theory.

Let  $\Omega$  be any set and  $\mathcal{F}$  a  $\sigma$ -algebra in  $\Omega$ . We say that a measure  $P$  is a *probability measure* on  $\mathcal{F}$  if  $P(\Omega) = 1$ .

The measure space triplet  $\{\Omega, \mathcal{F}, P\}$  is then called a *probability space*.

**Example 1:** The interval  $[0, 1]$ , the  $\sigma$ -algebra  $\mathcal{B}$ , and the Lebesgue measure,  $P_{Leb.}$ , makes up the probability space,  $\{[0, 1], \mathcal{B}, P_{Leb.}\}$ .

**Example 2:** Let  $\Omega = \mathbb{N} = \{1, 2, 3, \dots\}$  and  $\mathcal{F} = 2^{\mathbb{N}} =$  all possible subsets. We define the measure  $P$  as

$$P(\{1\}) = \frac{1}{2}, P(\{2\}) = \frac{1}{4}, \dots, P(\{n\}) = \frac{1}{2^n}, \dots \quad (33)$$

Check that this choice also gives a probability space!

A measurable function on a probability space is called a *random* (or *stochastic*) *variable*.

This random variable will turn out to be exactly what we know from before!



**Example:** Let  $Y$  be a standard Gaussian variable and  $\Phi(y) = \Pr(Y \leq y)$ . Let  $\Phi^{inv}$  be the inverse function,

$$\Phi^{inv} : [0, 1] \rightarrow \mathbb{R}. \quad (34)$$

If  $\omega$  is a uniform random variable on  $[0, 1]$ , then  $Y = \Phi^{inv}(\omega)$  is Gaussian. Thus, the function  $\Phi^{inv}$  defined on the probability space  $\{[0, 1], \mathcal{B}, P_{Leb}\}$  gives us a Gaussian random variable on  $\mathbb{R}$ . But is  $\Phi^{inv}$  really measurable? Yes, since it is continuous: In this case  $(\Phi^{inv})^{-1}(B)$  will be open for all open sets  $B \subset \mathbb{R}$ , and the open sets in  $[0, 1]$  are included in the Borel sets (in  $[0, 1]$ ) (Think about this for some time!).

This is the key difference between elementary (traditional) and advanced (modern) probability: In elementary probability, the random variables are "outcomes of experiments", whereas in advanced probability, the random variables are functions from some underlying probability space into the space of outcomes.

Elementary probability talks about random variables, but never mention that they are function values. However, who is picking the  $\omega$  for us?

In probability, random variables are usually written with capital letters. We shall do this here as well from now on.

Consider a random variable  $X$ . We can express statements about the outcomes, like

$$X \leq 0, 0 \leq X < 1, X > 10^6, X = 2, \dots, \quad (35)$$

as  $X \in B$  for some set  $B \in \mathcal{B} \subset \mathbb{R}$ . If we now think of  $X$  as a function  $X : \Omega \rightarrow \mathbb{R}$ , it is obvious that

$$\Pr(X \in B) = \Pr(X(\omega) \in B) = P(X^{-1}(B)). \quad (36)$$

For any  $X : \Omega \rightarrow \mathbb{R}$ , the  $\sigma$ -algebra generated by  $X$  is defined

$$\mathcal{H}_X \triangleq \{X^{-1}(B); B \in \mathcal{B} \subset \mathbb{R}\}. \quad (37)$$

If  $X$  is  $\mathcal{F}$ -measurable, then  $\mathcal{H}_X \subset \mathcal{F}$ .

*The  $\sigma$ -algebra  $\mathcal{H}_X$  represents the collection of all statements about or associated to  $X$ .*

By defining a measure  $\mu_X$  on the Borel sets in  $\mathbb{R}$  as

$$\mu_X(B) = P(X^{-1}(B)), \quad (38)$$

the triplet

$$\{\mathbb{R}, \mathcal{B}, \mu_X\} \quad (39)$$

*also becomes a probability space. The measure  $\mu_X$  is called the distribution of  $X$ .*

The *cumulative probability function* of  $X$ ,  $F_X(x)$ , is given by

$$F_X(x) = \mu_X((-\infty, x]). \quad (40)$$

Elementary probability	Advanced probability
Some "hidden mechanism"	Probability space $\{\Omega, \mathcal{F}, P\}$
$X$ is a random variable with values in $\mathbb{R}, \mathbb{Z}, \mathbb{R}^n, \mathbb{C}$ .	$X$ is a measurable function, $X : \Omega \rightarrow \mathbb{R}, \mathbb{Z}, \mathbb{R}^n, \mathbb{C}$ .
$E(X)$ exists ( $E( X  < \infty)$ )	$X \in L^1(\Omega, \mathcal{F}, P)$
$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$	$E(X) = \int_{\Omega} X(\omega) dP(\omega)$
$E( X ^2) < \infty$	$X \in L^2(\Omega, \mathcal{F}, P)$
$\text{Var}(X) = E((X - EX)^2)$	$\text{Var}(X) = \ X - EX\ _2^2$
Events connected with $X$	The $\sigma$ -algebra generated by $X$

Table 1: The correspondence between elementary and advanced probability theory.

If  $X$  has a nice distribution function,  $\mu_X$  may be expressed by the *probability density*,  $f_X$  of  $X$ ,

$$\mu_X(B) = \int_B f_X(x) dx. \quad (41)$$

Similar definitions also apply for stochastic variables with values in  $\mathbb{R}^n$  or  $\mathbb{C}$ .

We shall now consider integrals of random variables, and all these integrals will be Lebesgue integrals. First of all, if the random variable is *integrable*, that is  $\int_{\Omega} |X(\omega)| dP(\omega) < \infty$ , the *expectation of  $X$*  is defined as

$$E(X) = \int_{\Omega} X(\omega) dP(\omega). \quad (42)$$

It turns out (and in fact not so difficult to see) that we also have

$$E(X) = \int_{\mathbb{R}} x d\mu_X(x), \quad (43)$$

which, for nice variables, is just what we already know,

$$E(X) = \int_{\mathbb{R}} x f_X(x) dx. \quad (44)$$

This is the general pattern: *The abstract definitions always boil down to the well-know in the classical cases!*

A word about notation: When we are integrating a function  $f$  defined on a probability space  $\{\Omega, \mathcal{F}, P\}$ , it is common to write this in the short way as

$$\int f dP. \quad (45)$$

Thus, if there is no need to indicate  $\omega$  explicitly, it is omitted.

Some of the correspondences between elementary and advanced probability are summarized in Table 1.

**Example:** Consider a simple random variable  $X$  with just two outcomes,  $a$  and  $b$ . Set  $A = \{\omega ; X(\omega) = a\}$ . Then  $A^C = \{\omega ; X(\omega) = b\}$ . The only choices for  $X^{-1}(B)$  are  $\{\emptyset, \Omega, A, A^C\}$  and  $\mathcal{H}_X = \{\emptyset, \Omega, A, A^C\}$ .

The rule is: *Simple functions – simple  $\sigma$ -algebras!*

More generally, the  $\sigma$ -algebra generated by a collection of sets  $\mathcal{U}$  is equal to the intersection of all  $\sigma$ -algebras that contain  $\mathcal{U}$ . In this way, we may have  $\sigma$ -algebras generated by a set of random variables.

The following proposition is a consequence of the properties of measurable functions:

**Proposition 1:** *Let  $\mathcal{H}$  be a  $\sigma$ -algebra contained in  $\mathcal{F}$ . Then  $L^p(\Omega, \mathcal{H}, P)$  is a closed subspace of  $L^p(\Omega, \mathcal{F}, P)$  for  $1 \leq p \leq \infty$ .*

*Idea of proof:*

- Since  $X$  is  $\mathcal{H}$ -measurable, it is also  $\mathcal{F}$ -measurable since  $\mathcal{H} \subset \mathcal{F}$ .
- Measurability is closed under linear operations.
- If  $\{X_n\}$  is a sequence in  $L^p(\Omega, \mathcal{H}, P)$  converging to an element  $X$  in  $L^p(\Omega, \mathcal{F}, P)$ , there is a subsequence converging pointwise to  $X$ , that is,

$$X_{n_k}(\omega) \xrightarrow[k \rightarrow \infty]{} X(\omega) \text{ a.e.} \quad (46)$$

(This statement is general  $L^p$ -theory). Then  $X$  is  $\mathcal{H}$ -measurable and a member of  $L^p(\Omega, \mathcal{H}, P)$ .

We shall finally note a property we often use in proofs and various arguments

**Proposition 2:** *Let  $f$  and  $g$  be two  $\mathcal{H}$ -measurable functions, and assume that*

$$\int_H f dP = \int_H g dP \text{ for all } H \in \mathcal{H}.$$

*Then  $f = g$  a.e.*

*Proof:* The integral is linear, so that it is enough to prove that  $f = 0$  a.e. if  $\int_H f dP = 0$  for all  $H \in \mathcal{H}$ . Let  $A = \{\omega ; f(\omega) > 0\}$  and  $A_n = \{\omega ; f(\omega) > 1/n\}$ . Then  $A_1 \subset A_2 \subset \dots$  and  $A = \cup_{n=1}^{\infty} A_n$ . But  $A_n \in \mathcal{H}$  and since

$$0 = \int_{A_n} f dP \geq P(A_n) \frac{1}{n},$$

we must have that  $P(A_n) = 0$ . Hence,  $P(A) = \lim_{n \rightarrow \infty} P(A_n) = 0$ . A similar argument also applies for  $B = \{\omega; f(\omega) < 0\}$ .

## 6 Independence

*Independence* is probability's main contribution to the programme carried out by Kolmogorov. In fact, we could write

$$\text{"Independence + Measure Theory = Probability Theory!"} \quad (47)$$

In elementary probability we say that

$$A \text{ is independent of } B \stackrel{\Delta}{\iff} \Pr(A \cap B) = \Pr(A) \Pr(B), \quad (48)$$

and this definition of independence is kept for two sets  $A$  and  $B$  in the  $\sigma$ -algebra  $\mathcal{F}$  of a probability space,  $(\Omega, \mathcal{F}, P)$ ,

$$A \text{ is independent of } B \stackrel{\Delta}{\iff} P(A \cap B) = P(A) P(B). \quad (49)$$

Two  $\sigma$ -algebras are independent if all pairs of sets from the respective algebras are independent. When it comes to pairs of random variables defined on  $\Omega$ , the definition is therefore

$$X \text{ is independent of } Y \stackrel{\Delta}{\iff} \text{All pairs from } \mathcal{H}_X \text{ and } \mathcal{H}_Y \text{ are independent.}$$

In elementary probability, the corresponding definition is that the joint distribution of  $X$  and  $Y$  can be factorized, and this is now a *proposition* following from the definition.

Check out that if  $A$  and  $B$  are independent, then so are  $A^C$  and  $B$  (and hence  $A$  and  $B^C$ ), as well as  $A^C$  and  $B^C$ . Thus, convince yourself that  $A$  and  $B$  are independent if and only if  $\mathcal{H}_{\chi_A}$  and  $\mathcal{H}_{\chi_B}$  are independent.

Now consider the probability space  $(\Omega, \mathcal{F}, P)$  and an arbitrary collection of  $\sigma$ -algebras contained in  $\mathcal{F}$ . The collection consists of independent  $\sigma$ -algebras if all finite selections of sets  $S$ ,  $S = \{B_n\}_{n=1}^N$ , containing at most one set from each algebra, satisfy

$$P(B_1 \cap B_2 \cap \dots \cap B_N) = \prod_{n=1}^N P(B_n).$$

Observe that pairwise independence is not sufficient. In particular, an arbitrary collection  $\{X_i\}_{i \in I}$  of random variables defined on  $\Omega$  consists of independent random variables if the corresponding  $\sigma$ -algebras  $\{\mathcal{H}_{X_i}\}_{i \in I}$  are independent.

It turns out that the most convenient definition of independence for an arbitrary collection of events  $\{A_i\}_{i \in I}$  is to say that they are independent if the  $\sigma$ -algebras of their respective indicator functions,  $\{\mathcal{H}_{\chi_{A_i}}\}$ , are independent. This implies that for an arbitrary finite subset of  $\{A_i\}_{i \in I}$ , say  $\{A_n\}_{n=1}^N$ , we have

$$P(A_1 \cap A_2 \cap \dots \cap A_N) = \prod_{n=1}^N P(A_n) \quad (50)$$

as well as

$$P\left(A_1^{(C)} \cap A_2^{(C)} \cap \dots \cap A_N^{(C)}\right) = \prod_{n=1}^N P\left(A_n^{(C)}\right), \quad (51)$$

where  $(C)$  means taking the complement, - or not.

However, it may be proved that  $\{A_i\}_{i \in I}$  are indeed independent if an arbitrary finite subset of  $\{A_i\}_{i \in I}$  satisfies 50, and this is the most common definition of independence found in the

literature. Since the first definition clearly implies the second, they are, as they should be, equivalent.

**Example:** Consider  $\Omega = [0, 1]$  and the functions

$$\begin{aligned} X(\omega) &= 1, \\ Y(\omega) &= \begin{cases} 1, & 0 \leq \omega \leq 1/2, \\ -1 & 1/2 < \omega \leq 1. \end{cases} \end{aligned} \quad (52)$$

Check that  $\mathcal{H}_X = \{\Omega, \emptyset\}$  and  $\mathcal{H}_Y = \{\Omega, \emptyset, [0, 1/2], (1/2, 1]\}$ , and that all pairs from these two  $\sigma$ -algebras are independent!

*Note that a constant random variable is independent of everything!*

Independence will be a quite important concept for us in the following.

**Example:** In Øksendal, *Edition 5*, it is stated (and also considered in Exercise 2.5) that if  $\int_{\Omega} |XY| dP$ ,  $\int_{\Omega} |X| dP$ , and  $\int_{\Omega} |Y| dP$  are finite, and  $X$  and  $Y$  independent, then  $E(XY) = E(X) \cdot E(Y)$ , that is,

$$\int_{\Omega} XY dP = \int_{\Omega} X dP \cdot \int_{\Omega} Y dP. \quad (53)$$

In *Edition 6*, the condition  $\int_{\Omega} |XY| dP < \infty$  has been removed.

In general,  $\int_{\Omega} |XY| dP$  may be infinite even if  $\int_{\Omega} |X| dP$ , and  $\int_{\Omega} |Y| dP < \infty$  (consider  $X(\omega) = Y(\omega) = \omega^{-1/2}$  on  $[0, 1]$ !), but when  $X$  and  $Y$  are independent, this can not happen, and the condition  $\int_{\Omega} |XY| dP < \infty$  included in *Edition 5* is in fact superfluous. The proof of the identity in Eqn. 53 for bounded simple functions is indicated in Exercise 2.5, and this may be applied for arbitrary positive functions  $X$  and  $Y$  by applying the Monotone Convergence Theorem. For the general case, we write, as before,  $X = X^+ - X^-$  and observe that  $\mathcal{H}_{X^+} \subset \mathcal{H}_X$  since

$$(X^+)^{-1}(B) = X^{-1}(B \cap [0, \infty]) \subset \mathcal{H}_X. \quad (54)$$

The same applies to  $X^-$  and  $Y = Y^+ - Y^-$ . Thus,  $X^+$  and  $X^-$  are independent of  $Y^+$  and  $Y^-$ . Having already proved the relation for positive variables, we then obtain

$$\begin{aligned} E[XY] &= E[(X^+ - X^-)(Y^+ - Y^-)] \\ &= E[X^+Y^+ - Y^-X^+ - X^-Y^+ + X^-Y^-] \\ &\dots \text{fill in!} \dots \\ &= E(X) \cdot E(Y). \end{aligned} \quad (55)$$

In probability theory there is one little lemma which may be used for a lot of arguments (in fact, it should have been promoted to a Theorem long time ago!). This is the *Borel-Cantelli Lemma*. It is used several places in Øksendal's text and it is included here for a reference.

Let us consider the events  $\omega \in A_n$  for  $n = 1, 2, 3, \dots$ , and the event that  $\omega \in A_n$  for infinitely many  $n$ -s. This set may be written

$$A_{\infty} = \{\omega; \omega \in A_n \text{ for infinitely many } n\} = \bigcap_{n=1}^{\infty} \left( \bigcup_{k=n}^{\infty} A_k \right). \quad (56)$$

Note what this expression says: Regardless where we start to look (say at set number  $n$ ),  $\omega$  is in at least *some* of the following sets. Using the definition of  $A_\infty$ , the lemma goes as follows:

**Borel-Cantelli Lemma:**

(a) If  $\sum_{n=1}^{\infty} P(A_n) < \infty$ , then  $P(A_\infty) = 0$ .

(b) If all  $A_n$  are independent and  $\sum_{n=1}^{\infty} P(A_n) = \infty$ , then  $P(A_\infty) = 1$ .

(Note: Independence is crucial in (b). Just consider the case where  $A_n = A$  for all  $n$ -s, where  $0 < P(A) < 1$ ).

*Proof of (a):* Consider the function  $s = \sum_{n=1}^{\infty} \chi_{A_n}$  and note that  $\|s\|_1 = \sum_{n=1}^{\infty} P(A_n) < \infty$ . Since  $s$  is integrable, it has to be finite a.e., and  $s(\omega) = \#$  times  $\omega$  is in some  $A_n$ .

*Proof of (b):* This is somewhat tricky and requires the property of the measure listed above: If  $B_1 \supset B_2 \supset \dots$  and  $P(B_1)$  is finite (as it is here), then  $P(\cap_{n=1}^{\infty} B_n) = \lim_{n \rightarrow \infty} P(B_n)$ . It also requires an elementary result about infinite products: If  $0 < x_n < 1$ , then  $\prod_{n=1}^{\infty} (1 - x_n) = 0 \iff \sum_{n=1}^{\infty} x_n = \infty$ . Armed with these tools, the argument goes as follows (with  $B_n = \cup_{k=n}^{\infty} A_k$ ):

$$P\left(\bigcup_{k=n}^{\infty} A_k\right) = 1 - P\left(\bigcap_{k=n}^{\infty} A_k^C\right) \stackrel{\text{indep.}}{=} 1 - \prod_{k=n}^{\infty} P(A_k^C) = 1 - \prod_{k=n}^{\infty} (1 - P(A_k)) = 1, \quad (57)$$

since  $\sum_{k=n}^{\infty} P(A_k) = \infty$ .

(Do not worry if you do not grab this the first time you see it! The proof is not so essential for us).

## 7 Conditional expectation

*Conditional expectation* is another rather difficult concept in advanced probability. This section is a little/rather heavy, and you should not expect to understand everything at once. During the course we will return to this material several times.

You may remember the *conditional probability*, which typically is written  $P(A|B)$  and read *the conditional probability of A given the event B*. The conditional probability is computed using *Bayes Rule*,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (58)$$

(the rule is simple to visualize using diagrams). The event  $B$  represents some kind of knowledge we have and which influences the probability we are looking for. Similarly, the *conditional expectation of a random variable X given B* could be defined as

$$\frac{\int_B X dP}{P(B)}. \quad (59)$$

Let  $\{H_i\}_{i=1}^N \subset \mathcal{F}$  be a finite partition of  $\Omega$ , that is,  $\Omega = \cup_{i=1}^N H_i$ ,  $H_i \cap H_j = \emptyset$  for  $i \neq j$ , and  $P(H_i) > 0$ ,  $i = 1, \dots, N$ . This set defines a finite  $\sigma$ -algebra  $\mathcal{H}_N \subset \mathcal{F}$  consisting of all

possible unions formed from the sets; see B.Ø., Exercise 2.7. We may generalize the concept of conditional expectation of  $X$ , given the  $\sigma$ -algebra  $\mathcal{H}_N$ , by introducing the simple function,

$$Y = \sum_{i=1}^N \frac{\int_{H_i} X dP}{P(H_i)} \chi_{H_i}. \quad (60)$$

It is easy to verify that

$$\int_H Y dP = \int_H X dP \quad (61)$$

for all  $H \in \mathcal{H}_N$ . Moreover, since  $Y$  is constant on each of the  $H_i$ -sets, it is in fact  $\mathcal{H}_N$ -measurable (alternatively, it is  $\mathcal{H}_N$ -measurable since it is a sum of  $\mathcal{H}_N$ -measurable indicator functions). As observed in Section 5, if the integrals of two  $\mathcal{H}_N$ -measurable functions coincide on all sets in  $\mathcal{H}_N$ , they are equal a.e. Thus, any other  $\mathcal{H}_N$ -measurable function satisfying Eqn. 61 must be equal to  $Y$  a.e.

It turns out that the simple finite situation above may be extended to a general  $\sigma$ -algebra  $\mathcal{H} \subset \mathcal{F}$ . Assume that we have a  $\sigma$ -algebra  $\mathcal{H}$  and a function  $X \in L^1(\Omega, \mathcal{F}, P)$ . The integral of  $X$  over a set  $H \in \mathcal{H}$  will be a measure on  $\mathcal{H}$  (that is, a mapping from  $\mathcal{H}$  into the numbers  $\mathbb{R}, \mathbb{R}^n, \mathbb{C}$ , etc. depending on  $X$ ):

$$Q(H) = \int_H X dP. \quad (62)$$

This measure has the property that if  $P(H) = 0$ , then  $Q(H) = 0$  (why?). Thus, the measure  $Q$  on  $\mathcal{H}$  is what is called *absolutely continuous* with respect to  $P$  restricted to  $\mathcal{H}$ . A famous theorem in measure theory (The Radon-Nikodym Theorem) then says that there exist a unique function which is called the *conditional expectation* and written

$$E(X|\mathcal{H}), \quad (63)$$

such that

1.  $E(X|\mathcal{H})$  is  $\mathcal{H}$ -measurable,
2.  $\int_H E(X|\mathcal{H}) dP = \int_H X dP$  for all  $H \in \mathcal{H}$ .

The function is only unique up to values on a set of measure 0. It has all the properties of  $Y$  above, but there is in general no partition of sets on which  $Y$  takes constant values. The notation is strange, but so standard that it is impossible to change. It is confusing that an expectation is a function, but the notation is used everywhere.

**Example 1:** The simplest possible  $\sigma$ -algebra is  $\mathcal{H} = \{\emptyset, \Omega\}$  and the only  $\mathcal{H}$ -measurable functions are the constants. Check that it works to set  $E(X|\mathcal{H}) = E(X)$ ! This is logical: If we only have the trivial (or no!) other information, the conditional expectation is the same as the ordinary expectation.

**Example 2:** If  $X$  itself is  $\mathcal{H}$ -measurable,  $E(X|\mathcal{H}) = X$  a.e. (since  $X$  satisfies the definition and  $E(X|\mathcal{H})$  is unique).

**Example 3:**

We recover Bayes Rule if we consider the simple case that  $X = \chi_A$  and  $\mathcal{H} = \{\emptyset, \Omega, B, B^C\}$ , where  $0 < P(B) < 1$ . In this case  $Y = E(X|\mathcal{H})$  must satisfy the four equations

$$\begin{aligned} \int_B Y dP &= \int_B \chi_A dP = P(A \cap B), \\ \int_{B^C} Y dP &= \int_{B^C} \chi_A dP = P(A \cap B^C), \\ \int_\Omega Y dP &= \int_\Omega \chi_A dP = P(A), \\ \int_\emptyset Y dP &= \int_\emptyset \chi_A dP = 0. \end{aligned} \tag{64}$$

The function

$$Y(\omega) = \begin{cases} \frac{P(A \cap B)}{P(B)}, & \omega \in B, \\ \frac{P(A \cap B^C)}{P(B^C)}, & \omega \in B^C. \end{cases} \tag{65}$$

is  $\mathcal{H}$ -measurable, satisfies the equations, and fits the definition.

Conditional expectation has many (more or less!) obvious properties. Most properties below are also listed in Øksendal.

**Property 1:** *The operation  $X \rightarrow E(X|\mathcal{H})$  is linear, that is,  $E(aX + bY|\mathcal{H}) = aE(X|\mathcal{H}) + bE(Y|\mathcal{H})$ .*

*Idea of proof:*  $\int_H E(aX + bY|\mathcal{H}) dP = \int_H (aX + bY) dP = \text{fill in!} = \int_H (aE(X|\mathcal{H}) + bE(Y|\mathcal{H})) dP$ .

**Property 2:** *If  $X$  (or rather  $\mathcal{H}_X$ ) and  $\mathcal{H}$  are independent, then  $E(X|\mathcal{H}) = EX$  ( $\mathcal{H}$  provides no useful information!).*

*Proof:*

For any  $H \in \mathcal{H}$ ,

$$\begin{aligned} \int_H E(X|\mathcal{H}) dP &= \int_H X dP = \int_\Omega X \chi_H dP \\ &= \int_\Omega X dP \int_\Omega \chi_H dP \\ &= E(X) \cdot P(H) = \int_H (EX) dP. \end{aligned}$$

Note that this argument requires Eqn. 53.

**Property 3:** *If  $Y$  is  $\mathcal{H}$ -measurable, then*

$$E(YX|\mathcal{H}) = YE(X|\mathcal{H}). \tag{66}$$

*Idea of proof:* First prove it when  $Y$  is a simple  $\mathcal{H}$ -measurable function. The general result follows from a limiting argument also given in Øksendal.

**Property 4:** *If  $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$ , then  $E(X|\mathcal{G}) = E(E(X|\mathcal{H})|\mathcal{G})$ .*



*Idea of proof:* Let  $G \in \mathcal{G}$ , then

$$\int_G E(X|\mathcal{G}) dP = \int_G X dP = \int_G E(X|\mathcal{H}) dP = \int_G E(E(X|\mathcal{H})|\mathcal{G}) dP \quad (67)$$

since  $G \in \mathcal{G} \subset \mathcal{H}$ . Since both  $E(X|\mathcal{G})$  and  $E(E(X|\mathcal{H})|\mathcal{G})$  are  $\mathcal{G}$ -measurable, the equality of the functions follows from Proposition 2 at the end of the previous chapter.

**Property 5:** If  $X(\omega) \geq 0$  a.e., then  $E(X|\mathcal{H})(\omega) \geq 0$  a.e.

*Proof:* Let  $A = \{\omega; E(X|\mathcal{H})(\omega) < 0\}$ :

$$0 \geq \int_A E(X|\mathcal{H}) dP = \int_A X dP \geq 0. \quad (68)$$

Thus,  $\int_A E(X|\mathcal{H}) dP = 0$ . Note that  $A = \cup_{n=1}^{\infty} A_n$  where  $A_n = \{\omega; E(X|\mathcal{H})(\omega) < -1/n\}$ . However,  $P(A_n)$  has to be 0 for all  $n$ -s since

$$0 \leq \int_{A_n} X dP = \int_{A_n} E(X|\mathcal{H}) dP \leq -\frac{P(A_n)}{n}. \quad (69)$$

But  $P(A_n) \rightarrow P(A)$  by the properties of the measure.

**Property 6:**  $|E(X|\mathcal{H})(\omega)| \leq E(|X||\mathcal{H})(\omega)$  a.e.

*Proof:* Write  $X = X^+ - X^-$  and use Property 5:

$$\begin{aligned} |E(X|\mathcal{H})(\omega)| &= |E(X^+|\mathcal{H})(\omega) - E(X^-|\mathcal{H})(\omega)| \\ &\leq E(X^+|\mathcal{H})(\omega) + |E(X^-|\mathcal{H})(\omega)| \end{aligned} \quad (70)$$

$$= E(|X||\mathcal{H})(\omega). \quad (71)$$

The following paragraph requires some knowledge of functional analysis.

Conditional expectation has an interesting connection with the *best approximation*. In fact,  $E(X|\mathcal{H})$  is the best approximation of  $X$  given the information contained in  $\mathcal{H}$ . The  $L^2$ -case is actually quite simple to verify:

**Property 7:** Let  $X \in L^2(\Omega, \mathcal{F}, P)$  and  $\mathcal{H} \subset \mathcal{F}$ . Then  $E(X|\mathcal{H})$  is the best approximation in (i.e. the projection onto)  $L^2(\Omega, \mathcal{H}, P)$ .

*Proof:* Let  $M : L^2(\Omega, \mathcal{F}, P) \rightarrow L^2(\Omega, \mathcal{H}, P)$  be the projection operator (recall that  $L^2(\Omega, \mathcal{H}, P)$  is a closed subspace of  $L^2(\Omega, \mathcal{F}, P)$ !). Since  $MX$  is  $\mathcal{H}$ -measurable, we only have to verify property 2 in the definition:

$$\begin{aligned} \int_H (MX) dP &= \int_{\Omega} \chi_H (MX) dP \\ &= \langle \chi_H, MX \rangle = \langle M\chi_H, X \rangle \\ &= \langle \chi_H, X \rangle = \int_H X dP. \end{aligned} \quad (72)$$

Hence,  $X \rightarrow E(X|\mathcal{H})$  is a continuous, linear mapping in  $L^2$ . However, it turns out that the mapping  $X \rightarrow E(X|\mathcal{H})$  is *not continuous* in  $L^1$  (this is not obvious!). Nevertheless, if  $X_n(\omega) \rightarrow X(\omega)$  a.e. and  $|X_n(\omega)| \leq h(\omega)$ ,  $h \in L^1$ , then  $E(X_n|\mathcal{H}) \rightarrow E(X|\mathcal{H})$  a.e.

## 8 Characteristic Functions

The *characteristic function* of a (univariate) stochastic variable  $X$  is defined as

$$\phi_X(u) = E(\exp(iuX)). \quad (73)$$

For a multivariate stochastic variable,  $\mathbf{X} = (X_1, \dots, X_n)'$ , the characteristic function is a function of  $n$  variables  $\mathbf{u} = (u_1, \dots, u_n) \in \mathbb{R}^n$  defined as

$$\phi_{\mathbf{X}}(\mathbf{u}) = E(\exp(i\mathbf{u}'\mathbf{X})) = \int_{\mathbb{R}^n} e^{i\mathbf{u}'\mathbf{x}} d\mu_{\mathbf{X}}(\mathbf{x}). \quad (74)$$

This is thus the *Fourier transform* of the distribution of the variable. Since  $|\exp(i\mathbf{u}'\mathbf{X})| = 1$  for real variables, *the characteristic function always exists*. If  $\mathbf{X}$  has a probability density,  $f_{\mathbf{X}}(\mathbf{x})$ ,

$$\phi_{\mathbf{X}}(\mathbf{u}) = E(\exp(i\mathbf{u}'\mathbf{X})) = \int_{\mathbb{R}^n} e^{i\mathbf{u}'\mathbf{x}} f_{\mathbf{X}}(\mathbf{x}) d^n x. \quad (75)$$

Moreover, if  $\phi_{\mathbf{X}}(\mathbf{u})$  is an integrable function in  $\mathbb{R}^n$ , then, by the *inverse Fourier transform*,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-i\mathbf{u}'\mathbf{x}} \phi_{\mathbf{X}}(\mathbf{u}) d^n u. \quad (76)$$

(Note: It is customary to distinguish between  $\mathbb{R}^n$  for  $\mathbf{x}$  and  $\mathbf{u}$ , and you will often see the notation  $\mathbf{u} \in \hat{\mathbb{R}}^n$ ).

The characteristic function is a valuable tool when dealing with probability distributions. One particular feature is an elegant way to obtain the expectation of combination of the variables when we have an analytic expression for  $\phi_{\mathbf{X}}(\mathbf{u})$ . In particular, we observe that

$$\frac{\partial \phi_{\mathbf{X}}}{\partial u_j}(\mathbf{0}) = \left. \frac{\partial E(\exp(i\mathbf{u}'\mathbf{X}))}{\partial u_j} \right|_{\mathbf{u}=\mathbf{0}} = E(iX_j) = iEX_j \quad (77)$$

(the interchange of the derivative and  $E$  needs some justification we shall not go into). We therefore find the expectations by taking the derivative of the characteristic function, instead of an integration over the distribution for the variable. In general, this leads to the *very* useful formula

$$E(X_1^{n_1} X_2^{n_2} \dots X_N^{n_N}) = i^{-\sum_j n_j} \left. \frac{\partial^{\sum_j n_j}}{\partial u_1^{n_1} \partial u_2^{n_2} \dots \partial u_N^{n_N}} \phi_{\mathbf{X}}(\mathbf{u}) \right|_{\mathbf{u}=\mathbf{0}}. \quad (78)$$

Often it is convenient to write  $\phi_{\mathbf{X}}(\mathbf{u})$  as a Taylor series expansion before the derivatives are computed. The method is particularly efficient for Gaussian variables which we consider next.

## 9 Multivariate Gaussian Variables

Gaussian variables will play a central part throughout the course, and if multivariate Gaussian variables are completely new to you, you should look at one of the standard textbooks in Statistics, *e.g.* R. Johnson and D. Wichern: *Applied statistical analysis*, Prentice Hall.

We recall that a stochastic variable  $X$  is Gaussian if it has a density of the form

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

It follows by integration that  $EX = \mu$  and  $\text{Var } X = \sigma^2$ . The characteristic function of  $X$  is

$$\phi_X(u) = \exp\left(iu\mu - \frac{1}{2}\sigma^2u^2\right) \quad (79)$$

(You should try to compute integral in Eqn. 75. You may also consult a table of Fourier transforms).

A multivariate stochastic variable  $\mathbf{X} = (X_1, \dots, X_n)'$  is said to be *multivariate Gaussian* if all linear combinations of the form

$$Y = \sum_{i=1}^n \alpha_i X_i, \quad \alpha_i \in \mathbb{R}, \quad (80)$$

are Gaussian. This implies, in particular, that all components separately are Gaussian.

**Proposition:** *If  $\mathbf{X}$  is multivariate Gaussian, the characteristic function has the form*

$$\phi_{\mathbf{X}}(\mathbf{u}) = \exp\left(i\mathbf{u}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}\right). \quad (81)$$

**Proof:** Let  $Y = \sum_{i=1}^n u_i X_i = \mathbf{u}'\mathbf{X}$ . Then  $E(Y) = \sum_{i=1}^n u_i \mu_i = \mathbf{u}'\boldsymbol{\mu}$ , and

$$\begin{aligned} \text{Var } Y &= E(Y - \boldsymbol{\mu})^2 \\ &= E\left(\sum_{i,j=1}^n u_i u_j (X_i - \mu_i)(X_j - \mu_j)\right) \\ &= \sum_{i,j=1}^n u_i u_j \sigma_{ij} = \mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}, \end{aligned} \quad (82)$$

where  $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$ ,  $\sigma_{ij} = \text{Cov}(X_i, X_j)$ . The result follows by observing

$$\begin{aligned} \phi_{\mathbf{X}}(\mathbf{u}) &= E(\exp(i\mathbf{u}'\mathbf{X})) = E(\exp(iY)) \\ &= \phi_Y(1) = \exp\left(iEY - \frac{1}{2}\text{Var } Y\right) \\ &= \exp\left(i\mathbf{u}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}\right). \end{aligned} \quad (83)$$

In fact, if  $\mathbf{X} = (X_1, \dots, X_n)'$  has finite means and variances and the characteristic function has the form in Eqn. 81, then  $\mathbf{X}$  is multivariate Gaussian (See Øksendal, Theorem A.5).

If  $\boldsymbol{\Sigma}$  is non-singular it is positive definite, and hence

$$\lambda_{\min} |\mathbf{u}|^2 \leq \mathbf{u}'\boldsymbol{\Sigma}\mathbf{u} \quad (84)$$

for the smallest eigenvalue  $\lambda_{\min} > 0$ . Thus,  $\phi_{\mathbf{X}}$  will be integrable and  $\mathbf{X}$  has a density

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{-i\mathbf{u}'\mathbf{x}} \exp\left(i\mathbf{u}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}\right) d^n u \\ &= \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \end{aligned} \quad (85)$$

(This integral may be a bit tricky: Introduce  $\mathbf{u} = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$  and note that the new integral with respect to  $\mathbf{y}$  splits into a product of  $n$  similar one-dimensional integrals. Then use the result for the one-dimensional distribution).

Let us now assume that  $\mathbf{X}$  is a zero mean multivariate Gaussian variable with covariance matrix  $\boldsymbol{\Sigma}$  so that

$$\phi_{\mathbf{X}}(\mathbf{u}) = \exp\left(-\frac{\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}}{2}\right). \quad (86)$$

For a single variable  $X$  with standard deviation 1, we obtain

$$E\left(X^k\right) = -i^{-k} \frac{\partial^k \phi_X}{\partial u^k}(0) = -i^{-k} \frac{\partial^k e^{-\frac{u^2}{2}}}{\partial u^k}(0) = \begin{cases} 0, & k \text{ odd} \\ 1 \cdot 3 \cdots (k-1), & k \text{ even} \end{cases} \quad (87)$$

For a bivariate zero mean Gaussian variable with

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho \\ \rho & \sigma^2 \end{bmatrix}, \quad (88)$$

the corresponding bivariate characteristic function is

$$\phi(u, v) = \exp\left(-\frac{\mathbf{u}'\boldsymbol{\Sigma}\mathbf{u}}{2}\right) = \exp\left(-\frac{\sigma^2 u^2 + 2\rho uv + \sigma^2 v^2}{2}\right). \quad (89)$$

Hence, as an illustration,

$$E(X^2 Y^2) = \frac{\partial^4 E \exp(i(uX + vY))}{\partial u^2 \partial v^2} \Big|_{u,v=0} = \frac{\partial^4 \left( \exp\left(-\frac{\sigma^2 u^2 + 2\rho uv + \sigma^2 v^2}{2}\right) \right)}{\partial u^2 \partial v^2} \Big|_{u,v=0} = \sigma^4 + 2\rho^2$$

(use Maple, or expand the exponential to second order and keep only relevant terms).

The following two identities are often needed.

**Problem 1:** Assume  $E(\mathbf{X}) = 0$ . Prove that

$$E(X_1 X_2 X_3) = 0. \quad (90)$$

**Problem 2:** Assume  $E(\mathbf{X}) = 0$ . Prove the so-called *fourth-cumulant identity*:

$$E(X_1 X_2 X_3 X_4) = E(X_1 X_2) E(X_3 X_4) + E(X_1 X_3) E(X_2 X_4) + E(X_1 X_4) E(X_2 X_3). \quad (91)$$

Additional properties of Gaussian variables are found in Øksendal, Appendix A.

We close this section with a *very* important property of multivariate Gaussian variables:

*Multivariate Gaussian variables  $\{X_1, \dots, X_n\}$  are independent if and only if their covariance matrix  $\Sigma = \{\text{Cov}(X_i, X_j)\}$  is diagonal.*

Thus, multivariate Gaussian variables are independent if all pairs of variables are uncorrelated (and hence independent). It should be stressed that the multivariate Gaussian assumption is required for this to be true. The proof is simply to observe that if  $\Sigma$  is diagonal, the joint density splits into a product of the individual densities, which is the elementary definition of independence (Strictly speaking, the statement is true even if some of the variables have 0 variance, since constants are independent of anything and may be removed in the proof).