Book 6: Densities, Transformed Distributions, and Limit Theorems

Robert R. Reitano
Foundations of Quantitative Finance:
6. Densities, Transformed Distributions, and Limit Theorems

Robert R. Reitano
Brandeis International Business School
Waltham, MA 02454

January, 2019
Contents

Preface ix

to Dorothy and Domenic xi

Introduction xiii

1 Density Functions and Borel Measures 1
  1.1 Joint Density Functions 1
  1.1.1 Joint Distribution Functions 1
  1.1.2 A Digression into Integration Theory 4
    Riemann-Stieltjes Perspective 5
    Lebesgue-Stieltjes Perspective 6
    Riemann-Stieltjes vs. Lebesgue-Stieltjes 7
  1.1.3 Properties of Density Functions 9
  1.1.4 Existence of Density Functions 12
  1.2 Marginal Density Functions 15
  1.3 Independent Random Variables 18
  1.4 Conditional Density Functions 20

2 DFs of Transformed Random Vectors 25
  2.1 Product Integrals and Cavalieri’s Principle 25
  2.2 DFs of Sums and Ratios of RVs 29
    2.2.1 Sums of Independent Random Vectors 29
    2.2.2 Ratios of Independent Random Variables 35
  2.3 Densities of Transformed Random Vectors 39

3 Multivariate Normal Distribution 43
  3.1 Derivation and Properties 43
  3.2 Interesting Properties 52
## CONTENTS

3.2.1 Higher Moments ........................................... 53  
3.2.2 Independent vs. Uncorrelated ............................ 57  
3.2.3 Mean and Variance of a Random Normal Sample ....... 59  

3.3 Existence of Multivariate Normal Densities ............... 65  

4 Weak Convergence of Measures ............................... 69  
4.1 Measures on $\mathbb{R}$ ....................................... 69  
4.1.1 Portmanteau Theorem on $\mathbb{R}$ ........................ 69  
4.1.2 Applications .............................................. 73  
4.2 Measures on $\mathbb{R}^j$ ...................................... 74  
4.2.1 Portmanteau Theorem on $\mathbb{R}^j$ ..................... 74  
4.2.2 Applications .............................................. 80  
4.2.3 Scheffé’s Theorem ......................................... 94  
4.2.4 Prokhorov’s theorem ...................................... 100  

5 Expectations of Random Variables ............................ 105  
5.1 General Definition of $E[g(X)]$ ............................ 105  
5.2 Weak Convergence and Moment Limits ..................... 110  
5.3 Conditional Expectations ................................... 113  
5.3.1 Conditional Probability Measures ..................... 113  
5.3.2 Conditional Expectations ................................ 118  
General Conditional Expectations - An Introduction .......... 121  
5.3.3 Properties of Conditional Expectations ................. 127  

6 The Characteristic Function .................................. 133  
6.1 The Moment Generating Function .......................... 133  
6.2 The Characteristic Function ................................ 137  
6.3 Examples of Characteristic Functions ..................... 139  
6.3.1 Discrete Distributions ................................. 140  
6.3.2 Continuous Distributions ................................ 141  
6.4 Properties of Characteristic Functions on $\mathbb{R}$ ...... 144  
6.5 Properties of Characteristic Functions on $\mathbb{R}^n$ ....... 147  
6.5.1 The Cramér-Wold Device ............................... 160  
6.6 Bochner’s Theorem ........................................... 162
## CONTENTS

### 6.7 A Uniqueness of Moments Result .......................... 172

### 7 Applications of Characteristic Functions .......................... 175
  7.1 Central Limit Theorems .................................. 176
      7.1.1 The Classical Central Limit Theorem ................. 177
      7.1.2 Lindeberg’s Central Limit Theorem ................. 179
      7.1.3 Lyapunov’s Central Limit Theorem .................. 186
      7.1.4 A Central Limit Theorem on \( \mathbb{R}^n \) .............. 188
  7.2 Distribution Families Related Under Addition ............... 190
      7.2.1 Discrete Distributions ............................. 191
      7.2.2 Continuous Distributions ......................... 193
  7.3 Infinitely Divisible Distributions ............................ 195
  7.4 Distribution Families Related Under Multiplication .......... 209

### 8 Discrete Time Asset Models in Finance .......................... 211
  8.1 Spatial Models of Asset Prices ............................. 212
  8.2 Temporal Models of Asset Prices ............................ 214
      8.2.1 Additive Temporal Models .......................... 215
      8.2.2 Multiplicative Temporal Models ...................... 217
  8.3 Induced Models ........................................... 219
      8.3.1 Spatial and Additive Temporal Models .............. 219
      8.3.2 Spatial and Multiplicative Temporal Models ........ 221
      8.3.3 Simulating Asset Price Paths ....................... 224
  8.4 Harmonious Asset Models .................................... 225
      8.4.1 Identifiability of the Spatial Distribution ........ 226
      8.4.2 Scalability of the Temporal Distribution ........... 228
      8.4.3 Harmonious Additive Temporal Models ............... 231
      8.4.4 Harmonious Multiplicative Temporal Models ........ 236
  8.5 Limiting Distributions of Harmonious Asset Models .......... 240
      8.5.1 Perfectly Scalable Harmonious Models ............... 241
      8.5.2 Approximately Scalable Harmonious Models .......... 242
      Additive Temporal Models ................................. 243
      Multiplicative Temporal Models ......................... 248

### 9 Pricing of Financial Derivatives ................................. 253
  9.1 Discrete Time Pricing on a Binomial Lattice .................. 253
      9.1.1 Binomial Lattice Pricing of European Derivatives ...... 254
      9.1.2 Binomial Lattice Pricing of American Derivatives .... 262
  9.2 Limiting Asset Distribution: Risk Neutral Measure .......... 264
      9.2.1 Analysis of Risk Neutral Binomial Probability \( q \) ........ 264
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.2.2</td>
<td>Limiting Asset Distribution Under $q$</td>
<td>266</td>
</tr>
<tr>
<td>9.3</td>
<td>Limiting Price of European Derivatives</td>
<td>268</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Black-Scholes-Merton Option Pricing Formulas</td>
<td>275</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Properties of Black-Scholes-Merton Formulas</td>
<td>279</td>
</tr>
<tr>
<td></td>
<td>Price Convergence to the Payoff</td>
<td>279</td>
</tr>
<tr>
<td></td>
<td>Put-Call Parity</td>
<td>281</td>
</tr>
<tr>
<td></td>
<td>A Derivative’s &quot;Greeks&quot;</td>
<td>283</td>
</tr>
<tr>
<td></td>
<td>A Derivative’s &quot;Greeks&quot; - Lattice Approximations</td>
<td>289</td>
</tr>
<tr>
<td>9.4</td>
<td>Limiting Price of American Derivatives</td>
<td>292</td>
</tr>
<tr>
<td>9.5</td>
<td>Binomial Pricing of Path Dependent European Options</td>
<td>292</td>
</tr>
<tr>
<td>9.5.1</td>
<td>Binomial Path-Based Pricing</td>
<td>293</td>
</tr>
<tr>
<td></td>
<td>Standard European Options</td>
<td>293</td>
</tr>
<tr>
<td></td>
<td>Path-Dependent European Options</td>
<td>294</td>
</tr>
<tr>
<td>9.5.2</td>
<td>Approximate Path-Based Pricing</td>
<td>298</td>
</tr>
<tr>
<td></td>
<td>Multinomial Distribution</td>
<td>299</td>
</tr>
<tr>
<td></td>
<td>Approximate Path-Based Pricing of European Options</td>
<td>301</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>305</td>
</tr>
</tbody>
</table>
Preface
The idea for a reference book on the mathematical foundations of quantitative finance has been with me throughout my career in this field. But the urge to begin writing it didn’t materialize until shortly after completing my first book, Introduction to Quantitative Finance: A Math Tool Kit, in 2010. The one goal I had for this reference book was that it would be complete and detailed in the development of the many materials one finds referenced in the various areas of quantitative finance. The one constraint I realized from the beginning was that I could not accomplish this goal, plus write a complete survey of the quantitative finance applications of these materials, in the 700 or so pages that I budgeted for myself for my first book. Little did I know at the time that this project would require a multiple of this initial page count budget even without detailed finance applications.

I was never concerned about the omission of the details on applications to quantitative finance because there are already a great many books in this area that develop these applications very well. The one shortcoming I perceived many such books to have is that they are written at a level of mathematical sophistication that requires a reader to have significant formal training in mathematics, as well as the time and energy to fill in omitted details. While such a task would provide a challenging and perhaps welcome exercise for more advanced graduate students in this field, it is likely to be less welcome to many other students and practitioners. It is also the case that quantitative finance has grown to utilize advanced mathematical theories from a number of fields. While there are also a great many very good references on these subjects, most are again written at a level that does not in my experience characterize the backgrounds of most students and practitioners of quantitative finance.

So over the past several years I have been drafting this reference book, accumulating the mathematical theories I have encountered in my work in this field, and then attempting to integrate them into a coherent collection of books that develops the necessary ideas in some detail. My target readers would be quantitatively literate to the extent of familiarity, indeed comfort, with the materials and formal developments in my first book, and sufficiently motivated to identify and then navigate the details of the materials they were attempting to master. Unfortunately, adding these details supports learning but also increases the lengths of the various developments. But this book was never intended to provide a “cover-to-cover” reading challenge, but rather to be a reference book in which one could find detailed foundational materials in a variety of areas that support current questions and further studies in quantitative finance.
Over these past years, one volume turned into two, which then became a work not likely publishable in the traditional channels given its unforgiving size and likely limited target audience. So I have instead decided to self-publish this work, converting the original chapters into stand-alone books, of which there are now nine. My goal is to finalize each book over the coming year or two.

I hope these books serve you well.

I am grateful for the support of my family: Lisa, Michael, David, and Jeffrey, as well as the support of friends and colleagues at Brandeis International Business School.

Robert R. Reitano
Brandeis International Business School
to Dorothy and Domenic
Introduction

This is the sixth book in a series of several that will be self-published under the collective title of Foundations of Quantitative Finance. Each book in the series is intended to build from the materials in earlier books, with the first several alternating between books with a more foundational mathematical perspective, which was the case with the first, third and fifth book, and books which develop probability theory and some quantitative applications to finance, the focus of the second and fourth and now this sixth book. Latter books will be on stochastic processes. While providing many of the foundational theories underlying quantitative finance, this series of books does not provide a detailed development of these financial applications. Instead this series is intended to be used as a reference work for students, researchers and practitioners of quantitative finance who already have other sources for these detailed financial applications but find that such sources are written at a level which assumes significant mathematical expertise, which if not possessed can be difficult to acquire.

Because the goal of many books in quantitative finance is to develop financial applications from an advanced point of view, it is often the case that the needed advanced foundational materials from mathematics and probability theory are introduced and summarized, but without a complete and formal development that would of necessity take the respective authors too far astray from their intended objectives. And while there are a great many excellent books on mathematics and probability theory, a number of which are cited in the references, such books typically develop materials with a eye to comprehensiveness in the subject matter, and not with an eye toward efficiently curating and developing the theory needed for applications in quantitative finance.

Thus the goal of this series is to introduce and develop in some detail a number of the foundational theories underlying quantitative finance, with topics curated from a vast mathematical and probability literature for the ex-
press purpose of supporting applications in quantitative finance. In addition, the development of these topics will be found to be at a much greater level of detail than in most advanced quantitative finance books, and certainly in more detail that most advanced mathematics and probability theory texts.

The title of this sixth book, *Densities, Transformed Distributions, and Limit Theorems*, reflects the general scope of materials developed which on the one hand reflect the general integration theory of book 5, but also continue the themes of transformed distributions and limit theorems initiated in books 2 and 4.

Chapter 1 begins with a review of the earlier integration theories of Riemann, Lebesgue and their Stieltjes extensions from books 3 and 5, with emphasis on how these are related in probability theory applications. We then turn to a study of joint and related density functions and their properties, as well as address the density existence question which is characterized in terms of the absolute continuity of the associated Borel measure relative to Lebesgue measure. That this result echoes the Radon-Nikodým theorem of proposition 7.22 of book 5 is no coincidence, as it is simply an application of this earlier result to a probability theory question.

The transformation of distribution functions is studied in chapter 2, formalizing the introductory materials of book 4 with the development of a key tool, Cavalieri’s principle. Distribution functions of sums of random vectors and products of random variables are developed, as well as results on transformations of the associated density functions. The multivariate normal distribution is the focus of chapter 3, beginning with the derivation of the density function of an affine transformation of a random vector of independent unit normal variates. We then investigate linear transformations of general multivariate normal vectors, and the generation of normal vectors using the Cholesky decomposition of the associated covariance matrix $C$. The chapter ends with special properties of the normal distribution, and applications to normal samples.

General results on weak convergence of measures are developed in chapter 4, beginning with the portmanteau theorems, first in one dimension, then generalized. Applications investigated are the general continuous mapping theorem and its implication for convergence of random variables for the Mann-Wald theorem, then part 1 of the Cramér-Wold theorem and general versions of Slutsky’s theorem and the delta method, and finally Scheffé’s theorem and a general version of Prokhorov’s theorem.

With the aid of book 5’s integration theory, the theory of expectations is finally put on a solid foundation in chapter 5, with then also addresses weak convergence and moment limits. In this chapter we also develop the
theory of conditional expectations and its properties. Chapter 6 then turns to characteristic functions as an application of the Fourier analysis of book 5. Examples and properties are developed before turning to applications to part 2 of the Cramér-Wold theorem, Bochner’s theorem and a result on uniqueness of moments.

Chapter 7 then turns to the more extensive applications of characteristic functions, first turning to the classical central limit theorem and the versions of Lindeberg and Lyapunov, and a version in $\mathbb{R}^n$. We then study distribution families related under addition and a development of some results on infinitely divisible distributions.

The last two chapters focus on applications in finance. Chapter 8 investigates the notions of the temporal and spatial distributions of asset price models, introduces the notions of harmonious models, and then develops limiting distribution results for harmonious asset models. Chapter 9 addresses the binomial model pricing of financial derivatives and various limiting results, a derivatives' "Greeks," as well as the associated pricing approach for path dependent options and a study of convergence of path-based prices.
Chapter 1

Density Functions and Borel Measures

Joint and marginal distribution functions and their associated Borel measures were studied in books 2 and 4, and with the aid of the general integration theory of book 5 are only now in the position to formally introduce the associated density functions when they exist. In doing so, we finally justify the perhaps familiar mechanical manipulations seen in earlier books. Recall the following definition:

Definition 1.1 The Borel sigma algebra $\mathcal{B}(\mathbb{R}^n)$ is the smallest sigma algebra that contains the open sets. A Borel measure $\mu$ is a measure defined on $\mathcal{B}(\mathbb{R}^n)$ with the added property that $\mu(A) < \infty$ for all compact $A \in \mathcal{B}(\mathbb{R}^n)$.

These notions are named for Émile Borel (1871 – 1956).

1.1 Joint Density Functions

1.1.1 Joint Distribution Functions

Although joint distribution functions were introduced in definition 3.28 of book 2, we recall that definition here for completeness. Recall that a random variable $X$ is simply measurable function defined on a probability space $(\mathcal{S}, \mathcal{E}, \mu)$, so $X : \mathcal{S} \to \mathbb{R}$ and where measurability means that $X^{-1}(\mathcal{B}(\mathbb{R})) \subset \mathcal{E}$. A random vector is then a vector-valued function defined on this probability space with measurable component functions.
Definition 1.2 If $X_j : S \rightarrow \mathbb{R}$ are random variables defined on $(S, \mathcal{E}, \mu)$, $j = 1, 2, ..., n$, define the random vector $X = (X_1, X_2, ..., X_n)$ as the transformation or vector-valued function:

$$X : S \rightarrow \mathbb{R}^n,$$

with valued on $s \in S$ given by:

$$X(s) \equiv (X_1(s), X_2(s), ..., X_n(s)).$$

The joint distribution function (d.f.), or joint cumulative distribution function (c.d.f.) associated with $X$, denoted $F$ or $F_X$, is then defined on $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$ by:

$$F(x_1, x_2, ..., x_n) = \mu \left[ \bigcap_{j=1}^n X_j^{-1}(-\infty, x_j) \right]. \quad (1.1)$$

Notation 1.3 It is common in probability theory to say that $F(x_1, x_2, ..., x_n)$ is the probability that $X \in A(x_1, x_2, ..., x_n)$, denoted:

$$F(x_1, x_2, ..., x_n) = \Pr \left[ X \in A(x_1, x_2, ..., x_n) \right],$$

where $A(x_1, x_2, ..., x_n) \subset \mathbb{R}^n$ is defined as the unbounded right semi-closed rectangle:

$$A(x_1, x_2, ..., x_n) = \prod_{j=1}^n (-\infty, x_j). \quad (1.2)$$

This is a logical use of language since $X(s) \in A(x_1, x_2, ..., x_n)$ if and only if $s \in \bigcap_{j=1}^n X_j^{-1}(-\infty, x_j)$, and thus the probability of this $X$-event is by definition the $\mu$-measure or $\mu$-probability of this $s$-event as given in 1.1. Thus also:

$$F(x_1, x_2, ..., x_n) = \mu \left[ X^{-1} \left( A(x_1, x_2, ..., x_n) \right) \right]. \quad (1.3)$$

Remark 1.4 While this definition does not explicitly address measurability of the random vector $X$ relative to $\mathcal{B}(\mathbb{R}^n)$, proposition 3.32 of book 2 proved that if $X = (X_1, X_2, ..., X_n)$ with each $X_j : S \rightarrow \mathbb{R}$ a random variable on $(S, \mathcal{E}, \mu)$, then $X^{-1}(A) \in \mathcal{E}$ for all $A \in \mathcal{B}(\mathbb{R}^n)$. Conversely, if $X^{-1}(A) \in \mathcal{E}$ for all $A \in \mathcal{B}(\mathbb{R}^n)$ where $X = (X_1, X_2, ..., X_n)$, then each $X_j$ is a random variable on $(S, \mathcal{E}, \mu)$.

Joint distribution functions were seen to continuous from above and $n$-increasing in proposition 6.9 of book 2, recalling that this meant:
1.1 JOINT DENSITY FUNCTIONS

Definition 1.5 A function $F : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be **continuous from above at** $x$ if given $x^{(m)} \in \mathbb{R}^n$ with $x^{(m)}_i > x_i$ for all $i$ and $x^{(m)} \to x$ monotonically as $m \to \infty$:

$$F(x) = \lim_{m \to \infty} F(x^{(m)}). \quad (1.4)$$

Further, $F$ is said to be **$n$-increasing** if for any right semi-closed rectangle $\prod_{i=1}^n (a_i, b_i)$:

$$\sum_x sgn(x)F(x) \geq 0, \quad (1.5)$$

where each $x = (x_1, \ldots, x_n)$ in the summation is one of the $2^n$ vertices of this rectangle, so $x_i = a_i$ or $x_i = b_i$, and $sgn(x)$ is defined as $-1$ if the number of $a_i$-components of $x$ is odd, and $+1$ otherwise.

The significance of these two properties is that by proposition 8.14 of book 1, any such $F$ then **induces a measure** $\mu_F$ on the Borel sigma algebra $\mathcal{B}(\mathbb{R}^n)$ so that for all right semi-closed rectangles in $\mathbb{R}^n$:

$$\mu_F \left( \prod_{i=1}^n (a_i, b_i) \right) = \sum_x sgn(x)F(x). \quad (1.6)$$

The summation in 1.6 is defined as in 1.5 above. For this construction, $\mu_F$ is **not** uniquely defined by general such $F$ since replacing $F$ by $F + c$ for any $c \in \mathbb{R}$ obtains the same measure. However, if it is also the case that $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to \infty$ with apparent notation, and thus $F$ then has all the properties of a joint distribution function of a random vector $X$, then $\mu_F$ will be uniquely defined as a probability measure on $\mathbb{R}^n$ and:

$$F(x_1, x_2, \ldots, x_n) = \mu_F \left[ \prod_{i=1}^n (-\infty, x_i] \right]. \quad (1.7)$$

Remark 1.6 Comparing 1.3 and 1.7 to definition 3.9 of book 5, the measure $\mu_F$ defined on $\mathcal{B}(\mathbb{R}^n)$ is an example of a measure on the range space $\mathbb{R}^n$ **induced** by the transformation $X : S \rightarrow \mathbb{R}^n$, where $X \equiv (X_1, X_2, \ldots, X_n)$. In the notation of that definition, $\mu_F = \mu_X$, and thus for all $A \in \mathcal{B}(\mathbb{R}^n)$:

$$\mu_F(A) = \mu \left[ X^{-1} (A) \right]. \quad (1.8)$$

Conversely, if $\lambda$ is a probability measure on $\mathbb{R}^n$, there is an associated function $F_\lambda(x)$ **induced** by $\lambda$ and defined on $x \equiv (x_1, \ldots, x_n)$ by:

$$F_\lambda(x) = \lambda \left[ \prod_{i=1}^n (-\infty, x_i] \right].$$
CHAPTER 1 DENSITY FUNCTIONS AND BOREL MEASURES

By proposition 8.10 of book 1, $F_{\lambda}$ is continuous from above and $n$-increasing, and since $\lambda$ is a probability measure this assures that $F_{\lambda}(x) \rightarrow 0$ as $x \rightarrow -\infty$ and $F_{\lambda}(x) \rightarrow 1$ as $x \rightarrow \infty$. Thus by proposition 6.9 of book 2, $F_{\lambda}$ has all the properties of a joint distribution function of a random vector $X$. In addition, $F_{\lambda}$ is uniquely defined by $\lambda$ since if given $\lambda_1, \lambda_2$ such that $F_{\lambda_1}(x) = F_{\lambda_2}(x)$, then $\lambda_1 = \lambda_2$ on all rectangles $\prod_{i=1}^{n}(a_i, b_i]$ by proposition 8.9 of book 1, and thus by uniqueness of proposition 6.14 of book 1 it follows that $\lambda_1 = \lambda_2$.

In summary, probability measures on $\mathbb{R}^n$ and functions which have all the properties of joint distribution functions on $\mathbb{R}^n$ come in unique pairs.

One final question: Is every function on $\mathbb{R}^n$ which has all the properties of a joint distribution function on $\mathbb{R}^n$ actually a joint distribution function? The following result generalizes proposition 3.6 of book 2.

**Proposition 1.7** Let $F(x)$ be continuous from above and $n$-increasing on $\mathbb{R}^n$ and satisfy $F(x) \rightarrow 0$ as $x \rightarrow -\infty$ and $F(x) \rightarrow 1$ as $x \rightarrow \infty$. Here the notation $x \rightarrow \pm \infty$ means that $x_i \rightarrow \pm \infty$ for all $i$. Then there exists a probability space $(S, \mathcal{E}, \mu)$ and a random vector $X : (S, \mathcal{E}, \mu) \rightarrow (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), m^n)$, with $m^n$ Lebesgue measure, so that $F$ is the joint distribution function of $X$:

$$F_{X}(x) = F(x).$$

**Proof.** With the discussion above, this proof is surprisingly simple. Let $(S, \mathcal{E}, \mu) \equiv (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu_F)$ where $\mu_F$ is the unique probability measure on $\mathbb{R}^n$ induced by $F$, and let $X$ be the identity function, $X : x \rightarrow x$. Then $X^{-1}(\prod_{i=1}^{n}(-\infty, x_i]) = \prod_{i=1}^{n}(-\infty, x_i]$ and thus by 1.3 and 1.7:

$$F_{X}(x_1, x_2, \ldots, x_n) \equiv \mu_F\left[X^{-1}\left(\prod_{i=1}^{n}(-\infty, x_i]\right)\right] = F(x_1, x_2, \ldots, x_n).$$

$\blacksquare$

1.1.2 A Digression into Integration Theory

Because every distribution function has the fundamental properties of definition 1.5, and these are all that is required for the current discussion on integration theory, we will focus this section on joint distribution functions for notational convenience. However, nothing really changes in the application of the following discussion to other distribution functions, such as univariate distribution functions or the marginal or conditional distribution functions discussed below.
1.1 JOINT DENSITY FUNCTIONS

Riemann-Stieltjes Perspective

Because $F(x)$ is $n$-increasing and bounded, where $x \equiv (x_1, x_2, ..., x_n)$ for simplicity, it follows from proposition 4.69 of book 3 that for continuous and bounded $g(x)$, the Riemann-Stieltjes integral:

$$
\int_A g dF,
$$

is defined over every bounded or unbounded, right semi-closed rectangle $A \equiv \prod_{i=1}^n (a_i, b_i)$. Continuity of $g$ is defined to mean continuity on $A \equiv \prod_{i=1}^n [a_i, b_i]$, and so boundedness of $g$ is only an added restriction when $A$ is unbounded.

When $g \equiv 1$ that book’s definitions 4.53 and 4.64 obtain that $\int_A g dF$ is well defined to equal the summation in 1.6 for bounded such $A$, and proposition 4.69 then extends this to unbounded $A$ to produce:

$$
\mu_F(A) = \int_A dF.
$$

Letting $A \equiv A_{(x_1, x_2, ..., x_n)}$ :

$$
F(x_1, x_2, ..., x_n) = \int_{A_x} dF.
$$

Thus at least for rectangles, probability statements $X \in A$ are seen to equal Riemann-Stieltjes integrals of $g \equiv 1$ over $A$.

Also, proposition 4.75 of book 3 provides familiar results on the evaluation of integrals in 1.9 when $g$ is continuous and bounded and $A$ is a rectangle as above.

1. If $F(x)$ is $n$-times differentiable and $f(x) \equiv \frac{\partial^n F}{\partial x_1 \cdots \partial x_n}$ is continuous on $A$, then:

$$
\int_A g dF = (\mathcal{R}) \int_A g(x)f(x)dx,
$$

where the integral on the right is a Riemann integral.

2. Assume that $F(x) = \sum_{y_j \leq x} c_j$ where $y_j \leq x$ is shorthand for $y_{jk} \leq x_k$ for $1 \leq k \leq n$. Here $\{y_j\}_{j=1}^m \subset \mathbb{R}^n$ and $\{c_j\}_{j=1}^m$ are nonnegative with $\sum_{j=1}^m c_j = 1$, where if $m = \infty$ we assume that $\{y_j\}_{j=1}^m$ has no accumulation points. It is common to denote such $c_j$ by $f(y_j)$. Then:

$$
\int_A g dF = \sum_{y_j \in A} g(y_j)f(y_j).
$$
In both cases the function $f$ is called a **density function associated with** $F$. In the first case $f$ is typically referred to as a **continuous density function**, in the second a **discrete density function**, and both found throughout applications in probability theory.

**Lebesgue-Stieltjes Perspective**

Because $F(x)$ is continuous from above and $n$-increasing it induces a unique probability measure $\mu_F$ on the Borel sigma algebra $\mathcal{B}(\mathbb{R}^n)$ by proposition 8.14 of book 1. Thus by chapter 2 of book 5:

$$\int_A g d\mu_F,$$

(1.13)

is definable, though not necessarily finite, for all Borel measurable $g$ and $A \in \mathcal{B}(\mathbb{R}^n)$. That said, if both $g^+$ and $g^-$ are bounded, recalling definition 2.36 of book 5, the integral in 1.13 is well defined and finite since $\mu_F$ is a finite measure.

Letting $g \equiv 1$, it follows by definition that for all $A \in \mathcal{B}(\mathbb{R}^n)$:

$$\mu_F(A) = \int_A d\mu_F,$$

(1.14)

and for $A \equiv A_{(x_1,x_2,...,x_n)}$:

$$F(x_1, x_2, ..., x_n) = \int_{A_x} d\mu_F.$$

Thus for all $A \in \mathcal{B}(\mathbb{R}^n)$, probability statements $X \in A$ are seen to equal Lebesgue-Stieltjes integrals of $g \equiv 1$ over such $A$.

Proposition 3.6 of book 5 provides a familiar result on the evaluation of integrals in 1.13 in the special case where there exists a Lebesgue measurable function $f$ so that $\mu_F(A)$ can be represented:

$$\mu_F(A) = (\mathcal{L}) \int_A f(x) dx.$$  

(1.15)

Here the integral on the right is a Lebesgue integral. Any such Borel measure has the property that $\mu_F(A) = 0$ when $m(A) = 0$, where $m$ denotes Lebesgue measure. In this case we say that $\mu_F$ is **absolutely continuous with respect to** $m$, and is denoted $\mu_F \ll m$ in definition 7.3 of book 5. The Radon-Nikodým theorem of proposition 7.22 of book 5 then states that
if $\mu_F \ll \mu$ then there exists a Lebesgue measurable $f$ for which 1.15 is satisfied. Such $f$ is then uniquely defined $\mu$-a.e.

For distribution functions $F$ for which the induced Borel measure satisfies $\mu_F \ll \mu$, proposition 3.6 of book 5 then states that for all Borel measurable $g$ and $A \in \mathcal{B}(\mathbb{R}^n)$:

$$\int_A g d\mu_F = (\mathcal{L}) \int_A g(x) f(x) dx$$  \hspace{1cm} (1.16)

in the following sense. The function $g$ will be $\mu_F$-integrable if and only if $gf$ is Lebesgue integrable, and in this case the integrals agree for all such $A$.

**Riemann-Stieltjes vs. Lebesgue-Stieltjes**

It is apparent that the Lebesgue-Stieltjes model provides a far more robust set of tools for working with probability statements as well as a variety of associated integrals of Borel measurable functions over Borel measurable sets. The Riemann-Stieltjes model works best with probability statements on rectangles and the associated integrals of continuous functions. While it is possible to extend these latter applications it is difficult to do so without introducing the formal measure-theoretic ideas of the Lebesgue-Stieltjes theory. Rather than do this, it makes more sense to verify that one can use either approach, and that when both approaches apply, we will get the same result.

Comparing 1.10 and 1.14, $\mu_F(A)$ can be expressed in two ways for all bounded or unbounded rectangles $A$, and thus:

$$\int_A dF = \int_A d\mu_F.$$  \hspace{1cm} (1.17)

In particular, the distribution function $F(x_1, x_2, ..., x_n)$ can be expressed either way with $A \equiv A_{(x_1, x_2, ..., x_n)}$.

*The Lebesgue-Stieltjes integral on the right in 1.17 is also defined for all $A \in \mathcal{B}(\mathbb{R}^n)$ and equals $\mu_F(A)$ for all such $A$. The Riemann-Stieltjes integral on the left is not in general so defined based on the tools developed in book 3.*

For more general integrals, again we are restricted to a comparative result for rectangles $A$. Proposition 2.59 of book 5 states that for continuous
8 CHAPTER 1 DENSITY FUNCTIONS AND BOREL MEASURES

\( g \) on bounded \( A = \prod_{i=1}^{n}(a_i, b_i] \), where again continuity of \( g \) is defined to mean continuity on \( A = \prod_{i=1}^{n}[a_i, b_i] \), then:

\[
\int_A gdF = \int_A gd\mu_F. \tag{1.18}
\]

If \( g \) is \( \mu_F \)-integrable, then this equality extends to unbounded rectangles by Lebesgue’s dominated convergence theorem of proposition 2.43 of book 5. Define \( g_m = \chi_{A \cap A_m}g \) with \( A_m = \prod_{i=1}^{n}(-m, m] \) and \( \chi_{A \cap A_m} \) equal to 1 on \( A \cap A_m \) and 0 elsewhere. Then

\[
\int_A g_m d\mu_F \to \int_A g d\mu_F,
\]

and thus \( \int_A g dF \) is also well-defined by this limit.

The Lebesgue-Stieltjes integral on the right in 1.18 is also defined for all Borel measurable \( g \) and \( A \in B(\mathbb{R}^n) \). The Riemann-Stieltjes integral on the left is not in general so defined based on the tools developed in book 3.

In the special case in 1.18 where \( F \) is \( n \)-times differentiable and \( f(x) = \frac{\partial^n F}{\partial x_1 \partial x_2 \cdots \partial x_n} \) is continuous on bounded \( A = \prod_{i=1}^{n}[a_i, b_i] \), then for continuous \( g \):

\[
\int_A gdF = (R) \int_A g(x)f(x)dx = (L) \int_A g(x)f(x)dx. \tag{1.19}
\]

The first equality is 1.11 and the second follows from proposition 2.31 of book 3. This identity extends to unbounded rectangles if \( g(x)f(x) \) is absolutely Riemann integrable by that book’s proposition 2.64. But is this Lebesgue integral then equal to the Lebesgue-Stieltjes integral of 1.16? Put another way, is \( f \) as defined in terms of the derivative of \( F \) then equal to the measurable function that defines \( F \) in 1.15? The answer is "yes," and this follows by letting \( g = 1 \) in 1.19. By 1.10 we obtain that for all bounded and unbounded rectangles:

\[
\mu_F(A) = (L) \int_A f(x)dx.
\]

From this it follows by the uniqueness theorem of proposition 6.14 of book 1 that this identity holds for all \( A \in B(\mathbb{R}^n) \), and thus this \( f \) satisfies 1.15. In other words, by 1.16:

\[
(L) \int_A g(x)f(x)dx = \int_A g d\mu_F.
\]
1.1 JOINT DENSITY FUNCTIONS

The Lebesgue-Stieltjes integral on the right in 1.19 is also defined for all Borel measurable \( g \) and \( A \in B(\mathbb{R}^n) \). Neither the Riemann nor the Riemann-Stieltjes integral on the left is in general so defined based on the tools developed in book 3.

Finally, in the special case in 1.18 where \( F \) is a so-called saltus function (section 1.1, book4), \( F(x) = \sum_{y_j \leq x} c_j \) as above, then by 1.12 for continuous \( g \) and any rectangle \( A \):

\[
\int_A g dF = \sum_{y_j \in A} g(y_j) f(y_j),
\]

where we denote \( c_j \) by \( f(y_j) \). To prove that this summation also then equals \( \int_A g d\mu_F \), again let \( g = 1 \). Then by 1.10:

\[
\mu_F(A) = \sum_{y_j \in A} f(y_j)
\]

for all rectangles \( A \). Since \( \{y_j\}_{j=1}^m \) have no accumulation points by assumption, for fixed \( y_k = (y_{k1}, ..., y_{kn}) \) let \( A_k \equiv \prod_{i=1}^n (y_{ki} - \varepsilon, y_{ki} + \varepsilon) \) with \( \varepsilon \) small enough so that \( A_k \) contains no other \( y_j \) points. Then \( \mu_F(A_k) = f(y_k) \) and letting \( \varepsilon \to 0 \) and applying continuity from above of measures derives that \( \mu_F(y_k) = f(y_k) \). Similarly, \( \mu_F(y) = 0 \) for any \( y \notin \{y_j\}_{j=1}^m \). Thus \( \mu_F \) assigns all measure to \( \{y_j\}_{j=1}^m \) and by the definition of the Lebesgue-Stieltjes integral we obtain:

\[
\int_A g dF = \sum_{y_j \in A} g(y_j) f(y_j) = \int_A g d\mu_F. \tag{1.20}
\]

The Lebesgue-Stieltjes integral on the right in 1.20 is also defined for all Borel measurable \( g \) and \( A \in B(\mathbb{R}^n) \). The Riemann-Stieltjes integral on the left is in general not so defined based on the tools developed in book 3. Of course the summation in the middle is clearly well-defined for all such \( g \) and \( A \). The point is, for general \( g \) and \( A \) we cannot declare this to be the value of a Riemann-Stieltjes integral.

1.1.3 Properties of Density Functions

We begin with the definition of a density function.

**Definition 1.8** Given a joint distribution function \( F \) defined on \( \mathbb{R}^n \), a non-negative Lebesgue measurable function \( f \) is called the **density function associated with** \( F \) if for all \((x_1, x_2, ..., x_n)\), \( F \) defined by the Lebesgue integral:

\[
F(x_1, x_2, ..., x_n) = \int_{A(x_1, x_2, ..., x_n)} f dm^n. \tag{1.21}
\]
Here $A(x_1, x_2, ..., x_n)$ is given in 1.2 and $m^n$ denotes Lebesgue measure in the product space $(\mathbb{R}^n, \mathcal{M}_L^n, m^n)$. The function $f$ is sometimes called the joint density function associated with $F$.

**Remark 1.9** Existence of density functions is addressed in proposition 1.12 below. But note that when it exists, the density function $f$ is $m^n$-integrable since by Lebesgue’s monotone convergence theorem of proposition 2.21 of book 5:

$$\int_{\mathbb{R}^n} f \, dm^n = \lim_{x \to \infty} \int_{A(x_1, x_2, ..., x_m)} f \, dm^n = \lim_{x \to \infty} F(x_1, x_2, ..., x_n) = 1.$$ 

Thus by Fubini’s theorem of proposition 5.15 of book 5 this product integral can be written in the more appealing way as an iterated integral:

$$F(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(y_1, y_2, ..., y_n) \, dy_1 ... dy_n,$$

(1.22)

or as an iterated integral in any of the other $n! - 1$ orderings.

Also note that in general, a density function is not uniquely determined unless additional restrictions are imposed. For example, if $g = f$, $m^n$-a.e., then $g$ is also a density function associated with $F$. On the other hand, if one such $g$ is a continuous function, then this density function can be uniquely characterized with this additional constraint.

The next results on density functions will not surprise, and they reinforce the idea that probability statements can be expressed in terms of the integrals of density functions over the defining sets of such statements. We begin with:

**Proposition 1.10** If a distribution function $F$ defined on $\mathbb{R}^n$ has an associated density function $f$ and $\mu_F$ denotes the Borel measure induced by $F$, then for every bounded right semi-closed rectangle $A \equiv \prod_{j=1}^{n}(a_j, b_j)$:

$$\mu_F(A) = \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f(y_1, y_2, ..., y_n) \, dy_1 ... dy_n.$$ 

(1.23)
1.1 JOINT DENSITY FUNCTIONS

Proof. Since $f$ is $m^n$-integrable, the integral in 1.23 can be expressed by:

$$
\int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f(y_1, y_2, \ldots, y_n) dy_1 \cdots dy_n
$$

$$
= \int_{a_n}^{b_n} \cdots \int_{a_2}^{b_2} \left( \int_{-\infty}^{b_1} f(y_1, y_2, \ldots, y_n) dy_1 - \int_{-\infty}^{a_1} f(y_1, y_2, \ldots, y_n) dy_1 \right) dy_2 \cdots dy_n
$$

$$
= \int_{a_n}^{b_n} \cdots \int_{a_2}^{b_2} f(y_1, y_2, \ldots, y_n) dy_1 dy_2 \cdots dy_n
$$

$$
- \int_{a_n}^{b_n} \cdots \int_{a_2}^{a_1} f(y_1, y_2, \ldots, y_n) dy_1 dy_2 \cdots dy_n.
$$

Each of these integrals can be similarly expressed, for example:

$$
\int_{a_n}^{b_n} \cdots \int_{a_2}^{b_2} \int_{-\infty}^{b_1} f(y_1, y_2, \ldots, y_n) dy_1 dy_2 \cdots dy_n
$$

$$
= \int_{a_n}^{b_n} \cdots \int_{a_3}^{b_3} \int_{-\infty}^{b_2} f(y_1, y_2, \ldots, y_n) dy_1 dy_2 \cdots dy_n
$$

$$
- \int_{a_n}^{b_n} \cdots \int_{a_3}^{a_2} \int_{-\infty}^{b_1} f(y_1, y_2, \ldots, y_n) dy_1 dy_2 \cdots dy_n.
$$

By induction it then follows that:

$$
\int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f(y_1, y_2, \ldots, y_n) dy_1 \cdots dy_n
$$

$$
= \sum_x \text{sgn}(x) \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(y_1, y_2, \ldots, y_n) dy_1 \cdots dy_n
$$

$$
= \sum_x \text{sgn}(x) F(x_1, x_2, \ldots, x_n),
$$

where each $x = (x_1, \ldots, x_n)$ is one of the $2^n$ vertices of $\prod_{i=1}^n (a_i, b_i]$, so $x_i = a_i$ or $x_i = b_i$, and $\text{sgn}(x)$ is defined as $-1$ if the number of components of $x$ that equal $a_i$ is odd, and $+1$ otherwise.

Thus this integral equals $\mu_F[A]$ by 1.6.

The above result then generalizes:

**Proposition 1.11** If a distribution function $F$ defined on $\mathbb{R}^n$ has an associated density function $f$, and $\mu_F$ denotes the Borel measure induced by $F$, then for all $A \in \mathcal{B}(\mathbb{R}^n)$:

$$
\mu_F[A] = \int_A f dm^n.
$$

(1.24)
12CHAPTER 1 DENSITY FUNCTIONS AND BOREL MEASURES

Proof. This integral is well defined by:

$$\int_A f \, dm^n = \int_{\mathbb{R}^n} f \chi_A \, dm^n,$$

were \( \chi_A \) is the characteristic function of \( A \) and defined to equal 1 when \( x \in A \) and 0 otherwise, because \( f \chi_A \) is Lebesgue measurable and integrable since \( f \chi_A \leq f \). For the conclusion in 1.24, proposition 3.3 of book 5 assures that the set function:

$$\mu'[A] \equiv \int_A f \, dm^n$$

defines a measure on \( B(\mathbb{R}^n) \). By proposition 1.10, \( \mu' = \mu_F \) on the semi-algebra \( \mathcal{A}' \) of right semi-closed rectangles since if \( A \equiv \prod_{i=1}^{n} (a_i, b_i] \in \mathcal{A}' \), it follows from Fubini’s theorem that:

$$\int_A f \, dm^n = \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} f(y_1, y_2, ..., y_n) \, dy_1 \cdots dy_n.$$

This identity then extends to the algebra \( \mathcal{A} \) of finite disjoint unions of \( \mathcal{A}' \)-sets, and thus by the uniqueness theorem of proposition 6.14 of book 1, \( \mu' = \mu_F \) on \( \sigma(\mathcal{A}) \), the smallest sigma algebra generated by \( \mathcal{A} \). But as \( \sigma(\mathcal{A}) \) contains the open rectangles and thus the open sets, \( B(\mathbb{R}^n) \subset \sigma(\mathcal{A}) \). In fact \( B(\mathbb{R}^n) = \sigma(\mathcal{A}) \) though this is not required for this result. ■

1.1.4 Existence of Density Functions

The next result addresses the existence question for density functions. When \( n = 1 \) a complete characterization for general distribution functions was possible in proposition 1.3 of book 4. There it was seen that a distribution function \( F \) of one variable has a density function in the above sense if and only if \( F \) is absolutely continuous (definition 3.54, book 3), and in this case, \( f(x) = F'(x) \) m-a.e. As noted in remark 1.9, density functions are not uniquely defined but in the special case where \( F \) is continuously differentiable, \( F'(x) \) is continuous for all \( x \) and in this case \( f(x) = F'(x) \) is unique as the only continuous density function for \( F \).

For general \( n \) such a complete characterization of distribution functions appears elusive, though an existence result for the associated density functions is still possible. As seen in proposition 7.18 of book 5, a distribution function \( F \) of a single variable is absolutely continuous if and only if the induced Borel measure \( \mu_F \) is absolutely continuous with respect to Lebesgue
1.1 JOINT DENSITY FUNCTIONS

measure $m$, denoted $\mu_F \ll m$. Recall that this meant that $\mu_F(A) = 0$ if $m(A) = 0$. Thus the book 4 result above could be restated as: a distribution function $F$ of one variable has a density function in the above sense if and only if $\mu_F \ll m$.

**Proposition 1.12 (Existence of Density Functions)** Given a distribution function $F$ defined on $\mathbb{R}^n$ and $\mu_F$ the induced Borel measure, then $F$ has an associated density function $f$ in the sense of definition 1.8 if and only if $\mu_F \ll m^n$.

**Proof.** If $F$ has a density function then by definition of the integral (definition 2.40 of book 3), $\int_A f \, dm^n = 0$ when $m^n(A) = 0$. First, $\int_A h \, dm^n = 0$ for any simple function with $h \leq f$. Constructing increasing $h_m$ with $h_m \to f$ pointwise (example 2.53 of book 3), Lebesgue’s monotone convergence theorem then assures that $\int_A f \, dm^n = 0$ and thus $\mu_F \ll m^n$.

If $\mu_F \ll m^n$ then the existence of a nonnegative measurable function $f$ such that 1.24 is satisfied is the Radon-Nikodým theorem of proposition 7.22 of book 5. This result is applicable because $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), m^n)$ is a $\sigma$-finite measure space, and $\mu_F$ is a $\sigma$-finite (indeed, finite) measure. □

The next proposition provides a result on the differentiability of joint distribution functions and the relationship between such derivatives and the associated joint density function. This again extends the one variable result that $f(x) = F'(x)$ $m$-a.e. as noted above. Here we assume that $f$ is continuous, which in the one variable case assures that $F$ is continuously differentiable and thus absolutely continuous, and thus $f(x) = F'(x)$. For general $n$ an additional assumption is needed. The property identified in 1.25 is discussed in remark 1.14 below, but note that this assumption is vacuous when $n = 1$.

In the following, recall that for sets $A \subset B$ that $\bar{A} = B - A$.

**Proposition 1.13** Let $F(x_1, x_2, \ldots, x_n)$ be a distribution function with associated continuous density function $f(x_1, x_2, \ldots, x_n)$ and with the property that given any $I = (i_1, i_2, \ldots, i_m) \subset \{1, 2, \ldots, n\}$, there is an integrable function $g_I(x_{i_1}, x_{i_2}, \ldots, x_{i_m})$ so that:

$$f(x_1, x_2, \ldots, x_n) \leq g_I(x_{i_1}, x_{i_2}, \ldots, x_{i_m}), \text{ for all } x_j \in \bar{I}. \tag{1.25}$$

Then for all $(x_1, x_2, \ldots, x_n)$,

$$\frac{\partial^n F}{\partial x_1 \cdots \partial x_n} = f(x_1, x_2, \ldots, x_n). \tag{1.26}$$
Proof. We first prove that
\[ \frac{\partial F}{\partial x_1}(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_2} f(x_1, y_2, ..., y_n) dy_2 ... dy_n, \tag{(*)} \]
and then show that the general result follows by induction. The difference quotient for this partial derivative is:
\[ \frac{F(x_1 + \Delta x_1, \bar{x}) - F(x_1, \bar{x})}{\Delta x_1} = \int_{-\infty}^{\bar{x}} \left[ \frac{1}{\Delta x_1} \int_{x_1}^{x_1 + \Delta x_1} f(y_1, \bar{y}) dy_1 \right] d\bar{y}, \]
where \( \bar{x} \equiv (x_2, ..., x_n) \) and \( \bar{y} \equiv (y_2, ..., y_n) \) to simplify notation. Then:
\[
\left| \frac{F(x_1 + \Delta x_1, \bar{x}) - F(x_1, \bar{x})}{\Delta x_1} - \int_{-\infty}^{\bar{x}} f(x_1, \bar{y}) d\bar{y} \right| \\
\leq \int_{-\infty}^{\bar{x}} \left| \frac{1}{\Delta x_1} \int_{x_1}^{x_1 + \Delta x_1} f(y_1, \bar{y}) dy_1 - f(x_1, \bar{y}) \right| d\bar{y}.
\]
By the continuity of \( f \) and an application of the mean value theorem for integrals, the integrand of the \( d\bar{y} \)-integral on the right converges to 0 pointwise for every \( \bar{y} \) as \( \Delta x_1 \to 0 \). Also, this integrand is dominated by an integrable function of \( \bar{y} \), because \( f(y_1, \bar{y}) \leq g_1(\bar{y}) \) for all \( y_1 \), implies that:
\[
\left| \frac{1}{\Delta x_1} \int_{x_1}^{x_1 + \Delta x_1} f(y_1, \bar{y}) dy_1 - f(x_1, \bar{y}) \right| \leq 2g_1(\bar{y}),
\]
where \( g_1 = g_I \) for \( I = \{1\} \). An application of Lebesgue’s dominated convergence theorem obtains:
\[
\left| \frac{F(x_1 + \Delta x_1, \bar{x}) - F(x_1, \bar{x})}{\Delta x_1} - \int_{-\infty}^{\bar{x}} f(x_1, \bar{y}) d\bar{y} \right| \to 0
\]
proving (*).

Denoting this partial derivative by \( F_1 \), the above steps can be repeated to produce with \( \bar{x} = (x_3, ..., x_n) \):
\[
\left| \frac{F_1(x_1, x_2 + \Delta x_2, \bar{x}) - F_1(x_1, x_2, \bar{x})}{\Delta x_2} - \int_{-\infty}^{\bar{x}} f(x_1, x_2, \bar{y}) d\bar{y} \right| \\
\leq \int_{-\infty}^{\bar{x}} \left| \frac{1}{\Delta x_2} \int_{x_2}^{x_2 + \Delta x_2} f(x_1, y_2, \bar{y}) dy_2 - f(x_1, x_2, \bar{y}) \right| d\bar{y}.
\]
1.2 MARGINAL DENSITY FUNCTIONS

Pointwise convergence of the integrand to zero again follows from continuity, and the integrand is again dominated by an integrable function of $\bar{y} = (y_3, \ldots, y_n)$:

$$\left| \frac{1}{\Delta x_2} \int_{x_2}^{x_2+\Delta x_2} f(x_1, y_2, \bar{y}) dy_2 - f(x_1, x_2, \bar{y}) \right| \leq 2g_{12}(\bar{y}),$$

where $g_{12} = g_I$ for $I = \{1, 2\}$. Hence the second partial satisfies:

$$\frac{\partial^2 F}{\partial x_1 \partial x_2}(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_3} f(x_1, x_2, y_2, \ldots, y_n) dy_3 \cdots dy_n,$$

and the argument follows by induction. 

Remark 1.14 The condition in this theorem concerning the density function $f$ can be understood as follows. Let $(x_1, x_2, \ldots, x_n) \equiv (x, y)$ where $x = (x_{i_1}, x_{i_2}, \ldots, x_{i_m})$ and $y = (x_{j_1}, x_{j_2}, \ldots, x_{j_{n-m}})$ where $j_k \in \bar{I}$. By Fubini’s theorem of book 5, $f_x(y) \equiv f(x, y)$ is an integrable function of $y$ for almost all $x$, and $\int f(x, y) dx$ is an integrable function of $y$. Neither of these conclusions is adequate to support the above application of Lebesgue’s dominated convergence theorem, for which it was needed that $f(x, y)$ is bounded by an integrable function of $y$, uniformly in $x$.

So in the context of Fubini’s theorem, we required a strengthening of the conclusion that $f_x(y)$ is an integrable function of $y$ for almost all $x$, to an assumption that $f_x(y)$ is bounded by an integrable function of $y$ uniformly in $x$.

1.2 Marginal Density Functions

Marginal distribution functions were introduced in definition 3.34 of book 2:

Definition 1.15 Given a joint distribution function $F(x_1, x_2, \ldots, x_n)$ and $I = \{i_1, \ldots, i_m\} \subset \{1, 2, \ldots, n\}$, the marginal distribution function $F_I(x_{i_1}, x_{i_2}, \ldots, x_{i_m})$ is defined by:

$$F_I(x_{i_1}, x_{i_2}, \ldots, x_{i_m}) = \lim_{X' \to \infty} F(x_1, x_2, \ldots, x_n), \quad (1.27)$$

where $X' = (x_{j_1}, x_{j_2}, \ldots, x_{j_{n-m}})$ with $j_k \in \bar{I}$. 
CHAPTER 1 DENSITY FUNCTIONS AND BOREL MEASURES

There are \( 2^n - 2 \) proper marginal distribution functions defined by the proper subsets of \( \{1, 2, \ldots, n\} \), where "proper" means other that \( \emptyset \) and \( \{1, 2, \ldots, n\} \). In proposition 3.36 of book 2 it was shown that the marginal distribution function \( F_I \) defined in 1.27 is well defined in that it is independent of how \( X' \to \infty \). As a distribution function \( F_I \) is continuous from above and satisfies 1.5 for any index set \( I \subset \{1, 2, \ldots, n\} \). Hence, each such \( F_I \) induces a Borel measure \( \mu_{F_I} \) on \( \mathbb{R}^m \).

**Remark 1.16** If \( F(x_1, x_2, \ldots, x_n) \) is the joint distribution function of random variables \( \{X_j\}_{j=1}^n \) defined on \( (S, \mathcal{E}, \mu) \), \( j = 1, 2, \ldots, n \), and \( I = \{i_1, \ldots, i_m\} \subset \{1, 2, \ldots, n\} \), it is natural to wonder about the relationship between \( F_I(x_{i_1}, x_{i_2}, \ldots, x_{i_m}) \) and \( F(x_{i_1}, x_{i_2}, \ldots, x_{i_m}) \), the joint distribution function of \( \{X_{i_k}\}_{k=1}^m \). In fact:

\[
F_I(x_{i_1}, x_{i_2}, \ldots, x_{i_m}) = F(x_{i_1}, x_{i_2}, \ldots, x_{i_m}). \tag{1.28}
\]

That these are indeed equal is the essence of proposition 3.36 of book 2 as discussed in that book's remark 3.37. In essence, marginal distribution functions retain no "memory" of the random variables \( X_{jk} \) for \( j_k \in \bar{I} \), with which \( \{X_{i_k}\}_{k=1}^m \) were originally associated through \( F(x_1, x_2, \ldots, x_n) \).

The following result states that if \( F \) has an associated density function \( f \), then every marginal distribution \( F_I \) also has an associated density function \( f_I \) definable by 1.29.

**Proposition 1.17** If a distribution function \( F \) defined on \( (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \) has an associated density function \( f \), then for \( I = \{i_1, \ldots, i_m\} \subset \{1, 2, \ldots, n\} \), the marginal distribution function \( F_I \) has an associated density function \( f_I \) defined by:

\[
f_I(x_{i_1}, x_{i_2}, \ldots, x_{i_m}) = \int \cdots \int_{\mathbb{R}^{n-m}} f(x_1, x_2, \ldots, x_n) dx_{j_1} \ldots dx_{j_{n-m}}. \tag{1.29}
\]

As above, \( (x_{j_1}, x_{j_2}, \ldots, x_{j_{n-m}}) \) are defined by \( j_k \in \bar{I} \).

**Proof.** To simplify notation, assume that \( I = \{1, 2, \ldots, m\} \) and thus \( \bar{I} = \{m+1, \ldots, n\} \). By Fubini's theorem of proposition 5.15 of book 5 the iterated integral defining \( F \) can be reordered as follows, where \( X' = (x_{m+1}, x_{m+2}, \ldots, x_n) \):

\[
F_I(x_1, x_2, \ldots, x_m) = \lim_{X' \to \infty} \int_{-\infty}^{x_m} \cdots \int_{-\infty}^{x_1} \left[ \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_{m+1}} f(y_1, y_2, \ldots, y_n) dy_{m+1} \cdots dy_n \right] dy_1 \cdots dy_m.
\]
To justify taking the limiting operation inside the $dy_1...dy_m$-integral, define $A_N \subset \mathbb{R}^{n-m}$ by

$$A_N \equiv \prod_{k=m+1}^{n} (-\infty, N]_k,$$

where this interval notation is meant to suggest that $A_N$ is defined relative to the $\tilde{I}$ indexes. Define $f_N$ by

$$f_N(y_1, y_2, ..., y_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \chi_{A_N}(y_{m+1}, ..., y_n) f(y_1, y_2, ..., y_n) dy_{m+1}...dy_n.$$

Then with $f_I$ defined in 1.29, we have that $f_N(y_1, y_2, ..., y_m) \to f_I(y_1, y_2, ..., y_m)$ pointwise as $N \to \infty$. Since all functions in this sequence are dominated by $f_I(y_1, y_2, ..., y_m)$ which is integrable by Fubini’s theorem, we can apply Lebesgue’s dominated convergence theorem to obtain:

$$F_I(x_1, x_2, ..., x_m) = \int_{-\infty}^{x_m} \cdots \int_{-\infty}^{x_1} f_I(y_1, ..., y_m) dy_1...dy_m.$$ 

As $f_I(y_1, ..., y_m)$ is integrable, this iterated integral can then be expressed as in 1.21 and 1.29 follows.

Of all the $2^n - 2$ proper marginal distribution functions, by far the most important are the $n$ marginal distribution functions associated with the index set $I = \{j\}$ for $j = 1, 2, ..., n$. This is reinforced by the copula theory of chapter 7 of book 2. These special marginal distribution functions can be denoted $F_{(j)}(x_j)$, and there is no ambiguity if we denote these by $F_j(x_j)$ by remark 1.16. Then:

$$F_j(x_j) = \int_{-\infty}^{x_j} f_j(y_j) dy_j,$$ 

where by 1.29:

$$f_j(y_j) = \int_{\mathbb{R}^{n-1}} f(y_1, y_2, ..., y_n) dy_1...\widehat{dy}_j...dy_n,$$ 

and where $\widehat{dy}_j$ denotes that this integration variable is omitted.

**Corollary 1.18** Let $F(x_1, x_2, ..., x_n)$ denote joint distribution function of random variables $\{X_j\}_{j=1}^{n}$ with density function $f(x_1, x_2, ..., x_n)$. Then each distribution function $F_j(x_j)$ also has a density function given by 1.31, and:

$$F'_j(x_j) = f_j(x_j) \text{ m-a.e.}$$ (1.32)

**Proof.** The derivation of 1.31 is proposition 1.17. Since $f_j(y_j)$ is an integrable function by Fubini’s theorem, the marginal distributions in 1.30 are absolutely continuous and hence by proposition 3.37 of book 3, $F'_j(x_j)$ exists m-a.e. and 1.33 is satisfied.
In the general case of marginal distributions, we have the following result on derivatives of the associated distribution functions.

**Corollary 1.19** Let \( F(x_1, x_2, ..., x_n) \) be a distribution function with associated continuous density function \( f(x_1, x_2, ..., x_n) \) that satisfies the uniform integrability condition in 1.25 of proposition 1.13. Then for any \( I = \{i_1, ..., i_m\} \subset \{1, 2, ..., n\} \), the density function \( f_I(x_{i_1}, x_{i_2}, ..., x_{i_m}) \) defined in 1.29 also satisfies this condition, and hence:

\[
\frac{\partial^m F_I}{\partial x_{i_1}...\partial x_{i_m}} = f_I(x_{i_1}, x_{i_2}, ..., x_{i_m}).
\] (1.33)

**Proof.** Since \( F_I \) and \( f_I \) are again related by 1.22 by the proof of proposition 1.17, it is only necessary to prove that for any \( I \), that \( f_I(x_{i_1}, x_{i_2}, ..., x_{i_m}) \) satisfies the uniform integrability condition of 1.25 of proposition 1.13.

To simplify notation for this uniform integrability proof, again assume that \( I = \{1, ..., m\} \) and let \( K = \{1, ..., k\} \subset I \) be the first \( k \) index variables of \( I \). We show that \( f_I(x_1, ..., x_m) \) is uniformly integrable in \((x_1, ..., x_k)\), independent of \((x_{k+1}, ..., x_m)\). Define \( J = \{k+1, ..., m, m+1, ..., n\} \). Then since \( f(x_1, x_2, ..., x_n) \) satisfies the uniform integrability condition 1.25, there is an integrable function \( g_J(x_1, x_k, x_{m+1}, ..., x_n) \) so that for all \((x_{k+1}, ..., x_m)\):

\[
f(x_1, x_2, ..., x_n) \leq g_J(x_1, x_k, x_{m+1}, ..., x_n).
\]

Integrating with respect to \((x_{m+1}, ..., x_n)\) and applying 1.29 we find that for all \((x_{k+1}, ..., x_m)\):

\[
f_I(x_1, ..., x_m) \leq \int \cdots \int_{\mathbb{R}^{n-m}} g_J(x_1, ..., x_k, x_{m+1}, ..., x_n) \, dx_{m+1}...dx_n
\]

\[
\equiv \tilde{g}_J(x_1, ..., x_k).
\]

Now by Fubini’s theorem the function \( \tilde{g}_J \) is integrable, and the proof of 1.33 follows from proposition 1.13. ■

### 1.3 Independent Random Variables

Recall from definition 3.47 of book 2 the notion of independent random variables:

**Definition 1.20** If \( X_j : S \rightarrow \mathbb{R} \) are random variables on \((S, \mathcal{E}, \mu)\), \( j = 1, 2, ..., n \), we say that \( \{X_j\}_{j=1}^n \) are independent random variables if given \( \{A_j\}_{j=1}^n \subset \mathcal{B}(\mathbb{R}) \):

\[
\mu\left(\bigcap_{j=1}^n X_j^{-1}(A_j)\right) = \prod_{j=1}^n \mu\left(X_j^{-1}(A_j)\right).
\] (1.34)
Proposition 1.21 Let \( F(x_1, x_2, ..., x_n) \) denote the joint distribution function of independent random variables \( \{X_j\}_{j=1}^n \) with associated distribution functions \( \{F_j(x_j)\}_{j=1}^n \). Then \( F \) has a density function if and only if all \( F_j \) have density functions, and then:

\[
f(x_1, x_2, ..., x_n) = \prod_{j=1}^n f_j(x_j), \quad m^n\text{-a.e.} \tag{1.36}
\]

**Proof.** If all \( F_j \) have density functions which we denote \( f_j \), then \( F_j(x_j) = \int_{-\infty}^{x_j} f_j(y_j)\,dy_j \). From 1.35:

\[
F(x_1, x_2, ..., x_n) = \prod_{j=1}^n \int_{-\infty}^{x_j} f_j(y_j)\,dy_j \\
= \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} \left[ \prod_{j=1}^n f_j(y_j) \right] dy_1...dy_n \\
= \int_{A(x_1, x_2, ..., x_m)} \prod_{j=1}^n f_j\,dm^n.
\]

The middle step is notational, while the last step follows from Tonelli’s theorem of book 5 since \( \prod_{j=1}^n f_j \) is nonnegative and measurable. Tonelli’s theorem also then assures the \( m^n \)-integrability of \( \prod_{j=1}^n f_j \) because the iterated integral is finite. Thus \( F \) has a density function, and 1.36 follows since density functions are uniquely defined \( m^n \)-a.e.

Conversely, if \( f \) is the density function for \( F \), then Fubini’s theorem applies since \( f \) is integrable:

\[
F(x_1, x_2, ..., x_n) \equiv \int_{A(x_1, x_2, ..., x_m)} f\,dm^n \\
= \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(y_1, y_2, ..., y_n)\,dy_1...dy_n.
\]
This last expression also equals \( \prod_{j=1}^{n} F_j(x_j) \), and thus letting \( x_1, \ldots, x_n \rightarrow \infty \) and applying Fubini again obtains:

\[
F_j(x_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{x_j} f(y_1, y_2, \ldots, y_n) dy_1 \cdots dy_n = \int_{-\infty}^{x_j} \left[ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_1, y_2, \ldots, y_n) dy_1 \cdots \widehat{dy}_j \cdots dy_n \right] dy_j.
\]

Now

\[
f_j(y_j) \equiv \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_1, y_2, \ldots, y_n) dy_1 \cdots \widehat{dy}_j \cdots dy_n \quad (1.37)
\]

is measurable and integrable by Tonelli’s theorem, and is therefore a density function for \( F_j \). That 1.36 is satisfied now follows from the first result by 1.35. ■

**Corollary 1.22** Let \( F(x_1, x_2, \ldots, x_n) \) denote the joint distribution function of independent random variables \( \{X_j\}_{j=1}^{n} \). If \( I = \{i_1, \ldots, i_m\} \subset \{1, 2, \ldots, n\} \), then \( F_I \) has a density function if and only if all \( F_{i_k} \) have density functions, and then:

\[
f_I(x_{i_1}, x_{i_2}, \ldots, x_{i_m}) = \prod_{k=1}^{m} f_{i_k}(x_{i_k}), \ m \text{-a.e.} \quad (1.38)
\]

**Proof.** By remark 1.16, \( F_I(x_{i_1}, x_{i_2}, \ldots, x_{i_m}) \) is the joint distribution function of \( \{X_{i_k}\}_{k=1}^{m} \), and since independence of \( \{X_j\}_{j=1}^{n} \) assures independence of \( \{X_{i_k}\}_{k=1}^{m} \) by definition, proposition 1.21 obtains the result. ■

### 1.4 Conditional Density Functions

Conditional distribution functions were introduced in definition 3.39 of book 2, and reflected the notion of a conditional probability measure from that book’s definition 1.31.

**Definition 1.23** Let \( X : S \rightarrow \mathbb{R}^n \) be the random vector \( X = (X_1, X_2, \ldots, X_n) \) defined on \( (S, \mathcal{E}, \mu) \), \( J = \{j_1, \ldots, j_m\} \subset \{1, 2, \ldots, n\} \) and \( X_J = (X_{j_1}, X_{j_2}, \ldots, X_{j_m}) \). Given a Borel set \( B \in \mathcal{B}(\mathbb{R}^m) \) with \( \mu \left[ X_J^{-1}(B) \right] \neq 0 \), the conditional distribution function of \( X \) given \( X_J \in B \), denoted \( F(x_1, x_2, \ldots, x_n|X_J \in B) \) is given by the conditional probability measure:

\[
F(x_1, x_2, \ldots, x_n|X_J \in B) \equiv \mu \left[ X^{-1} \left( A_{(x_1, x_2, \ldots, x_n)} \right) \right| X_J^{-1}(B) .
\]

Thus:

\[
F(x_1, x_2, \ldots, x_n|X_J \in B) = \mu \left[ X^{-1} \left( A_{(x_1, x_2, \ldots, x_n)} \right) \cap X_J^{-1}(B) \right] / \mu \left[ X_J^{-1}(B) \right] , \quad (1.39)
\]
1.4 CONDITIONAL DENSITY FUNCTIONS

where \( A(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} (-\infty, x_i] \) as in 1.2.

This distribution function is sometimes denoted \( F_{J\mid B}(x_1, x_2, \ldots, x_n) \).

If the joint distribution function \( F(x_1, x_2, \ldots, x_n) \) has a density function \( f(x_1, x_2, \ldots, x_n) \), then it follows from 1.8 and then 1.24 that for all \( A \in \mathcal{B}(\mathbb{R}^n) \):

\[
\mu \left[ X^{-1}(A) \right] = \int_{A} f dm^n.
\]

Similarly, for all \( B \in \mathcal{B}(\mathbb{R}^m) \):

\[
\mu \left[ X^{-1}_{J}(B) \right] = \int_{B} f_J dm^m
\]

where \( f_J \) is the marginal density function for \( J \equiv \{j_1, \ldots, j_m\} \) and defined in 1.29. Since

\[
X^{-1}(A(x_1, x_2, \ldots, x_n)) \cap X^{-1}_{J}(B) = X^{-1}(A(x_1, x_2, \ldots, x_n) \cap (B)),
\]

we obtain the following:

**Proposition 1.24** Assume that the joint distribution function \( F(x_1, x_2, \ldots, x_n) \) has a density function \( f(x_1, x_2, \ldots, x_n) \), and that for \( X_J \equiv (X_{j_1}, X_{j_2}, \ldots, X_{j_m}) \) and \( B \in \mathcal{B}(\mathbb{R}^m) \) that \( \mu \left[ X^{-1}_{J}(B) \right] \neq 0 \) as in definition 1.23. Then \( F_{J\mid B}(x_1, x_2, \ldots, x_n) \) has a density function \( f_{J\mid B}(x_1, x_2, \ldots, x_n) \equiv f(x|x \in B) \) given by:

\[
f(x|x \in B) = \frac{f(x) \chi_B(x)}{\int_{B} f_J dm^m}.
\]

(1.40)

Here \( \chi_B(x_1, x_2, \ldots, x_n) = 1 \) if \((x_{j_1}, x_{j_2}, \ldots, x_{j_m}) \in B\) and is 0 otherwise, and \( f_J \) is the marginal density function as in 1.29.

**Proof.** Note that \( f_{J\mid B} \) is well defined since the denominator is nonzero by assumption that \( \mu \left[ X^{-1}_{J}(B) \right] \neq 0 \) as noted above. The above results can be applied in 1.39 to obtain:

\[
F(x_1, x_2, \ldots, x_n|X_J \in B) = \int_{A(x_1, x_2, \ldots, x_n) \cap (B)} f dm^n \left/ \int_{B} f_J dm^m \right.
\]

\[
= \int_{A(x_1, x_2, \ldots, x_n)} f \chi_B dm^n \left/ \int_{B} f_J dm^m \right.
\]

\[
= \int_{A(x_1, x_2, \ldots, x_n)} f_{J\mid B} dm^n.
\]
Remark 1.25 In general one does not encounter a marginal distribution function where $J = \{1, 2, \ldots, n\}$ in definition 1.23. But formally, $F_J = F$ in this case since the limit $X' \to \infty$ is vacuous in 1.27, and so too $f_J = f$. However it is not uncommon for a conditional distribution function to have $J = \{1, 2, \ldots, n\}$. See example 1.28 for a simple case.

Corollary 1.26 If the joint distribution function $F(x_1, x_2, \ldots, x_n)$ has a density function $f(x_1, x_2, \ldots, x_n)$, then with $X_J \equiv (X_{j_1}, X_{j_2}, \ldots, X_{j_m})$ and $B \in \mathcal{B}(\mathbb{R}^m)$ with $\mu [X_J^{-1}(B)] \neq 0$, if $\mu_{F_{J|B}}$ denotes the Borel measure induced by $F_{J|B}$, then for all $A \in \mathcal{B}(\mathbb{R}^n)$:

$$
\mu_{F_{J|B}}[A] = \int_A f_{J|B} dm^n. \tag{1.41}
$$

Proof. This follows immediately from propositions 1.24 and 1.11.

Exercise 1.27 Assume that $F(x) = \sum_{y_j \leq x} c_j$ where $y_j \leq x$ is shorthand for $y_{j_k} \leq x_k$ for $1 \leq k \leq n$. Here $\{y_j\}_{j=1}^N \subset \mathbb{R}^n$ and $\{c_j\}_{j=1}^N$ are nonnegative with $\sum_{j=1}^N c_j = 1$, where if $N = \infty$ we assume that $\{y_j\}_{j=1}^N$ has no accumulation points. Denote $c_j$ by $f(y_j)$.

1. Given $J \equiv \{j_1, \ldots, j_m\} \subset \{1, 2, \ldots, n\}$ and $X_J \equiv (X_{j_1}, X_{j_2}, \ldots, X_{j_m})$, define the marginal density function $f_J(x_{j_1}, x_{j_2}, \ldots, x_{j_m})$.

2. What does $\mu [X_J^{-1}(B)] \neq 0$ mean in this context?

3. Derive the analogous formula for $f_{J|B}(x_1, x_2, \ldots, x_n) \equiv f(x|x \in B)$ assuming that $\mu [X_J^{-1}(B)] \neq 0$:

$$
f(x|x \in B) = \frac{f(x) \chi_B(x)}{\sum_B f_J(x_{j_1}, x_{j_2}, \ldots, x_{j_m})}, \tag{1.42}
$$

where the summation is over all $(x_{j_1}, x_{j_2}, \ldots, x_{j_m}) \in B$.

4. Derive the analogous formula for $\mu_{F_{J|B}}[A]$ with $A \in \mathcal{B}(\mathbb{R}^n)$ assuming that $\mu [X_J^{-1}(B)] \neq 0$:

$$
\mu_{F_{J|B}}[A] = \sum_A f_{J|B}(x). \tag{1.43}
$$

Example 1.28 1. Let $F$ be a distribution function for a random variable $X$, and let $B = (x_0, \infty)$. To have $\mu [X^{-1}(B)] \neq 0$ implies by 1.8
1.4 CONDITIONAL DENSITY FUNCTIONS

that \( \mu_F[B] \neq 0 \), and since \( \mu_F[\bar{B}] \equiv \mu_F[(-\infty, x_0)] = F(x_0) \), this is equivalent to \( F(x_0) < 1 \). Then by 1.39:

\[
F(x|X \in B) \equiv \mu\left[X^{-1}((-\infty, x)) \cap X^{-1}((x_0, \infty))\right]/\mu\left(X^{-1}((x_0, \infty))\right) = \mu_F[(x_0, x)]/[1-F(x_0)].
\]

Thus:

\[
F(x|X \in B) = \begin{cases} 
0, & x < x_0, \\
\frac{F(x)-F(x_0)}{1-F(x_0)}, & x \geq x_0.
\end{cases}
\]

If \( F \) has a density function \( f \), then by 1.40:

\[
f(x|x \in B) = \frac{f(x)\chi_B(x)}{1-F(x_0)},
\]

and thus:

\[
f(x|X \in B) = \begin{cases} 
0, & x < x_0, \\
\frac{f(x)}{1-F(x_0)}, & x \geq x_0.
\end{cases}
\]

This density function also follows directly from 1.40 since \( J = 1 \) and \( f_J = f \).

2. If \( f(x, y) \) is discrete, \( B = \{y\} \) for some \( y \) with \( f_2(y) \neq 0 \) where \( f_2 \equiv f_J \) with \( J = \{2\} \), then from 1.42:

\[
f((x, y)|(x, y) \in B) = \frac{f(x, y)}{f_2(y)},
\]

a familiar formula from elementary probability theory.

A similar formula is possible when \( f(x, y) \) is a continuously differentiable density function, even though \( \mu[Z^{-1}(B)] = 0 \) in this case where \( Z \equiv (X, Y) \). The derivation in example 3.42 of book 2 begins with \( B^+ \equiv [y, y + \Delta y] \), and then investigates the marginal distribution function when \( \Delta y \to 0 \).
Chapter 2

DFs of Transformed Random Vectors

In sections 1.3 – 1.4 of book 4 we studied the distribution functions of general transformations of a single random variable, as well as examples which implicitly involved a joint distribution function. The details of this latter analysis were deferred to the present book since they required the tools of book 5. In this section we justify the manipulations introduced in book 4 for sums and ratios of random variables, as well as derive a general result on the joint distribution function of transformed random vectors.

2.1 Product Integrals and Cavalieri’s Principle

We begin with a development of the general framework needed for the applications of this section. Let $X, Y$ be independent random vectors defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$:

$X : \mathcal{S} \to \mathbb{R}^j, \quad Y : \mathcal{S} \to \mathbb{R}^k.$

Recalling definition 1.20, this means that for $A \in \mathcal{B}(\mathbb{R}^j)$ and $B \in \mathcal{B}(\mathbb{R}^k)$,

$\lambda[\{X(s) \in A\} \cap \{Y(s) \in B\}] = \lambda[\{X(s) \in A\}] \lambda[\{Y(s) \in B\}].$

Let $F_X(x)$ and $F_Y(y)$ be the associated distribution functions, and $
mu = \mu_{F_X}$ and $\nu = \nu_{F_Y}$ the induced Borel measures on $\mathbb{R}^j$ and $\mathbb{R}^k$, respectively, defined as in chapter 6 of book 2. Then $(\mathbb{R}^j, \mathcal{M}_\mu(\mathbb{R}^j), \mu)$ and $(\mathbb{R}^k, \mathcal{M}_\nu(\mathbb{R}^k), \nu)$ are complete measure spaces which contain the associated Borel sigma algebras, so $\mathcal{B}(\mathbb{R}^j) \subset \mathcal{M}_\mu(\mathbb{R}^j)$ and $\mathcal{B}(\mathbb{R}^k) \subset \mathcal{M}_\nu(\mathbb{R}^k)$.

25
By the assumption of independence, the random vector \((X,Y): \mathcal{S} \rightarrow \mathbb{R}^j \times \mathbb{R}^k = \mathbb{R}^{j+k}\) has distribution function given by proposition 3.53 of book 2:

\[
F(x,y) = F_X(x)F_Y(y).
\]

The Borel measure defined on \(\mathbb{R}^{j+k}\) induced by \(F\) is in fact the product measure \(\mu \times \nu\), the same measure as would be obtained from the product space defined by \((\mathbb{R}^j, \mathcal{M}_\mu(\mathbb{R}^j), \mu)\) and \((\mathbb{R}^k, \mathcal{M}_\nu(\mathbb{R}^k), \nu)\) using chapter 7 of book 1.

To prove this it is enough to show that the set functions induced by these distribution functions and defined on measurable rectangles \(\prod_{i=1}^{j+k}(a_i, b_i)\) satisfy this relationship:

\[
\mu_F \left( \prod_{i=1}^{j+k}(a_i, b_i) \right) = \mu_{F_X} \left( \prod_{i=1}^{j}(a_i, b_i) \right) \mu_{F_Y} \left( \prod_{i=j+1}^{j+k}(a_i, b_i) \right). \tag{2.1}
\]

These set functions are defined as in 1.6, and details are left in the following exercise.

**Exercise 2.1** Given the right semi-closed rectangle \(A = \prod_{i=1}^{j+k}(a_i, b_i)\), define \(\mu_F\) as in 1.6:

\[
\mu_F \left( \prod_{i=1}^{j+k}(a_i, b_i) \right) = \sum_{(x,y)} \text{sgn}(x,y)F(x,y),
\]

where the summation over \((x,y)\) is over all vertexes of \(A\) and where \(x\) denotes the first \(j\) components, and \(y\) the last \(k\) components. As always, \(\text{sgn}(x,y)\) equals \(-1\) if the total number of \(a_i\)-components in \((x,y)\) is odd, and equals \(+1\) otherwise. First show that \(\text{sgn}(x,y) = \text{sgn}(x)\text{sgn}(y)\). Then prove that:

\[
\mu_F \left( \prod_{i=1}^{j+k}(a_i, b_i) \right) = \sum_x \text{sgn}(x)F_X(x) \sum_y \text{sgn}(y)F_Y(y),
\]

where the first sum is over the vertexes of \(A_X \equiv \prod_{i=1}^{j}(a_i, b_i)\), the second over the vertexes of \(A_Y \equiv \prod_{i=j+1}^{j+k}(a_i, b_i)\). Hint: Note for any fixed \(x\) that \(\sum_{(x,y)}\) is a summation over all the vertexes of \(A_Y\) and thus \(\sum_{(x,y)} = \sum_x \sum_y\).

By the uniqueness theorem of proposition 6.14 of book 1, the complete product space induced by \(F\), \((\mathbb{R}^{j+k}, \mathcal{M}_{\mu \times \nu}(\mathbb{R}^{j+k}), \mu \times \nu)\) is the product space of \((\mathbb{R}^j, \mathcal{M}_\mu(\mathbb{R}^j), \mu)\) and \((\mathbb{R}^k, \mathcal{M}_\nu(\mathbb{R}^k), \nu)\), the measure spaces induced by \(F_X\) and \(F_Y\). And as before, \(\mathcal{B}(\mathbb{R}^{j+k}) \subset \mathcal{M}_{\mu \times \nu}(\mathbb{R}^{j+k})\).

For the next result, recall the definition of the cross-sections of \(A \in \mathcal{B} (\mathbb{R}^{j+k})\) in definition 5.8 of book 5:

- **x-cross-section**: \(A_x = \{y | (x,y) \in A\}\), \(y\)-cross-section: \(A_y = \{x | (x,y) \in A\}\).
2.1 PRODUCT INTEGRALS AND CAVALIERI’S PRINCIPLE

Proposition 2.2 (Cavalieri’s Principle) Let \( X, Y \) be independent random vectors defined on a probability space \( (S, \mathcal{E}, \lambda) \):

\[
X : S \to \mathbb{R}^j, \quad Y : S \to \mathbb{R}^k,
\]

with induced Borel measures \( \mu \) and \( \nu \) defined on \( \mathbb{R}^j \) and \( \mathbb{R}^k \), respectively. Then for all \( A \in \mathcal{B}(\mathbb{R}^{j+k}) \):

\[
\Pr[(X, Y) \in A] = \int_{\mathbb{R}^j} \nu(A_x) d\mu(x) = \int_{\mathbb{R}^k} \mu(A_y) d\nu(y). \quad (2.2)
\]

Further, given \( B \in \mathcal{B}(\mathbb{R}^j) \),

\[
\Pr[X \in B, (X, Y) \in A] = \int_B \nu(A_x) d\mu(x), \quad (2.3)
\]

and analogously for \( B \in \mathcal{B}(\mathbb{R}^k) \):

\[
\Pr[Y \in B, (X, Y) \in A] = \int_B \mu(A_y) d\nu(y).
\]

Proof. If \( A \in \mathcal{B}(\mathbb{R}^{j+k}) \), then recalling notation 1.3 and 1.8 of remark 1.6:

\[
\Pr[(X, Y) \in A] = \lambda[(X, Y)^{-1}(A)] \equiv \mu \times \nu[A].
\]

This product measure of \( A \) is finite and given by:

\[
\mu \times \nu[A] = \int_{\mathbb{R}^{j+k}} \chi_A(x, y) d(\mu \times \nu),
\]

where \( \chi_A \) denotes the characteristic function of \( A \). Now \( A \in \mathcal{B}(\mathbb{R}^{j+k}) \subset \mathcal{M}_{\mu \times \nu}(\mathbb{R}^{j+k}) \) and \( \chi_A(x, y) \) is nonnegative and measurable relative to \( \mathcal{B}(\mathbb{R}^{j+k}) \) and hence \( \mathcal{M}_{\mu \times \nu}(\mathbb{R}^{j+k}) \). Thus we can apply either Fubini’s or Tonelli’s theorem of book 5 since both \( (\mathbb{R}^j, \mathcal{M}_\mu(\mathbb{R}^j), \mu) \) and \( (\mathbb{R}^k, \mathcal{M}_\nu(\mathbb{R}^k), \nu) \) are \( \sigma \)-finite, indeed finite, and complete. From either we conclude that:

\[
\mu \times \nu[A] = \int_{\mathbb{R}^k} \left[ \int_{\mathbb{R}^j} \chi_A(x, y) d\mu(x) \right] d\nu(y) = \int_{\mathbb{R}^j} \left[ \int_{\mathbb{R}^k} \chi_A(x, y) d\nu(y) \right] d\mu(x).
\]

Fubini’s theorem assures that \( \chi_A(\cdot, y) \) is \( \nu \)-integrable \( \mu \text{-a.e.} \) and \( \chi_A(x, \cdot) \) is \( \mu \)-integrable \( \nu \text{-a.e.} \). Now by definition:

\[
\mu(A_y) = \int_{\mathbb{R}^j} \chi_A(x, y) d\mu(x), \quad \nu(A_x) = \int_{\mathbb{R}^k} \chi_A(x, y) d\nu(y),
\]
which by Fubini’s theorem are defined defined \( \nu \)-a.e. and \( \mu \)-a.e., respectively, and are integrable. This proves 2.2.

A similar application of Fubini provides 2.3:

\[
\Pr \left[ X \in B, \ (X,Y) \in A \right] = \lambda \left[ (X,Y)^{-1}(A) \cap (B \times \mathbb{R}^k) \right] = \mu \times \nu \left[ A \cap (B \times \mathbb{R}^k) \right].
\]

Thus:

\[
\Pr \left[ X \in B, \ (X,Y) \in A \right] = \int_{\mathbb{R}^{j+k}} \chi_B(x) \chi_A(x,y) d(\mu \times \nu)
= \int_{\mathbb{R}^j} \chi_B(x) \nu(A_x) d\mu(x)
= \int_B \nu(A_x) d\mu(x).
\]

\[\blacksquare\]

**Remark 2.3** The above result is on the one hand a probabilistic interpretation of Fubini-Tonelli, and on the other hand is a statement that has an intuitive application in what is known as **Cavalieri’s principle.** This principle is named for **Bonaventura Cavalieri** (1598 – 1647) who developed these ideas before the formality of an integral calculus was developed. In the context of the volume of two 3-dimensional objects labeled \( A \) and \( A' \), this principle states the following. If the area of every horizontal cross-section of these objects agrees, where cross-sections are defined as the intersections of the objects with a complete set of parallel planes, then the volumes agree. Formally, if \( \mu(A_x) = \mu(A'_x) \) for all \( x \) then \( \text{Vol}(A) = \text{Vol}(A') \). The same result applies to the respective areas of two plane figures.

**Example 2.4** 1. The formula for the area of a triangle is \( \frac{1}{2}bh \), where \( b \) denotes the length of the base, and \( h \) the vertical height measured from the line that contains the base. This area is the same independent of the location of the vertex opposite the base, as long as \( h \) is fixed. Similarly, the volume of a circular cone is \( \frac{1}{3}\pi r^2h \), where \( r \) is the radius of the base and \( h \) the vertical height measured from the plane that contains the base. Again the volume is independent of the location of the top vertex, as long as \( h \) is fixed. These are both applications of Cavalieri’s principle since one observes that cross-sectional results agree in each case.

2. For a probability theory application, let \( X \) and \( Y \) denote independent random variables with \( X \) uniformly distributed on \([0,1]\), and \( Y \)
2.2 DFS OF SUMS AND RATIOS OF RVS

a standard normal variable. Define \( A = \{ Y \leq X \} \). Given \( x \in [0, 1] \), \( A_x = \{ Y \leq x \} \) and \( \nu(A_x) = \Phi(x) \) where \( \Phi \) is the distribution function for the standard normal. Thus by 2.2:

\[
\mu \times \nu [A] = \int_0^1 \Phi(x)dx \approx 0.68437.
\]

Alternatively, \( A_y = \{ y \leq X \} \) and:

\[
\mu(A_y) = \begin{cases} 
1, & y \leq 0, \\
1 - y, & 0 \leq y \leq 1, \\
0, & 1 \leq y,
\end{cases}
\]

so also:

\[
\mu \times \nu [A] = \int_{-\infty}^1 \mu(A_y)d\nu.
\]

Now the Borel measure \( \nu \) is defined relative to the standard normal distribution function:

\[
\nu[(-\infty, x]] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2}dy,
\]

and by proposition 3.6 of book 5:

\[
\mu \times \nu [A] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^1 \mu(A_y)e^{-y^2/2}dy
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-y^2/2}dy + \frac{1}{\sqrt{2\pi}} \int_0^1 (1 - y)e^{-y^2/2}dy
\]

\[
\approx 0.68437.
\]

2.2 DFs of Sums and Ratios of RVs

2.2.1 Sums of Independent Random Vectors

Book 4 introduced results for the distribution and density functions for sums of independent random variables in the special cases of absolutely continuous and discrete random variables. This introduction was somewhat informal, but identified the formal mathematics that this development required. With the aid of the mathematical tools of book 5, we now fill in these details and develop the general case.
Let $X$ and $Y$ be independent random vectors on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$, 
\[ X, Y : \mathcal{S} \rightarrow \mathbb{R}^j, \]
with respective joint distribution functions $F_X$ and $F_Y$ and induced Borel measures $\mu$ and $\nu$ on $\mathbb{R}^j$. If taken as a combined random vector $(X, Y) : \mathcal{S} \rightarrow \mathbb{R}^{2j}$ with joint distribution function $F$, the Borel measure on $\mathbb{R}^{2j}$ induced by $F$ is given by $\mu \times \nu$ due to independence as noted in 2.1 and the subsequent discussion. That is, the Borel space induced by $F$ is the product of the Borel spaces induced by $F_X$ and $F_Y$. This implies that for $A \in \mathcal{B}(\mathbb{R}^j)$:
\[
\lambda \left[ (X, Y)^{-1} [A] \right] \equiv (\mu \times \nu) [A]. \tag{2.4}
\]

Define $Z : \mathcal{S} \rightarrow \mathbb{R}^j$ by $Z \equiv X + Y$. Formally, $Z = f(X, Y)$ with $f(x, y) = x + y$:
\begin{align*}
(X, Y) : (\mathcal{S}, \mathcal{E}, \lambda) &\rightarrow (\mathbb{R}^{2j}, \mathcal{B}(\mathbb{R}^{2j}), m^{2j}), \\
f : (\mathbb{R}^{2j}, \mathcal{B}(\mathbb{R}^{2j}), m^{2j}) &\rightarrow (\mathbb{R}^j, \mathcal{B}(\mathbb{R}^j), m^j).
\end{align*}

Given a Borel set $B \in \mathcal{B}(\mathbb{R}^j)$, we will with the aid of 2.2 determine:
\[
\Pr[Z \in B] \equiv \lambda \left[ Z^{-1}[B] \right],
\]
or more formally:
\[
\Pr[Z \in B] = \lambda \left[ (X, Y)^{-1} f^{-1}[B] \right].
\]

If $B \in \mathcal{B}(\mathbb{R}^j)$ then $f^{-1}(B) \in \mathcal{B}(\mathbb{R}^{2j})$ since $f(x, y) = x + y$ is a Borel measurable function. Thus by 2.4:
\[
\Pr[Z \in B] = (\mu \times \nu) \left[ f^{-1}[B] \right]. \tag{2.5}
\]

Defining $A \equiv f^{-1}(B)$, so $A = \{(x, y) | x + y \in B\}$, the cross-sections of $A$ are:
\[
A_x = \{y | y \in B - x\}, \quad A_y = \{x | x \in B - y\},
\]
where $B - x$ and $B - y$ are defined as translates of the Borel set $B$ by $x$ and $y$ respectively. For example:
\[
B - x \equiv \{b - x | b \in B\}.\]
Definition 2.5 Denote by $\mu \ast \nu$ the set function defined on $B \in \mathcal{B}(\mathbb{R}^j)$ by the identity:

$$(\mu \ast \nu)[B] \equiv (\mu \times \nu)[f^{-1}(B)]. \quad (2.6)$$

The "measure" $\mu \ast \nu$ is called the **convolution of the Borel measures** $\mu$ and $\nu$.

Exercise 2.6 Show directly that $\mu \ast \nu$ is a Borel measure. Alternatively, this exercise can be avoided by noting that in the terminology and notation of definition 3.9 of book 5:

$$(\mu \ast \nu) \equiv (\mu \times \nu)_f.$$  

In other words, $\mu \ast \nu$ is the measure on $\mathcal{B}(\mathbb{R}^j)$ **induced by the measurable transformation** $f : (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), m^2) \to (\mathbb{R}^1, \mathcal{B}(\mathbb{R}), m^1)$, and it is indeed a measure by that book's proposition 3.10.

By 2.6 $(\mu \ast \nu)[B] \equiv (\mu \times \nu)[A]$, and by 2.2 recalling the cross-sections above:

$$(\mu \ast \nu)[B] = \int_{\mathbb{R}^j} \mu(B - y)d\nu(y) = \int_{\mathbb{R}^j} \nu(B - x)d\mu(x), \quad (2.7)$$

where the latter integral is sometimes denoted $\nu \ast \mu$. Thus

$$\mu \ast \nu = \nu \ast \mu. \quad (2.8)$$

Exercise 2.7 Prove that convolution is also associative:

$$(\mu \ast \nu) \ast \omega = \mu \ast (\nu \ast \omega). \quad (2.9)$$


Proposition 2.8 Let $X$ and $Y$ be independent random vectors on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$, so $X, Y : \mathcal{S} \to \mathbb{R}^j$, with respective joint distribution functions $F_X$ and $F_Y$ and respective induced Borel measures $\mu$ and $\nu$. Define $Z \equiv X + Y$. Then expressed as a Lebesgue-Stieltjes integral, the joint distribution function of $Z$ is given by:

$$F_Z(z) = \int_{\mathbb{R}^j} F_X(z - y)d\nu(y) = \int_{\mathbb{R}^j} F_Y(z - x)d\mu(x), \quad (2.10)$$

or in alternative notation:

$$F_Z(z) = \int_{\mathbb{R}^j} F_X(z - y)dF_Y(y) = \int_{\mathbb{R}^j} F_Y(z - x)dF_X(x). \quad (2.11)$$
CHAPTER 2 DFS OF TRANSFORMED RANDOM VECTORS

If \( F_Y \) has an associated joint density function \( f_Y \), then \( F_Z \) in 2.10 is given by the Lebesgue integral:

\[
F_Z(z) = \int_{\mathbb{R}^j} F_X(z - y)f_Y(y)dy,
\]

(2.12)

with an analogous result if \( F_X \) has a joint density function.

Proof. Let \( z = (z_1, \ldots, z_j) \) and \( B = \prod_{i=1}^j (-\infty, z_i] \). Then by 2.5 and 2.6 \( F_Z(z) = (\mu * \nu) |B| \). Also in 2.7 \( \mu(B - y) = F_X(z - y) \), and so:

\[
F_Z(z) = \int_{-\infty}^{\infty} F_X(z - y)dv(y),
\]

the first expression in 2.10. The second expression in 2.10 is derived analogously.

If \( Y \) has a joint density function, then \( \nu(A) = \int_A f_Y(t)dt \) as a Lebesgue integral by 1.24, and so by proposition 3.6 of book 5,

\[
\int_{\mathbb{R}^j} F_X(z - y)dv(y) = \int_{\mathbb{R}^j} F_X(z - y)f_Y(y)dy,
\]

which is 2.12. \( \blacksquare \)

Remark 2.9 For the proof of the next result, we require a proof that \( f_X(z - y) \) is Lebesgue measurable on \( \mathbb{R}^2 \) when \( f_X(x) \) is Lebesgue measurable on \( \mathbb{R}^j \). The case \( j = 1 \) was proved in proposition 6.16 and example 6.17 of book 5. The proofs given there apply in this general context by making the following modifications:

1. Proof of Proposition 6.16 (book 5):

   (a) The reference to proposition 2.42 of book 1 must be changed to reference corollary 7.23 of book 1.

   (b) The reference to proposition 3.30 of book 1 must be changed to reference proposition 1.5 of book 5.

   (c) The reference to proposition 3.49 of book 1 must be changed to reference proposition 1.18 of book 5.

   (d) The reference to proposition 3.47 of book 1 must be changed to reference corollary 1.10 of book 5.
2.2 DFS OF SUMS AND RATIOS OF RVS

2. Proof in Example 6.17 (book 5): This derivation requires one change: the reference to proposition 2.43 of book 1 must be changed to reference proposition 7.26 of book 1. Also recall that all measures are continuous from above using the proof of proposition 2.44 of book 1, and changing only notation.

Corollary 2.10 Let $X$ and $Y$ be given independent random vectors on a probability space $(S, \mathcal{E}, \lambda)$ as in proposition 2.8. If both $F_X$ and $F_Y$ have density functions, then $F_Z$ has a density function expressed as a Lebesgue integral:

$$f_Z(z) = \int_{\mathbb{R}^j} f_X(z-y)f_Y(y)dy,$$

with an analogous representation with the roles of $f_X$ and $f_Y$ reversed.

Proof. If $X$ has a density function, then with $A_z \equiv \prod_{i=1}^j (-\infty, z_i]$:

$$F_X(z-y) \equiv \int_{A_{z-y}} f_X(t)dt = \int_{A_z} f_X(t-y)dt.$$

The second integral follows from proposition 3.34 of book 5 where $T : x \rightarrow x - y$. So by 2.12:

$$F_Z(z) = \int_{\mathbb{R}^j} \int_{A_z} f_X(t-y)f_Y(y)dtdy.$$

By remark 2.9, $f_X(t-y)$ is Lebesgue measurable as a function defined on $\mathbb{R}^2$, while $f_Y(y)$ is Lebesgue measurable on $\mathbb{R}^j$ and thus $\mathbb{R}^{2j}$ by definition of a density function. Since this integrand is nonnegative and this iterated integral is finite by Cavalieri’s theorem, an application of Tonelli’s theorem of proposition 5.22 of book 5 allows the reversal of the iterated integrals to obtain:

$$F_Z(z) = \int_{A_z} \left[ \int_{\mathbb{R}^j} f_X(t-y)f_Y(y)dy \right] dt.$$

Thus $f_Z(z)$ as given in 2.13 is the density function associated with $F_Z$ if this function of $z$ can be shown to be Lebesgue measurable. But this then also follows from Tonelli’s theorem. The expression with the roles of $f_X$ and $f_Y$ reversed is derived by a change of variables.

Notation 2.11 The expression for the distribution function $F_Z$ in 2.12 gives the convolution of $F_X$ and $f_Y$, denoted $F_X * f_Y(z)$ in definition 6.14 of book 5:

$$F_Z(z) = F_X * f_Y(z).$$
CHAPTER 2 DFS OF TRANSFORMED RANDOM VECTORS

This earlier definition applied to the convolution of integrable functions on $\mathbb{R}$, so $j = 1$, but as noted above this convolution is well defined on $\mathbb{R}^j$ if one function is integrable and one is bounded.

Analogously, the expression for the density function $f_Z$ in 2.12 gives the convolution of $f_X$ and $f_Y$ in the sense of book 5:

$$f_Z(z) = f_X * f_Y(z).$$

In this case of book 5 with $j = 1$, proposition 6.18 of book 5 assures that if both $f_X$ and $f_Y$ are Lebesgue integrable, then so too is $f_X * f_Y$, and by 6.10 of that result the density function $f_Z(z)$ integrates to 1 as required. The proof of the generalization of this result to $\mathbb{R}^j$ is left as an exercise.

In the more general case of 2.11, the joint distribution function of $Z$ is represented as:

$$F_Z(z) = F_X * F_Y(z). \quad (2.14)$$

Of course since neither is integrable, this convolution cannot exist in the sense of definition 6.14 of book 5, and thus it is always the case that such notation is to be interpreted in the sense of 2.11.

Proposition 2.12 Let $f(x)$ and $g(x)$ be integrable functions on the Lebesgue measure space $(\mathbb{R}^j, \mathcal{M}^j, m^j)$. Then:

1. The function $f(x - y)g(y)$ is $m^j(y)$-integrable for almost all $x$. That is,

$$\int |f(x - y)g(y)| \, dm^j(y) < \infty, \ m^j\text{-a.e.}$$

2. Defining $f * g(x)$ as in 2.13, then $f * g(x)$ is $m^j$-integrable with:

$$\int |f * g(x)| \, dm^j(x) \leq \int |f(x)| \, dm^j \int |g(y)| \, dm^j, \quad (2.15)$$

and

$$\int f * g(x) \, dm^j(x) = \int f(x) \, dm^j \int g(y) \, dm^j. \quad (2.16)$$

**Proof.** Left as an exercise to check that proposition 6.18 of book 5 generalizes since Fubini’s and Tonelli’s theorems still apply. ■

We provide one example here, noting that many examples can be found in chapter 1 of book 4. The primary objective of this section was to formally justify the earlier derivations.
2.2 DFS OF SUMS AND RATIOS OF RVS

Example 2.13 Sums of Exponentials are Gamma: Let \( \{X_k\} \) be independent exponentially distributed random variables with parameter \( \lambda \) and density function as in 1.20 of book 4, so \( f(x) = \lambda e^{-\lambda x} \). We generalize example 1.18 and solve exercise 1.19 of book 4 to show that \( Z \equiv \sum_{k=1}^{n} X_k \) has a gamma distribution as in 1.22 of that book with parameters \( \lambda \) and \( \alpha = n \). To show this, denote by \( f_k(z) \) the density function of the sum of \( k \) independent exponential variables, then by 2.13 it follows that

\[
f_k(z) = \int_0^\infty f_{k-1}(z-x)f_1(x)dx.
\]

However, if \( g_k(z) \) denotes the density function of a gamma variable with parameters \( \lambda \) and \( \alpha = k \), then recalling that \( \Gamma(k-1) = (k-2)! \), a calculation shows that for \( z > 0 \):

\[
\int_0^\infty g_{k-1}(z-x)g_1(x)dx = \frac{\lambda^k}{(k-2)!} \int_0^z (z-x)^{k-2}e^{-(z-x)}e^{-\lambda x}dx
\]

\[
= \frac{\lambda^k e^{-\lambda z}}{(k-2)!} \int_0^z w^{k-2}dw
\]

\[
= g_k(z).
\]

Since \( f_1(x) = g_1(x) = f(x) \), we have by induction that \( f_k(z) = g_k(z) \).

2.2.2 Ratios of Independent Random Variables

Book 4 also introduced formulas for the distribution and density functions for ratios of independent random variables in the special cases of absolutely continuous and discrete random variables. This introduction was again somewhat informal, but referenced the formal mathematics required. With the aid of the mathematical tools of book 5, we now fill in these details.

Let \( X \) and \( Y \) be independent random variables on a probability space \((\mathcal{S}, \mathcal{E}, \lambda)\),

\[
X, Y : \mathcal{S} \to \mathbb{R},
\]

with respective distribution functions \( F_X \) and \( F_Y \) and induced Borel measures \( \mu \) and \( \nu \). Define \( Z : \mathcal{S} \to \mathbb{R} \) by \( Z \equiv X/Y \), and assume that \( \lambda \{Y^{-1}[0]\} \equiv \nu [0] = 0 \). Formally, \( Z = g(X,Y) \) with \( g(x,y) = x/y \):

\[
(X,Y) : (\mathcal{S}, \mathcal{E}, \lambda) \to (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), m^2), \quad g : (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), m^2) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), m).
\]

Given a Borel set \( B \in \mathcal{B}(\mathbb{R}) \), we will with the aid of 2.2 determine:

\[
\Pr[Z \in B] = \lambda \{Z^{-1}[B]\},
\]
or more formally:

\[ \Pr[Z \in B] = \lambda \left[ (X, Y)^{-1} g^{-1}[B] \right]. \]

If \( B \in \mathcal{B}(\mathbb{R}) \) then \( g^{-1}(B) \in \mathcal{B}(\mathbb{R}^2) \) since \( g(x, y) = x/y \) is a continuous and thus Borel measurable function for \( y \neq 0 \). Thus as in 2.5:

\[ \Pr[Z \in B] = (\mu \times \nu) [g^{-1}[B]]. \]

Defining \( A \equiv g^{-1}(B) \), so \( A = \{(x, y)|x/y \in B\} \), note that \( A \subset \mathbb{R}^2 - \{(0, y)|y \neq 0\} \) for all such \( B \). Now since \( \nu[0] = 0 \), it follows that

\[ (\mu \times \nu) [A] = (\mu \times \nu) [A - (0, 0)], \]

and thus:

\[ (\mu \times \nu) [A] = (\mu \times \nu) [A^+] + (\mu \times \nu) [A^-], \quad (2.17) \]

where \( A^+ \), respectively \( A^- \), denote \( A \) restricted to \( y > 0 \), respectively, \( y < 0 \). Also, the \( x \)-cross-sections of \( A^\pm \) are:

\[ A^+_y = \{x|x \in By, \ y > 0\}, \quad A^-_y = \{x|x \in By, \ y < 0\}, \]

where \( By \) is defined as the dilation/contraction of the Borel set \( B \) by \( y \). In other words:

\[ By \equiv \{by|b \in B\}. \]

**Proposition 2.14** Let \( X \) and \( Y \) be independent random variables on a probability space \((S, \mathcal{E}, \lambda)\), so \( X, Y : S \rightarrow \mathbb{R} \), with respective distribution functions \( F_X \) and \( F_Y \) and induced Borel measures \( \mu \) and \( \nu \) where \( \lambda[Y^{-1}[0]] = \nu[0] = 0 \). Define \( Z \equiv X/Y \). Then expressed as a Lebesgue-Stieltjes integral:

\[ F_Z(z) = \int_{-\infty}^{0} \left[ 1 - F_X(yz^-) \right] d\nu(y) + \int_{0}^{\infty} F_X(yz) d\nu(y), \quad (2.18) \]

where \( F_X(yz^-) \) denotes the left limit of \( F_X \) at \( yz \).

If \( Y \) has a density function \( f_Y(y) \), then expressed as a Lebesgue integral:

\[ F_Z(z) = \int_{-\infty}^{0} \left[ 1 - F_X(yz^-) \right] f_Y(y)dy + \int_{0}^{\infty} F_X(yz)f_Y(y)dy, \quad (2.19) \]

**Proof.** For \( B \in \mathcal{B}(\mathbb{R}) \), let \( A \equiv g^{-1}(B) \) in 2.17, and note that by 2.5 and then 2.2:

\[ \Pr[Z \in B] = (\mu \times \nu) [A^+] + (\mu \times \nu) [A^-] \]

\[ = \int_{-\infty}^{0} \mu(A^+_y) d\nu(y) + \int_{0}^{\infty} \mu(A^-_y) d\nu(y). \]
These integrals are well defined since \( \nu[0] = 0 \). By taking \( B = (-\infty, z] \), note that:

\[
A^+_y = (-\infty, zy], \quad A^-_y = [zy, \infty),
\]

and thus \( \nu(A^-_y) = \nu(zy, \infty) \equiv 1 - F_X(yz^-) \) and \( \mu(A^+_y) \equiv F_X(yz) \). This then produces 2.18.

If \( Y \) has a density function, \( F_Y(y) = \int_{-\infty}^{y} f_Y(t) dt \) as a Lebesgue integral, and so 2.19 follows by proposition 3.6 of book 5.

**Remark 2.15** Note that 2.19 is equivalent to 1.50 of book 4 when \( Y \) has range \((0, \infty)\). Similarly 2.20 is 1.52 of book 4 in this case.

**Corollary 2.16** Let \( X \) and \( Y \) be given independent random variables on a probability space \((S, \mathcal{E}, \lambda)\) as in proposition 2.14 and define \( Z = X/Y \). If both \( F_X \) and \( F_Y \) have density functions then \( F_Z \) has a density function expressed as a Lebesgue integral:

\[
f_Z(z) = \int_{-\infty}^{\infty} |y| f_X(yz) f_Y(y) dy.
\]

**Proof.** If \( X \) has a density function \( f_X \), then \( F_X \) is continuous by proposition 3.33 of book 3, so by 2.19:

\[
F_Z(z) = \int_{-\infty}^{0} \left[ \int_{y}^{\infty} f_X(x) dx \right] f_Y(y) dy + \int_{0}^{\infty} \left[ \int_{-\infty}^{y} f_X(x) dx \right] f_Y(y) dy.
\]

Define a transformation \( T \) on \( \mathbb{R} \) by \( T : x \to yx \), so \( |T'(x)| = |y| \). Then by the Lebesgue change of variable formula of proposition 3.34 of book 5, and \( B = [yz, \infty) \) with \( y < 0 \):

\[
\int_{yz}^{\infty} f_X(x) dx = \int_{B} f_X(x) dx = \int_{T^{-1}B} f_X(yx) |y| dx = -\int_{-\infty}^{z} y f_X(yx) dx.
\]

Similarly with \( B = (-\infty, yz] \) and \( y > 0 \):

\[
\int_{-\infty}^{y} f_X(x) dx = \int_{B} f_X(x) dx = \int_{T^{-1}B} f_X(yx) |y| dx = \int_{-\infty}^{z} y f_X(yx) dx.
\]

Combining:

\[
F_Z(z) = -\int_{-\infty}^{0} \int_{-\infty}^{z} y f_X(yx) f_Y(y) dy dx dy + \int_{0}^{\infty} \int_{-\infty}^{z} y f_X(yx) f_Y(y) dy dx dy.
\]
As both integrands are nonnegative and Lebesgue measurable, and these iterated integrals are finite, we can apply Tonelli’s theorem of book 5 to produce:

\[ F_Z(z) = \int_{-\infty}^{z} \left[ \int_{0}^{\infty} y f_X(yx) f_Y(y) dy - \int_{-\infty}^{0} y f_X(yx) f_Y(y) dy \right] dx \]

and thus:

\[ f_Z(z) = \int_{0}^{\infty} y f_X(yz) f_Y(y) dy - \int_{-\infty}^{0} y f_X(yz) f_Y(y) dy \]

\[ = \int_{-\infty}^{\infty} |y| f_X(yz) f_Y(y) dy. \]

\[ \text{Example 2.17} \quad \text{Assume that } X \text{ and } Y \text{ are independent unit normal variables with density given in 3.2 by } \phi(x) = \exp \left(-\frac{x^2}{2}\right) / \sqrt{2\pi}, \text{ and define } Z = X/Y. \text{ Since } \lambda \left[ Y^{-1}[0]\right] = \nu \left[ 0 \right] = 0, \text{ we can apply 2.20 to conclude that:} \]

\[ f_Z(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} |y| \exp \left(-\left(1 + z^2\right) \frac{y^2}{2}\right) dy \]

\[ = \frac{1}{\pi} \int_{0}^{\infty} y \exp \left(-\left(1 + z^2\right) \frac{y^2}{2}\right) dy. \]

This is a Lebesgue integral, but because the integrand is continuous and exists as a Riemann integral, proposition 2.64 of book 3 assures that these integral values agree. This integral can be evaluated using a simple substitution of \( u = \left(1 + z^2\right) \frac{y^2}{2} \) (recall example 3.1 of book 5), and so \( du = \left(1 + z^2\right) y dy \) and:

\[ f_Z(z) = \frac{1}{\pi \left(1 + z^2\right)} \int_{0}^{\infty} \exp(-u) du \]

\[ = \frac{1}{\pi \left(1 + z^2\right)} . \]

Recalling the Cauchy density function in 3.64 of book 4 and named for Augustin Louis Cauchy (1789 – 1857), we have derived that the ratio of independent standard normal variates is standard Cauchy.
2.3 DENSITIES OF TRANSFORMED RANDOM VECTORS

2.3 Densities of Transformed Random Vectors

Section 1.3 of book 4 introduced the question of determining the distribution function for the random variable \( Y \equiv g(X) \), where \( X : S \to \mathbb{R} \) is a random variable defined on \( (S, \mathcal{E}, \lambda) \) with associated distribution function \( F(x) \), and \( g : \mathbb{R} \to \mathbb{R} \) is Borel measurable. Formulas for \( F_Y(y) \) were derived in 1.38 for monotonically increasing \( g \); and in 1.40 for monotonically decreasing \( g \). When \( F(x) \) is absolutely continuous with density function \( f(x) \), the density function for \( Y \) exists when \( g \) is monotonic and \( g^{-1} \) differentiable, and as in 1.42:

\[
f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|.
\] (2.21)

In this section we generalize the results on density functions to random vectors defined on \( (S, \mathcal{E}, \lambda) \) with the aid of the tools of chapter 3 of book 5. To this end, let \( X = (X_1, X_2, \ldots, X_n) \) be a random vector defined on \( (S, \mathcal{E}, \lambda) \), so

\[ X : S \to \mathbb{R}^n, \]

with distribution function \( F(x_1, x_2, \ldots, x_n) \), and let \( T : \mathbb{R}^n \to \mathbb{R}^n \) be a continuously differentiable, one-to-one transformation with \( \det(T'(x)) \neq 0 \). Recall that \( T'(x) \) denotes the Jacobian matrix for \( T \), and is defined as follows.

For \( x = (x_1, x_2, \ldots, x_n) \), let

\[ T(x) = (t_1(x), t_2(x), \cdots, t_n(x)), \]

where \( t_i : \mathbb{R}^n \to \mathbb{R} \) for all \( i \). We will say that \( T \) is continuous, differentiable, continuously differentiable, etc., if the component functions \( \{t_i\}_{i=1}^n \) have such properties. When this transformation has differentiable component functions, the Jacobian matrix associated with \( T \), denoted \( T'(x) \), is defined as:

\[
T'(x) \equiv \begin{pmatrix}
\frac{\partial t_1}{\partial x_1} & \frac{\partial t_1}{\partial x_2} & \cdots & \frac{\partial t_1}{\partial x_n} \\
\frac{\partial t_2}{\partial x_1} & \frac{\partial t_2}{\partial x_2} & \cdots & \frac{\partial t_2}{\partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial t_n}{\partial x_1} & \frac{\partial t_n}{\partial x_2} & \cdots & \frac{\partial t_n}{\partial x_n}
\end{pmatrix}.
\] (2.22)

The Jacobian determinant associated with \( T \), denoted \( \det(T'(x)) \), is defined as the determinant of \( T'(x) \). This matrix and its determinant are named for Carl Gustav Jacob Jacobi (1804 – 1851), an early developer of determinants and their applications in analysis.
CHAPTER 2 DFS OF TRANSFORMED RANDOM VECTORS

Notation 2.18 The Jacobian matrix is denoted in many ways:
\[ T'(x) \equiv \frac{\partial(t_1, t_2, \ldots, t_n)}{\partial(x_1, x_2, \ldots, x_n)} = \left( \frac{\partial(t_1, t_2, \ldots, t_n)}{\partial(x_1, x_2, \ldots, x_n)} \right) = \left( \frac{\partial T}{\partial x} \right). \]

By 3.26 of proposition 3.34 of book 5, for any nonnegative Borel measurable function \( g : \mathbb{R}^n \to \mathbb{R}^n \), \( g(x) \) is Lebesgue integrable if and only if \( g(Tx) \mid \det(T'(x)) \mid \) is Lebesgue integrable, and when integrable,
\[
\int_A g(x) dm^n = \int_{T^{-1}A} g(Tx) \mid \det(T'(x)) \mid dm^n \tag{2.23}
\]
for all \( A \in \mathcal{B}(\mathbb{R}^n) \).

We now have the following generalization of 2.21.

Proposition 2.19 Assume that the distribution function \( F(x_1, x_2, \ldots, x_n) \) of \( X \equiv (X_1, X_2, \ldots, X_n) \) has an associated joint density function \( f(x_1, x_2, \ldots, x_n) \), and define \( Y = (Y_1, Y_2, \ldots, Y_n) \equiv g(X) \) where \( g : \mathbb{R}^n \to \mathbb{R}^n \) is continuously differentiable and one-to-one. Then the density function for \( Y \) is given by:
\[
f_Y(y) = f \left( g^{-1}(y) \right) \left\mid \det \left( \frac{\partial g^{-1}}{\partial y} \right) \right\mid, \quad m^n\text{-}a.e., \tag{2.24}
\]
where \( \frac{\partial g^{-1}}{\partial y} \) denotes the Jacobian matrix of \( g \) as defined in 2.22.

Proof. That \( (X_1, X_2, \ldots, X_n) \) has a joint density function \( f(x_1, x_2, \ldots, x_n) \) means that:
\[
F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(u_1, u_2, \ldots, u_n) du_1 du_2 \cdots du_n.
\]

Letting \( Y \equiv g(X) \) as above, then because \( g \) is continuous and hence Borel measurable, \( g^{-1}(A) \in \mathcal{B}(\mathbb{R}^n) \) for all \( A \in \mathcal{B}(\mathbb{R}^n) \). Further,
\[
(Y_1, Y_2, \ldots, Y_n) \in A \text{ if and only if } (X_1, X_2, \ldots, X_n) \in g^{-1}A,
\]
so as measurable sets in \( S \),
\[
\lambda \left[ (Y_1, Y_2, \ldots, Y_n)^{-1}(A) \right] = \lambda \left[ (X_1, X_2, \ldots, X_n)^{-1}(g^{-1}A) \right] = \int_{T^{-1}A} f(u_1, u_2, \ldots, u_n) dm^n,
\]
where we define the transformation \( T \equiv g^{-1} \).
2.3 DENSITIES OF TRANSFORMED RANDOM VECTORS

Because \( g \) is continuously differentiable and one-to-one, \( T \) is continuously differentiable by the inverse function theorem (the statement is summarized in section 3.3.3 of book 5). It now follows from 2.23 that:

\[
\lambda [(Y_1, Y_2, ..., Y_n)^{-1}(A)] = \int_A f(Tx) |\det(T'(x))| \, dm^n.
\]

Letting \( A_y \equiv \prod_{i=1}^n (-\infty, y_i] \) and noting that \( F_Y(y_1, y_2, ..., y_n) \equiv \lambda [(Y_1, Y_2, ..., Y_n)^{-1}(A_y)] \), we have:

\[
F_Y(y_1, y_2, ..., y_n) = \int_{A_y} f(Tx) |\det(T'(x))| \, dm^n.
\]

This integrand is nonnegative and Lebesgue measurable, and thus by definition 1.8 it follows that the density function of \( Y = g(X) \) is given \( m^n \)-a.e. by:

\[
f_Y(y) = f(Ty) |\det(T'(y))|.
\]

Substituting \( T = g^{-1} \) obtains 2.24: ■

As an application we generalize somewhat proposition 1.25 of book 4. Gamma and Beta distribution functions were discussed in section 1.2.2 of book 4.

**Proposition 2.20** Let \((X_1, X_2)\) be a random vector of independent gamma random variables with parameters \(\alpha_1, \lambda\) and \(\alpha_2, \lambda\), respectively. Define the random vector

\[
(Y_1, Y_2) \equiv g(X_1, X_2) = \left( \frac{X_1}{X_1 + X_2}, X_1 + X_2 \right).
\]

Then \( Y_1 \) and \( Y_2 \) are independent random variables, where \( Y_1 \) is beta with parameters \(\alpha_1\) and \(\alpha_2\), and \( Y_2 \) is gamma with parameters \(\alpha_1 + \alpha_2, \lambda\).

**Proof.** By 1.22 of book 4 and 1.38 we have that for \(x_1 > 0, x_2 > 0:\)

\[
f(x_1, x_2) = \lambda^{\alpha_1+\alpha_2} x_1^{\alpha_1-1} x_2^{\alpha_2-1} e^{-\lambda(x_1+x_2)} / [\Gamma(\alpha_1)\Gamma(\alpha_2)],
\]

where the gamma function \(\Gamma(\alpha)\) is defined for \(\alpha > 0:\)

\[
\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx.
\]

The inverse transformation is defined by \(g^{-1}(y_1, y_2) = (y_1 y_2, (1 - y_1) y_2)\), and so:

\[
\det \left( \frac{\partial g^{-1}}{\partial y} \right) = \det \begin{pmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{pmatrix} = y_2.
\]
CHAPTER 2 DFS OF TRANSFORMED RANDOM VECTORS

Now:
\[ g(\{x_1, x_2\} | x_1 > 0, x_2 > 0 \} = \{(y_1, y_2) | y_1 > 0, 0 < y_2 < 1\}, \]

and so from 2.24 we derive that
\[
\begin{align*}
f_Y(y_1, y_2) &= \lambda^{\alpha_1 + \alpha_2} (y_1 y_2)^{\alpha_1 - 1} ((1 - y_1) y_2)^{\alpha_2 - 1} y_2 e^{-\lambda y_2} / \Gamma(\alpha_1) \Gamma(\alpha_2) \\
&= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \Gamma(\alpha_2)} y_1^{\alpha_1 - 1} (1 - y_1)^{\alpha_2 - 1} \times \lambda^{\alpha_1 + \alpha_2} y_2^{\alpha_1 + \alpha_2 - 1} e^{-\lambda y_2} / \Gamma(\alpha_1 + \alpha_2).
\end{align*}
\]

Thus \( f_Y(y_1, y_2) \) is a product of density functions: beta with parameters \( \alpha_1 \) and \( \alpha_2 \) in \( y_1 > 0 \), and gamma with parameters \( \alpha_1 + \alpha_2, \lambda \) in \( y_2 \in (0, 1) \). Thus \( F_Y(y_1, y_2) \) is a product of the associated distribution functions, and independence follows from proposition 3.53 of book 2.
Chapter 3

Multivariate Normal Distribution

This chapter reflects a more extended application of the prior chapter on transformations of random vectors, but now to define and investigate the multivariate normal distribution. As in the one dimensional case discussed in section 1.2.2 of book 4, this distribution function is often called the multivariate Gaussian distribution, named for Carl Friedrich Gauss (1777 – 1855).

3.1 Derivation and Properties

In this section we derive various results relating to collections of normal variates. Recall that by convention we interpret vectors as column matrices (or column vectors) so if $A$ is an $n \times n$ matrix and $X$ and $\mu$ are $n$-vectors, then $AX$ is well defined and $AX + \mu$ is a (column) $n$-vector. Also given a matrix $A$, not necessarily square, $A^T$ denotes the transpose of $A$ and is defined in components by $A^T_{ij} = A_{ji}$. It is an exercise to check that $(AB)^T = B^T A^T$. Again this does not assume that $A$ and $B$ are square, only that $AB$ is well defined.

Introduced in books 2 and 4, the normal density function $f_N(x)$ depends on a location parameter $\mu \in \mathbb{R}$ and a scale parameter $\sigma > 0$, and is defined by:

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

(3.1)

where $\exp y = e^y$ to simplify notation. This is often call the Gaussian
probability density after Carl Friedrich Gauss (1777 – 1855) who was one of the codiscoverers of this formula. When \( \mu = 0 \) and \( \sigma = 1 \), this is known as the unit normal or standard normal density, and often denoted \( \phi(x) \):

\[
\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{3.2}
\]

See section 3.3 for a discussion of the following result when \( A \) is not invertible, or is an \( m \times n \) matrix with \( m \neq n \).

**Proposition 3.1 (Multivariate Normal distribution)** Let \( X \equiv (X_1, X_2, ..., X_n) \) be \( n \) independent unit normals with density functions as in 3.2. Let \( Y \equiv (Y_1, Y_2, ..., Y_n) \) be defined by \( Y = g(X) \equiv AX + \mu \), where \( A \) is an invertible linear transformation on \( \mathbb{R}^n \) and \( \mu = (\mu_1, \mu_2, ..., \mu_n) \) a constant vector. Then with \( y = (y_1, y_2, ..., y_n) \), the probability density function of \( Y \) is given:

\[
f_Y(y) = (2\pi)^{-n/2} |\det C|^{-1/2} \exp\left[ -\frac{1}{2}(y - \mu)^T C^{-1}(y - \mu) \right], \tag{3.3}
\]

where \( C = AA^T \), and \( \det C \) denotes the determinant of \( C \).

**Proof.** First express the density function of \( X \) in a format suitable to apply a transformation. Since \( f(x_1, x_2, ..., x_n) = \prod_{i=1}^n \phi(x_i) \) by proposition 1.21:

\[
f(x) = (2\pi)^{-n/2} \exp\left[ -\frac{1}{2} x^T I x \right],
\]

where \( I \) denotes the identity matrix, \( x^T \) the row matrix transpose of the column matrix \( x \), and thus:

\[
x^T I x = \sum_{i=1}^n x_i^2.
\]

The inverse of the transformation \( g \) is the linear transformation \( g^{-1}(y) = A^{-1}(y - \mu) \), and so the Jacobian determinant of \( g^{-1} \) is the determinant of the matrix which represents \( A^{-1} \). Recall that \( \det A^{-1} = 1/\det A \), and this is well defined since \( A \) is invertible and hence \( \det A \neq 0 \). Applying 2.24:

\[
f_Y(y) = (2\pi)^{-n/2} |\det A^{-1}| \exp\left[ -\frac{1}{2} (A^{-1}(y - \mu))^T I (A^{-1}(y - \mu)) \right]
\[
= \frac{1}{(2\pi)^{n/2}|\det A|} \exp\left[ -\frac{1}{2} (y - \mu)^T C^{-1}(y - \mu) \right].
\]

Using various matrix manipulations:

\[
C^{-1} = (A^{-1})^T I A^{-1} = (A^{-1})^T A^{-1} = (A^T)^{-1} A^{-1} = (AA^T)^{-1},
\]
3.1 DERIVATION AND PROPERTIES

and so \( C \equiv AA^T \). Then \( \det C = \det AA^T = (\det A)^2 \), so \( [\det C]^{-1/2} \) is well defined and the result follows.

**Definition 3.2** Generalizing definition 3.9 of book 4, define the moment generating function of \( Y \equiv (Y_1, Y_2, ..., Y_n) \) with distribution function \( F_Y(y) \) by:

\[
M_Y(t) \equiv \int_{\mathbb{R}^n} e^{ty} d\mu_{F_Y} \tag{3.4}
\]

when this Lebesgue-Stieltjes integral exists for \( |t| < t_0 \) for some \( t_0 > 0 \). Here \( t = (t_1, t_2, ..., t_n) \), \( |t|^2 = \sum_{j=1}^{n} t_j^2 \), and \( y \cdot t = \sum_{j=1}^{n} y_j t_j \) is the usual dot product, or inner product on \( \mathbb{R}^n \).

If \( Y \equiv (Y_1, Y_2, ..., Y_n) \) has a density function \( f_Y(y) \), then by proposition 3.6 of book 5 this moment generating function is equivalently defined:

\[
M_Y(t) \equiv \int_{\mathbb{R}^n} e^{ty} f_Y(y) dy, \tag{3.5}
\]

and this integral exists if and only if the integral in 3.4 exists.

**Remark 3.3** Note that the above definition already reflects a change of variables for expediency. That is, if \( Y \) is defined on a probability space \( (\mathcal{S}, \mathcal{E}, \lambda) \), the moment generating function is formally defined by:

\[
M_Y(t) \equiv \int_{\mathcal{S}} e^{tY} d\lambda. \tag{3.6}
\]

The Borel measure \( \lambda_Y \) on \( \mathbb{R}^n \) induced by \( Y \), as given in definition 3.9 of book 5, is then seen to be the Borel measure \( \mu_{F_Y} \) induced by the distribution function \( F \). Hence the integral in 3.6 becomes the integral in 3.4 using the change of variable result in proposition 3.14 of book 5.

**Exercise 3.4** Prove that the moment generating function of the multivariate normal \( Y \equiv (Y_1, Y_2, ..., Y_n) \) with density given in 3.3 is:

\[
M_Y(t) = \exp \left[ \mu \cdot t + \frac{1}{2} t^T Ct \right]. \tag{3.7}
\]

Hint: First show that for \( X \equiv (X_1, X_2, ..., X_n) \) as in proposition 3.1 that:

\[
M_X(t) = \exp \left[ \frac{1}{2} t^T t \right],
\]

where \( t^T t \equiv t \cdot t \). Then show that if \( Y = AX + \mu \), with \( A \) as in proposition 3.1:

\[
M_Y(t) = M_X(A^T t) \exp [\mu \cdot t].
\]
Remark 3.5 Corollary 3.58 of book 4 noted that for a distribution function \( F \) of a single random variable \( X \), the existence of \( M_X(t) \) on \((-t_0,t_0)\) for \( t_0 > 0 \) assured that \( F \) is the only distribution with this moment generating function. Its proof required the tools of characteristic functions to be discussed below in chapter 6, and this previous result will be formally settled with proposition 6.44.

A similar conclusion is true for the multivariate moment generating function, its proof again requiring chapter 6 results, but we will not formally develop this result in this book. We will instead in this chapter simply assume for expediency that the multivariate normal distribution is characterized by the moment generating function given in 3.7. A typical application of this result is found in the situation where we derive \( M_Y(t) \) and see that it has the form 3.7 with \( C \) invertible. Given uniqueness, we would then be able to declare that \( Y \) is a random vector with density function given in 3.3. See proposition 3.9.

To compensate the reader for this logic gap, it should be noted that in chapter 6 we will instead prove that characteristic functions uniquely determine distribution functions, and assign as an exercise the derivation of the characteristic function of the above multivariate normal. As a further exercise, the reader can then reformulate the above application in terms of characteristic functions, and then obtain the desired conclusion of proposition 3.9 without moment generating functions and this "leap of faith."

The other application of this uniqueness result is to justify the following definition. In definition 6.26, the following definition will be reformulated in terms of characteristic functions.

Definition 3.6 A random vector \( Y \equiv (Y_1,Y_2,...,Y_n) \) is said to have a multivariate normal distribution, or a multivariate Gaussian distribution, if there exists an \( n \times n \) symmetric, positive semidefinite matrix \( C \) and an \( n \)-vector \( \mu \) so that the moment generating function of \( Y \) exists for all \( t \) and has the form in 3.7:

\[
M_Y(t) = \exp \left[ \mu \cdot t + \frac{1}{2} t^T Ct \right].
\] (3.8)

The matrix \( C \) is said to be symmetric if \( C = C^T \), while positive semidefinite means that \( x^T C x \geq 0 \) for all \( x \).

Remark 3.7 By remark 3.5, if \( Y \) has a multivariate normal distribution by definition 3.6 and \( C \) is invertible, then the distribution function of \( Y \) has the density function given in 3.3. But note that this definition does not require
that $C$ be invertible, only symmetric and positive semidefinite. See section 3.3 for more on this.

**Example 3.8** By exercise 3.4 it follows that if $X \equiv (X_1, X_2, ..., X_n)$ are $n$ independent unit normals, $A$ an $n \times n$ invertible matrix, $A : \mathbb{R}^n \to \mathbb{R}^n$, and $\mu \equiv (\mu_1, \mu_2, ..., \mu_n)$ a constant vector, then $Y \equiv AX + \mu$ has a multivariate normal distribution. In this case the associated $C$ is invertible since $C = AA^T$ obtains that $C^{-1} = (AA^T)^{-1} = (A^{-1})^T A^{-1}$.

More generally by proposition 3.9 below, if $Y \equiv (Y_1, Y_2, ..., Y_n)$ is multivariate normally distributed by definition 3.6 and $B : \mathbb{R}^n \to \mathbb{R}^m$ is an $m \times n$ matrix, then $BY$ is multivariate normally distributed on $\mathbb{R}^m$.

**Proposition 3.9** Let $Y \equiv (Y_1, Y_2, ..., Y_n)$ be multivariate normally distributed by definition 3.6 and $B : \mathbb{R}^n \to \mathbb{R}^m$ an $m \times n$ matrix. Then $BY$ is multivariate normally distributed on $\mathbb{R}^m$ with

$$B = B, \quad CB = CB^T.$$ 

**Proof.** Let $W = TY$ where we define the transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ by $T : Y \to BY$. We determine $M_W(t)$ with an application of proposition 3.14 of book 5 with $g(w) = e^{w^T}$ defined on $\mathbb{R}^m$. Formally:

$$T : (\mathbb{R}^n, \mu_{F_Y}, \mathcal{B}(\mathbb{R}^n)) \to (\mathbb{R}^m, \mu_T, \mathcal{B}(\mathbb{R}^m)),$$

where the induced measure $\mu_T$ is defined on $D \in \mathcal{B}(\mathbb{R}^m)$ by:

$$\mu_T(D) \equiv \mu_{F_Y}(T^{-1}D).$$

Note that this definition does not require that $B$ be an invertible linear transformation, and indeed it can not be if $n \neq m$. Instead, here as in 3.7 of book 5 this general inverse is defined by:

$$T^{-1}D = \{x \in \mathbb{R}^n | Tx \in D\}.$$

Now $T^{-1}D \in \mathcal{B}(\mathbb{R}^n)$ since $T$ is continuous, and $\mu_T$ is a probability measure since $T^{-1}\mathbb{R}^m = \mathbb{R}^n$, and thus let $F_W$ denote the induced distribution function as in chapter 8 of book 1. Then by definition, $\mu_T = \mu_{F_W}$.

From 3.10 of proposition 3.14 of book 5 with $t \equiv (t_1, ..., t_m)$:

$$M_W(t) = \int_{\mathbb{R}^m} e^{t^Tw} d\mu_{F_W} = \int_{\mathbb{R}^n} e^{t^T(Ty)} d\mu_{F_Y} = \int_{\mathbb{R}^n} e^{(B^Tt)^Ty} d\mu_{F_Y}.$$
noting that \( t \cdot (By) = t^T By = (B^T t) \cdot y \). Thus by 3.8:

\[
M_W(t) = \exp \left[ \mu \cdot B^T t + \frac{1}{2} (B^T t)^T C (B^T t) \right] = \exp \left[ (B \mu) \cdot t + \frac{1}{2} t^T (BCB^T) t \right].
\]

By definition 3.6, this moment generating function is that of a multivariate normal distribution with \( \mu_B = B \mu \) and \( C_B = BCB^T \).

**Example 3.10** This proposition provides the simplest means of obtaining the various marginal distributions of a multivariate normal. For example, let \( B : \mathbb{R}^n \to \mathbb{R}^1 \) be defined by \( B : (Y_1, Y_2, ..., Y_n) \to Y_j \). Formally, \( B \) is given by a \( 1 \times n \) matrix: \( B = (0, ..., 0, 1, 0, ..., 0) \) with 1 in component \( j \), recalling the standard convention that \( (Y_1, Y_2, ..., Y_n) \) is treated as a column vector. Then \( Y_j \) has a normal distribution with parameters \( \mu_j \) and \( \sigma_j^2 \), which we recall from 3.65 of book 4 are the mean and variance of \( Y_j \) (see exercise 3.11 below). Similarly, given \( B : \mathbb{R}^n \to \mathbb{R}^2 \) with \( B : (Y_1, Y_2, ..., Y_n) \to (Y_i, Y_j) \), this latter vector is multivariate normal with vector \( \mu \equiv (\mu_i, \mu_j) \) and matrix \( C = \begin{pmatrix} c_{ii} & c_{ij} \\ c_{ij} & c_{jj} \end{pmatrix} \).

**Exercise 3.11** Show that if \( Y \) is a random vector with density function given in 3.3, then the means, variances and covariances of the variates are given by:

\[
E[Y_i] = \mu_i, \quad \text{Var}[Y_i] = c_{ii}, \quad \text{Cov}[Y_i, Y_j] = c_{ij},
\]

where \( (c_{ij}) \) denote the components of \( C \). These moments are defined in chapter 3 of book 4. Recall the mean and variance calculations require the marginal density function \( f(y_i) \), while the covariance requires the joint density \( f(y_i, y_j) \), all obtainable from example 3.10. Specifically:

\[
\text{Var}[Y_i] \equiv E[(Y_i - \mu_i)^2], \quad \text{Cov}[Y_i, Y_j] \equiv E[(Y_i - \mu_i) (Y_j - \mu_j)],
\]

and these are commonly denoted \( \sigma_i^2 \) and \( \sigma_{ij} \).

Prove this same result if \( Y \) is a multivariate normal random vector by definition 3.6. Here you will need to assume the result of proposition 6.44 of chapter 6, that if \( M_Y(t) \) exists for \( |t| < t_0 \) for \( t_0 > 0 \), then the moment \( E[y_1^{m_1} \cdots y_n^{m_n}] \) exists for all nonnegative integers \( m_j \), and with
3.1 DERIVATION AND PROPERTIES

\[ m \equiv \sum_{j=1}^{n} m_j : \]

\[ E[y_1^{m_1} \cdots y_n^{m_n}] = \left. \frac{\partial^{m} M_Y(t)}{\partial t_1^{m_1} \cdots \partial t_n^{m_n}} \right|_{t=0}. \quad (3.11) \]

**Remark 3.12** Given exercise 3.11, the density function \( f_Y(y) \) in 3.3 is called the **multivariate normal density** with **mean vector** \( \mu \) and **covariance matrix** \( C \). Important properties of \( C \) are that it is **symmetric**, so \( C = C^T \), and **positive definite**, so \( x^T C x > 0 \) unless \( x = 0 \). This latter property follows because:

\[ x^T C x = x^T A A^T x = (A^T x)^T A^T x = \| A^T x \|^2, \]

where \( \| A^T x \| \) denotes the usual **Euclidean norm** of the vector \( A^T x \), so for example:

\[ \| x \|^2 \equiv \sum_{i=1}^{n} x_i^2. \quad (3.12) \]

Thus \( C \) is positive definite when \( A \) is invertible since then \( A A^T x = 0 \) if and only if \( A^T x = 0 \) if and only if \( x = 0 \).

In the special case where \( A \) is diagonal, \( A_{ii} = \sigma_i \) (by exercise 3.11) and \( A_{ij} = 0 \) for \( i \neq j \), then \( C^{-1} \) is also diagonal with \( (C^{-1})_{ii} = 1/\sigma_i^2 \) and so \( \det C^{1/2} = \prod_{i=1}^{n} \sigma_i \). In this case the density for \( Y \) is the density for \( n \) independent normals with means \( \mu_i \) and standard deviations \( \sigma_i \):

\[ f_Y(y) = (2\pi)^{-n/2} [\prod_{i=1}^{n} \sigma_i]^{-1} \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu_i)^2 / \sigma_i^2 \right]. \quad (3.13) \]

Thus:

\[ f_Y(y) = \prod_{i=1}^{n} f_N(y_i) \]

with \( f_N(y_i) \) given in 3.1, and \( (Y_1, Y_2, \ldots, Y_n) \) are independent normal variates by proposition 3.53 of book 2.

The final result addresses the following question and extends the discussion of chapter 4 of book 4.

**Example 3.13** How can one simulate samples of normal vectors,

\[ \{(Y_1^{(j)}, Y_2^{(j)}, \ldots, Y_n^{(j)})\}_{j=1}^{N}, \]

with prescribed mean vector \( \mu \) and covariance matrix \( C \)?

If we could simulate samples of independent unit normal vectors,

\[ \{(X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)})\}_{j=1}^{N}, \]
and also factor the matrix \( C = AA^T \) with an invertible matrix \( A \), then proposition 3.1 gives the answer. With apparent notation simply define \( Y^{(j)} \equiv AX^{(j)} + \mu \), and then by exercise 3.11, \( \{ (Y^{(1)}_1, Y^{(1)}_2, \ldots, Y^{(n)}_n) \}_{j=1}^N \) has the appropriate distribution and moments.

As it turns out, it is relatively easy to generate independent unit normal variates using a variety of mathematical programming languages (including Microsoft Excel). But the factorization of \( C \) involves an obvious challenge. A general \( n \times n \) matrix \( A \) has \( n^2 \) components, while since symmetric, meaning \( c_{ij} = c_{ji} \), the matrix \( C \) only provides \( n(n+1)/2 \) constraints. Requiring \( A \) to also be symmetric reduces the unknowns to \( n(n+1)/2 \), and so the equation \( C = AA^T \) is thus theoretically solvable. But a little thought reveals that the resulting system of equations in these unknowns is not linear but second order. For example, the first equation is:

\[
\sum_{i=1}^{n} a_{1i}^2 = c_{11}.
\]

The following proposition provides an especially convenient factorization of positive definite \( C \) which makes this system of equations solvable iteratively. In particular, the above first equation reduces to \( a_{11}^2 = c_{11} \), an easy calculation since \( c_{11} \equiv \sigma_1^2 > 0 \). This result generalizes for positive semidefinite \( C \), but without the uniqueness property. We do not develop this more general result.

**Proposition 3.14 (Cholesky decomposition)** Given symmetric and positive definite matrix \( C \), there exists a unique invertible, lower triangular matrix \( L \) so that \( C = LL^T \). By lower triangular is meant that \( l_{ij} = 0 \) for \( j > i \).

**Proof.** The existence proof is by mathematical induction. Uniqueness then follows from the example 3.16 algorithm below for constructing such decompositions iteratively. The existence result is true for \( n = 1 \) where \( C \equiv (c) \), since positive definite implies \( c > 0 \) and thus \( L \equiv (\sqrt{c}) \). So assume the existence result is true for matrices that are \( (n-1) \times (n-1) \) and let \( C \) be a given \( n \times n \) symmetric and positive definite matrix. Then

\[
C = \begin{pmatrix}
C' & y \\
y^T & c_{nn}
\end{pmatrix},
\]

where \( C' \) is \( (n-1) \times (n-1) \) and symmetric, \( y \) is an \( (n-1) \)-vector and \( c_{nn} \) a constant. To be positive definite requires \( C' \) positive definite and \( c_{nn} > 0 \),
0, and this follows by considering $x^T C x$ with $x = (x_1, x_2, ..., x_{n-1}, 0)$ and $x = (0, 0, ..., 0, x_n)$, respectively. Also by assumption, $C' = L L^T$ where $L$ is $(n-1) \times (n-1)$ lower triangular.

To factor $C$, we require:

$$
\begin{pmatrix}
C' & y \\
y^T & c_{nn}
\end{pmatrix} =
\begin{pmatrix}
L & 0 \\
z^T & b
\end{pmatrix}
\begin{pmatrix}
L^T & z \\
0 & b
\end{pmatrix},
$$

where $z$ and 0 denote $(n-1)$-vectors, and $b > 0$ a constant. A calculation obtains:

$$
\begin{pmatrix}
C' & y \\
y^T & c_{nn}
\end{pmatrix} =
\begin{pmatrix}
L L^T & L z \\
(Lz)^T & \|z\|^2 + b^2
\end{pmatrix},
$$

where $\|z\|$ is the Euclidean norm in 3.12. Since $C' = L L^T$ by assumption, the proof is complete if we can solve for $z$ and $b$:

$$
L z = y, \quad \|z\|^2 + b^2 = c_{nn}.
$$

Since $L$ is invertible, $z \equiv L^{-1} y$, and the second equation for $b$ requires $c_{nn} - \|z\|^2 > 0$, and this is true by positive definiteness of $C$ as we prove next. Specifically, we prove that if $C = \begin{pmatrix} L L^T & L z \\ (Lz)^T & c_{nn} \end{pmatrix}$ is positive definite, then $c_{nn} > \|z\|^2$. To this end, consider the vector $x = \begin{pmatrix} (L^{-1} L)^{-1} z \\ 1 \end{pmatrix}$. Since $L^{-1} = (L^{-1})^T$, it follows that $C x = \begin{pmatrix} 0 \\ \|z\|^2 - c_{nn} \end{pmatrix}$ and thus $x^T C x = c_{nn} - \|z\|^2$. Since $x^T C x > 0$ for any such $x$, the result follows.

Combining:

$$
\begin{pmatrix}
C' & y \\
y^T & c_{nn}
\end{pmatrix} =
\begin{pmatrix}
L & 0 \\
(L^{-1} y)^T & c_{nn} - \|L^{-1} y\|^2
\end{pmatrix}
\begin{pmatrix}
L^T & L^{-1} y \\
0 & c_{nn} - \|L^{-1} y\|^2
\end{pmatrix}.
$$

\[\blacksquare\]

**Remark 3.15** This representation of $C$ is called the Cholesky decomposition of $C$, named for André-Louis Cholesky (1875 – 1918) who developed this approach for real matrices, the current application. This result is
in fact true more generally for complex matrices where the requirement of
being symmetric, \( c_{ij} = c_{ji} \), is replaced by being self-adjoint, \( c_{ij} = \overline{c_{ji}} \) with \( \overline{c_{ji}} \) denoting the complex conjugate of \( c_{ji} \).

**Example 3.16** Proposition 3.14 provides an existence result on a lower
diagonal matrix \( L \) such that \( C = LL^T \), so we now investigate its calculation
given a covariance matrix \( C \). Because this calculation will be shown to be
iterative, this assures the uniqueness of \( L \).

To simplify notation, let \( L_i \) denote the \( i \)th row of \( L \) so \( L_i \equiv (L_{i1}, L_{i2}, ..., L_{in}) \),
noting that since lower triangular, \( L_{ij} = 0 \) for \( j > i \). The system of \( n(n+1)/2 \)
equations, using dot product notation is:

\[
L_i \cdot L_j = c_{ij}, \quad 1 \leq j \leq i \leq n.
\]

The first equation is \( L_1 \cdot L_1 = c_{11} \), or \( l_{11}^2 = c_{11} \), and since \( c_{11} > 0 \) and \( l_{1j} = 0 \)
for \( j > 1 \) we are done defining \( L_1 \). For \( L_2 \) there are 2 equations:

\[
L_2 \cdot L_1 = c_{21}, \quad L_2 \cdot L_2 = c_{22}.
\]

The first equation has one unknown, \( l_{21} \), since \( L_1 \) has only one nonzero com-
ponent, and the second equation then has only \( l_{22} \) unknown. This determines
\( L_2 \) since \( l_{2j} = 0 \) for \( j > 2 \).

For general \( L_i \) there are \( i \) equations, \( i - 1 \) of which involve \( L_j \) for \( j < i \),
all of which are known, and the last is \( L_i \cdot L_i = c_{ii} \). Each successive equation
contains only one unknown since \( L_j \) has only \( j \) nonzero components, and
thus each is readily solvable. The one potential question is the existence of
a real solution to the last equation, \( L_i \cdot L_i = c_{ii} \), since this reduces to:

\[
l_{ii}^2 = c_{ii} - \sum_{k=1}^{i-1} l_{ik}^2,
\]

and involves a square root. That the right hand expression is positive follows
from the existence proof since this equals \( c_{nn} - \|L^{-1}y\|^2 \) in the notation of
that proof.

### 3.2 Interesting Properties

In this section we develop three interesting results on the multivariate
normal distribution. The first is the general formula for the higher central
moments of this random vector which generalizes 3.9, and which provides
an important special case known as **Isserlis’ theorem**.
3.2 INTERESTING PROPERTIES

The second result states that the component normal variates of multivariate normal $\mathbf{Y} \equiv (Y_1, Y_2, \ldots, Y_n)$ are independent if and only if they are uncorrelated. Independent implies uncorrelated, but the reverse implication is certainly not true in general for random variables. We provide an example where this reverse implication fails, even for normal random variables. An apparent paradox? See below.

The third result has applications in sampling theory. Specifically, that the sample mean and sample variance of a normal sample are independent random variables, with distribution functions identified below.

3.2.1 Higher Moments

If the random variable $Y$ has a normal distribution with parameters $\mu$ and $\sigma^2$, it was seen in 3.65 of book 4 that $E \left[ (Y - \mu)^{2m+1} \right] = 0$ for all $m$, while:

$$E \left[ (Y - \mu)^{2m} \right] = \frac{\sigma^{2m}(2m)!}{2^m m!}.$$  

Thus in particular $E[Y] = \mu$ and $E \left[ (Y - \mu)^2 \right] = \sigma^2$. In this section we generalize this formula to a multivariate normal context.

Let $\mathbf{Y} \equiv (Y_1, Y_2, \ldots, Y_n)$ be multivariate normally distributed by definition 3.6, and thus with a little rearrangement:

$$E \left[ e^{t(Y-\mu)} \right] = \exp \left[ \frac{1}{2} t^T \mathbf{C} t \right].$$

Using Taylor series and an application of Lebesgue’s dominated convergence theorem to $E \left[ e^{t(Y-\mu)} \right]$, as justified in the proof of proposition 3.17 below, we obtain:

$$\sum_{p=0}^{\infty} \frac{E \left[ (t \cdot (Y - \mu))^p \right]}{p!} = \sum_{m=0}^{\infty} \frac{(t^T \mathbf{C} t)^m}{2^m m!}. \quad (3.14)$$

The next proposition uses this identity to generalize the above results on moments.

**Proposition 3.17** Let $\mathbf{Y} \equiv (Y_1, Y_2, \ldots, Y_n)$ be multivariate normally distributed by definition 3.6 with positive semidefinite matrix $\mathbf{C} \equiv (c_{ij})_{1 \leq i,j \leq n}$ and vector $\mu \equiv (\mu_1, \mu_2, \ldots, \mu_n)$. Given nonnegative integers $(M_1, M_2, \ldots, M_n)$ define $M \equiv \sum_{j=1}^{n} M_j$. If $M = 2m + 1$ is odd, then:

$$E \left[ \prod_{j=1}^{n} (Y_j - \mu_j)^{M_j} \right] = 0. \quad (3.15)$$
For $M = 2m$:

$$E\left[\prod_{j=1}^{n} (Y_j - \mu_j)^{M_j}\right] = \frac{\prod_{j=1}^{n} M_j!}{2^m} \sum_{(m_{ij})} \frac{\prod_{i,j=1}^{n} m_{ij}!}{\prod_{i,j=1}^{n} m_{ij}!}, \quad (3.16)$$

where the summation is over all nonnegative integer $n^2$-tuples $(m_{ij})_{1 \leq i,j \leq n}$ such that:

$$\sum_{i, j=1}^{n} m_{i,j} = m,$$

and for all $j$:

$$\sum_{i=1}^{n} m_{i,j} + \sum_{i=1}^{n} m_{j,i} = M_j.$$

**Proof.** To justify (3.14), note that because $e^{t(Y - \mu)} \leq e^{t(Y - \mu)} + e^{-t(Y - \mu)}$:

$$E\left[e^{t(Y - \mu)}\right] \leq e^{-t\mu}M_Y(t) + e^{t\mu}M_Y(-t),$$

and thus $e^{t(Y - \mu)}$ is integrable by definition 3.6. Define $g_p(Y)$ by:

$$g_p(Y) = \sum_{p=0}^{P} \frac{(t \cdot (Y - \mu))^p}{p!},$$

and note that

$$E[g_p(Y)] = \sum_{p=0}^{P} \frac{E[(t \cdot (Y - \mu))^p]}{p!}.$$  

Now $g_p(Y) \to e^{t(Y - \mu)}$ pointwise, and for all $P$:

$$E[|g_p(Y)|] \leq E\left[\sum_{p=0}^{P} \frac{|(t \cdot (Y - \mu))^p|}{p!}\right] \leq E\left[e^{t(Y - \mu)}\right].$$

Thus (3.14) follows by Lebesgue’s dominated convergence theorem of proposition 2.43 of book 5:

$$E\left[e^{t(Y - \mu)}\right] = \lim_{P \to \infty} \sum_{p=0}^{P} \frac{E[(t \cdot (Y - \mu))^p]}{p!}.$$

By the multinomial theorem of 3.27 of book 4 applied to $t \cdot (Y - \mu) \equiv \sum_{j=1}^{n} t_j(Y_j - \mu_j)$:

$$E[(t \cdot (Y - \mu))^p] = \sum_{(p_1, ..., p_n)} \frac{p!}{\prod_{j=1}^{n} p_j!} E\left[\prod_{j=1}^{n} (Y_j - \mu_j)^{p_j}\right] \prod_{j=1}^{n} t_j^{p_j},$$

where the summation is over all nonnegative integer $n^2$-tuples $(m_{ij})_{1 \leq i,j \leq n}$ such that:

$$\sum_{i, j=1}^{n} m_{i,j} = m,$$
3.2 INTERESTING PROPERTIES

where the summation is over all nonnegative integer \( n \)-tuples \((p_1, ..., p_n)\) with \( \sum_{j=1}^{n} p_j = p \). Similarly applying this result to \( t^T C t \equiv \sum_{i=1}^{n} \sum_{j=1}^{n} c_{i,j} t_i t_j \), and with analogous notation:

\[
(t^T C t)^m = \sum_{(m_{ij})} \frac{m!}{\prod_{i,j=1}^{n} m_{i,j}!} \prod_{i,j=1}^{n} c_{i,j}^{m_{i,j}} \prod_{i,j=1}^{n} (t_i t_j)^{m_{i,j}}. \quad (**)
\]

Because power series expansions are unique on their domain of convergence, \(|t| < \varepsilon\) say, we can insert these expressions into 3.14 and then equate the coefficients of the various \( t \)-monomials from these expansions. From (*) we see that every \( t \)-monomial of the power series on the right in 3.14 has even degree, and thus the coefficients of all odd degree monomials for the power series on the left must be \( 0 \). From the above expression for \( E[(t \cdot (Y - \mu))^p] \) this proves 3.15.

By the same argument, for each \( m \) the coefficients of the \( t \)-monomials in the expression on the right of 3.14 must equal the coefficients of the \( t \)-monomials in the expression on the left for \( p = 2m \). This obtains that for each \( m \):

\[
\frac{1}{2m} \sum_{(m_{ij})} \frac{1}{\prod_{i,j=1}^{n} m_{i,j}!} \prod_{i,j=1}^{n} c_{i,j}^{m_{i,j}} \prod_{i,j=1}^{n} (t_i t_j)^{m_{i,j}} \quad (***)
\]

\[
= \sum_{(M_1, ..., M_n)} \prod_{j=1}^{n} M_j! E \left[ \prod_{j=1}^{n} (Y_j - \mu_j)^{M_j} \right] \prod_{j=1}^{n} t_j^{M_j},
\]

where \( \sum_{i,j=1}^{n} m_{i,j} = m \) and \( \sum_{j=1}^{n} M_j = 2m \). To compare coefficients, note that for each \( j \):

\[
\prod_{i=1}^{n} (t_i t_j)^{m_{i,j}} = \prod_{i=1}^{n} t_j^{m_{i,j}} \prod_{i=1}^{n} t_i^{m_{i,j}} = t_j^{\prod_{i=1}^{n} m_{i,j}} \prod_{i=1}^{n} t_i^{m_{i,j}},
\]

and thus:

\[
\prod_{j=1}^{n} \prod_{i=1}^{n} (t_i t_j)^{m_{i,j}} = \prod_{j=1}^{n} t_j^{\prod_{i=1}^{n} m_{i,j}} \cdot \prod_{i=1}^{n} \prod_{j=1}^{n} t_i^{m_{i,j}} = \prod_{j=1}^{n} t_j^{\prod_{i=1}^{n} m_{i,j} + \prod_{j=1}^{n} m_{j,i}}.
\]

To finish the proof, fix the \((M_1, ..., M_n)\)-term in the sum on the right in (**). By the above calculation, the coefficient of \( \prod_{j=1}^{n} t_j^{M_j} \) must equal the sum of the coefficients on the left for all \((m_{ij})\) so that \( \prod_{i=1}^{n} m_{i,j} + \prod_{j=1}^{n} m_{j,i} = M_j \). This obtains 3.16 and the proof is complete. \( \blacksquare \)

Remark 3.18 Note that for given \((M_1, ..., M_n)\), the summation in 3.16 has potentially many repetitions of the same value. For any given \( n^2 \)-tuple \((m_{ij})\),
if there exists \( m_{kl} \) so that \( m_{kl} \neq m_{lk} \), then simply switching these values obtains another term in the summation with the same value as the given \((m_{ij})\) because \( C \) is symmetric and thus \( c_{kl} = c_{lk} \). Indeed if there are \( L \) such pairs in the given \( n^2 \)-tuple \((m_{ij})\), there will be \( 2^L \) terms in the summation with this same value.

This observation is at the heart of the proof of the following result, known as Isserlis' theorem, and named for Leon Isserlis (1881–1966). It is often stated under the assumption that \( \mu = 0 \).

**Proposition 3.19 (Isserlis' theorem )** Let \( Y \equiv (Y_1, Y_2, ..., Y_n) \) be multivariate normally distributed by definition 3.6 with positive semidefinite matrix \( C \equiv (c_{ij})_{1 \leq i,j \leq n} \) and vector \( \mu \equiv (\mu_1, \mu_2, ..., \mu_n) \). Then \( E \left[ \prod_{j=1}^{n} (Y_j - \mu_j) \right] = 0 \) if \( n \) is odd, whereas if \( n = 2m \):

\[
E \left[ \prod_{j=1}^{2m} (Y_j - \mu_j) \right] = \sum'_{(i_k,j_k)_{k=1}^{m}} \prod_{k=1}^{m} c_{i_k,j_k},
\]

where the \( \sum' \)-summation is over all distinct \( m \)-tuples \((i_k,j_k)_{k=1}^{m}\) so that \( \bigcup_{k=1}^{m} (i_k,j_k) = (1,2,\ldots,2m) \).

In particular, the number of terms in this summation is:

\[
N_m = \frac{(2m)!}{2^m m!}.
\]

**Proof.** The statement for \( n \) odd is proposition 3.17. If \( n = 2m \), then \( M_j = 1 \) for all \( j \) above, and thus for all \( j \):

\[
\sum_{i=1}^{n} m_{i,j} + \sum_{i=1}^{n} m_{j,i} = 1. \quad (**)
\]

This implies that for each \( j \), \( m_{i,j} = 0 \) and there is exactly one \( i \) so that \( m_{i,j} = 1 \) or \( m_{j,i} = 1 \). Considering \((m_{ij})_{1 \leq i,j \leq n} \) as a matrix, this implies that for every \( j \), there is exactly one non-zero element in either the \( j \)th row or \( j \)th column of this matrix, and this cannot be the element on the diagonal.

This restriction assures that if \((i_k,j_k)_{k=1}^{m}\) are arbitrarily chosen to satisfy these \( n \) constraints, that \( \bigcup_{k=1}^{m} (i_k,j_k) = (1,2,\ldots,2m) \). To see this note that if this is not the case then there is at least one index \( j \) so that both \((i_k,j)\), and either \((j,j_k)\) or \((i,j)\), are in this collection. But then in either case:

\[
\sum_{i=1}^{n} m_{i,j} + \sum_{i=1}^{n} m_{j,i} = 2,
\]

contradicting (**).
3.2 INTERESTING PROPERTIES

It now follows from 3.16 that

$$E \left[ \prod_{j=1}^{n} (Y_j - \mu_j) \right] = \frac{1}{2^{m}} \sum_{(i_k,j_k)_{k=1}^{m}} \prod_{i,j=1}^{n} c_{i,j}^{m_{i,j}},$$

where the summation is over all such index collections $(i_k,j_k)_{k=1}^{m}$ for which the associated $(m_{i,j})$ satisfy $(\ast)$. Now recall remark 3.18. For any such given collection, there are exactly $2^m$ collections which obtain the same value of $\prod_{i,j=1}^{n} c_{i,j}^{m_{i,j}}$. These are defined as the collections $(i'_k,j'_k)_{k=1}^{m}$ for which either $(i'_k,j'_k) = (i_k,j_k)$ or $(i'_k,j'_k) = (j_k,i_k)$ for all $k$. Identifying all such collections defined by such subsets one obtains the count $2^m$, and thus 3.17 follows by restricting the summation to distinct collections of index pairs.

For 3.18, note that there are $(2^{m})!$ orderings of the $(1,2,...,2m)$ which by sequentially grouping into pairs can be initially identified with $(2^{m})!$ collections of distinct indexes $(i_k,j_k)_{k=1}^{m}$, where by construction, $\bigcup_{k=1}^{m} (i_k,j_k) = (1,2,...,2m)$. The collections are not all distinct of course, since the ordering of the $m$ pairs, and the index order within the $m$ pairs, does not affect the given collection. For any given collection identified, there will be and additional $2^{m} - 1$ collections with the $m$ pairs in the same order, but with from 1 to $m$ pairs of indexes in reverse order. Then given any such collection, there will be $m!$ collections with the same pairs in some order.

Remark 3.20 The number of terms in the summation grows very fast with $m$: $N_1 = 1$, $N_2 = 3$, $N_3 = 15$, $N_4 = 105$, and in general $N_{m+1} = (2m + 1)N_m$. Using Stirling’s approximation in 3.91 of book 4, as $m \to \infty$:

$$N_m \sim 2^{m+1/2} \left( \frac{m}{e} \right)^m.$$

Example 3.21 With $m = 2$:

$$E \left[ \prod_{j=1}^{4} (Y_j - \mu_j) \right] = c_{12}c_{34} + c_{13}c_{24} + c_{14}c_{23}.$$

3.2.2 Independent vs. Uncorrelated

A corollary of proposition 3.1 is that for the multivariate normal distribution, the component normal variates are independent if and only if they are uncorrelated. Recall from book 4 that the correlation between variates $Y_i, Y_j$ is defined:

$$\text{Corr}[Y_i, Y_j] \equiv \frac{\text{Cov}[Y_i, Y_j]}{\sigma_i \sigma_j},$$

(3.19)
with $\text{Cov}[Y_i, Y_j]$ defined in 3.10. This correlation is often denoted $\rho_{ij}$. By corollary 3.50 of book 4:

$$-1 \leq \rho_{ij} \leq 1,$$

and *uncorrelated* means that $\rho_{ij} = 0$, or equivalently (as long as $\sigma_i$ and $\sigma_j$ exist), $\text{Cov}[Y_i, Y_j] = 0$.

In general, independence is a much stronger condition than is the property of being uncorrelated, in that independence assures that $\text{Cov}[Y_i, Y_j] = 0$ by 1.36. That the opposite implication fails, indeed even for general normal variates, see example 3.23 that follows.

**Corollary 3.22** Let $Y = (Y_1, Y_2, \ldots, Y_n)$ be multivariate normally distributed as in 3.3 with $C$ positive definite. Then $\{Y_i\}_{i=1}^n$ are independent if and only if $c_{ij} = 0$ for $i \neq j$. Thus by 3.9, component normal variates of a multivariate normal distribution are independent if and only if they are uncorrelated.

**Proof.** If $c_{ij} = 0$ for $i \neq j$ then $C$ is diagonal and hence $f_Y(y) = \prod_{i=1}^n f_{Y_i}(y_i)$ where $f_{Y_i}(y_i)$ is the normal density with mean $\mu_i$ and variance $c_{ii}$. Thus $F_Y(y) = \prod_{i=1}^n F_{Y_i}(y_i)$ and this assures independence by proposition 3.53 of book 2.

Conversely, if $\{Y_i\}_{i=1}^n$ are independent then $f_Y(y) = \prod_{i=1}^n f_{Y_i}(y_i)$ and so $c_{ij} = \text{Cov}[Y_i, Y_j] = 0$ for $i \neq j$ and thus $C$ is diagonal. $\blacksquare$

**Example 3.23** It is important to note that in general uncorrelated normals need not be independent. The corollary above specifies only that this must be the case when these normals are the components of a multivariate normally distributed random vector.

As a general example, let $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$ and define:

$$g(c) = \int_{-c}^c x^2 \varphi(x)dx.$$  

Then $g(0) = 0$, $g$ is continuous and strictly increasing, and $g(c) \to 1$ as $c \to \infty$ since $g(\infty)$ is the variance of the standard normal. Thus by the intermediate value theorem there is a $c'$ with $g(c') = 1/2$. Numerically one obtains $c' \approx 1.5382$.

Now let $X$ have a standard normal distribution, and define $Y = X$ if $|X| \leq c'$, and $Y = -X$ otherwise. Then by symmetry $Y$ has a standard normal distribution, the details of which are left as an exercise. Also:

$$E[XY] = \int_{|x| \leq c'} x^2 \varphi(x)dx - \int_{|x| > c'} x^2 \varphi(x)dx = 0.$$
3.2 INTERESTING PROPERTIES

Hence \(X\) and \(Y\) are uncorrelated normals. But \(X\) and \(Y\) are not independent, since for example
\[
\Pr(|X| \leq c', |Y| \leq c'] = \Pr(|X| \leq c') \neq \Pr(|X| \leq c' | \Pr(|Y| \leq c'].
\]

3.2.3 Mean and Variance of a Random Normal Sample

The notion of a random sample was formalized in definition 4.1 of book 2, which we recall here:

**Definition 3.24** Given a probability space \((\mathcal{S}, \mathcal{E}, \mu)\) and a random variable \(X : \mathcal{S} \rightarrow \mathbb{R}\), a collection of random variables \(\{X_j\}_{j=1}^N\) defined on a probability space \((\mathcal{S}', \mathcal{E}', \mu')\), where \(N\) can be finite or infinite, is said to be a sample of \(X\) or a random sample of \(X\), if this collection is independent and identically distributed (i.i.d.):

1. \(\{X_j\}_{j=1}^N\) are independent if given \((i_1, \ldots, i_m) \subset (1, 2, \ldots, N)\) and \(\{A_j\}_{j=1}^m \subset \mathcal{B}(\mathbb{R})\), the Borel sigma algebra:
\[
\mu'[\bigcap_{j=1}^m X_{i_j}^{-1}(A_j)] = \prod_{j=1}^m \mu'[X_{i_j}^{-1}(A_j)],
\]  
(3.20)

2. \(X_j\) and \(X\) are identically distributed if given \(A \in \mathcal{B}(\mathbb{R})\), then for any \(j\):
\[
\mu'[X_j^{-1}(A)] = \mu[X^{-1}(A)].
\]  
(3.21)

The space \((\mathcal{S}', \mathcal{E}', \mu')\) is called a sample space.

As noted in definition 4.3 (Alternate), conditions 1 and 2 can be restated.

If \(F_j\) denotes the distribution function of \(X_j\) and \(F\) the distribution function of \(X\):

**Definition 3.25** 1. \(\{X_j\}_{j=1}^N\) are independent if given \((i_1, \ldots, i_m) \subset (1, 2, \ldots, N)\) and \(F(x_{i_1}, \ldots, x_{i_m})\) the joint distribution function of \(\{X_{i_j}\}_{j=1}^m\), then:
\[
F(x_{i_1}, \ldots, x_{i_m}) = \prod_{j=1}^m F_{i_j}(x_{i_j}).
\]  
(3.22)

2. \(X_j\) and \(X\) are identically distributed if for any \(j\):
\[
F_j(x) = F(x),
\]  
(3.23)
CHAPTER 3 MULTIVARIATE NORMAL DISTRIBUTION

Given a probability space \((\mathcal{S}, \mathcal{E}, \mu)\) and a random variable \(X : \mathcal{S} \rightarrow \mathbb{R}\), proposition 4.4 of book 2 proves the existence of \(\{X_j\}_{j=1}^N\) and \((\mathcal{S}', \mathcal{E}', \mu')\) for any \(N\), finite or infinite. The rest of that chapter then focused on the empirical generation of such samples, a discussion extended in example 3.13 above.

Returning to the application at hand, let \(X\) be a normal variate with mean \(\mu\) and variance \(\sigma^2\), and \(\{X_j\}_{j=1}^N\) a random sample of such variates.

**Definition 3.26** A statistic is the numerical value obtained when a Borel measurable function on \(\mathbb{R}^N\) is evaluated on a random sample \(\{X_j\}_{j=1}^N\). Put another way, a statistic is a random variable defined on the sample space \((\mathcal{S}', \mathcal{E}', \mu')\).

Given a finite random sample \(\{X_j\}_{j=1}^n\), statistics of interest are:

**Definition 3.27** 1. The **sample mean**, denoted \(m\) (for \(\mu\)) or \(\bar{X}\):

\[
m \equiv \frac{1}{n} \sum_{j=1}^n X_j.
\]

(3.24)

2. The **sample variance**, denoted \(s^2\) (for \(\sigma^2\)):

\[
 s^2 \equiv \frac{1}{n} \sum_{j=1}^n (X_j - m)^2.
\]

(3.25)

3. The **unbiased sample variance**, denoted (here) \(\hat{s}^2\):

\[
 \hat{s}^2 \equiv \frac{1}{n - 1} \sum_{j=1}^n (X_j - m)^2.
\]

(3.26)

**Note:** There does not seem to be a standard notational convention for the statistic in 3.

**Remark 3.28** The above definition may initially surprise a reader who, based on elementary probability theory, considered "statistics" to be simply the numerical results of calculations applied to samples. While this is certainly the case, what makes the calculations of these numbers meaningful is that one is often able to make probabilistic statements about the extent to which these values reflect the true parameters that these calculations seek to estimate.
3.2 INTERESTING PROPERTIES

For example, if \( \{X_j\}_{j=1}^N \) is a sample from an unknown distribution which is assume to have a mean \( \mu \), then using properties of expectations (see section 3.2.3 of book 4 for example):

\[
E[m] = \frac{1}{n} \sum_{j=1}^{n} E[X_j] = \mu,
\]

recalling that by the i.i.d. property of definition 3.24, \( E[X_j] = \mu \) for all \( j \).

The statistic \( m \) is then said to be an **unbiased estimator for** \( \mu \), meaning that its expected value equals the theoretical value of the parameter that it is intended to estimate. Other probabilistic results for \( m \) can be found in the **laws of large numbers of books 2 and 4** and the **central limit theorems** of books 4 and this book.

For 2 and 3, note that if \( \{X_j\}_{j=1}^N \) is a sample from an unknown distribution which is assume to have a mean \( \mu \) and variance \( \sigma^2 \):

\[
(X_j - m)^2 = \left( X_j - \frac{1}{n} \sum_{k=1}^{n} X_k \right)^2
= \left( \frac{n-1}{n} X_j - \frac{1}{n} \sum_{k\neq j} X_k \right)^2
= \left( \frac{n-1}{n} \right)^2 X_j^2 - \frac{2(n-1)}{n^2} \sum_{k\neq j} X_j X_k
+ \frac{1}{n^2} \sum_{k\neq j} X_k^2 + \frac{1}{n^2} \sum_{k,l:k\neq l} X_l X_k.
\]

Taking expectations, \( E[X_i^2] = \sigma^2 + \mu^2 \) for all \( i \), and by independence \( E[X_i X_k] = \mu^2 \) if \( k \neq i \). Thus, noting that the summations have \( n-1, n-1, \) and \( (n-1)(n-2) \) terms, respectively, obtains:

\[
E[(X_j - m)^2] = \left[ \left( \frac{n-1}{n} \right)^2 + \frac{n-1}{n^2} \right] (\sigma^2 + \mu^2)
+ \left[ \frac{(n-1)(n-2)}{n^2} - \frac{2(n-1)(n-1)}{n^2} \right] \mu^2
= \left( \frac{n-1}{n} \right) \sigma^2.
\]

Hence:

\[
E[s^2] = \left( \frac{n-1}{n} \right) \sigma^2,
\]

making the statistic \( s^2 \) a **biased estimator for** \( \sigma^2 \), while for \( \tilde{s}^2 \):

\[
E[\tilde{s}^2] = \sigma^2
\]
and thus as the name implies, this statistic is an unbiased estimator for \( \sigma^2 \).

The next result provides an interesting conclusion that for a normal sample, \( m \) and \( s^2 \) are independent random variables.

**Proposition 3.29** If \( \{X_j\}_{j=1}^n \) is a random sample from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), then \( m \) and \( s^2 \) are independent random variables.

**Proof.** Note that

\[
\sum_{k=1}^n (X_k - m) = 0
\]

implies that

\[
s^2 = \frac{1}{n} \left[ \sum_{j=2}^n (X_j - m)^2 + \left( \sum_{j=2}^n (X_j - m) \right)^2 \right].
\]

If we prove that \( m \) is independent of \((X_2 - m, \ldots, X_n - m)\), and then \( m \) will be independent from \( s^2 \) by proposition 3.56 of book 2.

To this end, define a random vector \( Y \) on sample space \((S', \mathcal{E}', \mu')\) by:

\[
Y : S' \to (m, X_2 - m, \ldots, X_n - m),
\]

where \( m = \frac{1}{n} \sum_{j=1}^n X_j \) as in 3.24. Note that \( Y = BX \) with \( X \equiv (X_1, X_2, \ldots, X_n) \) and \( B \) an \( n \times n \) matrix given by:

\[
B = \begin{pmatrix}
\frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\
-\frac{1}{n} & 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\
-\frac{1}{n} & -\frac{1}{n} & 1 - \frac{1}{n} & \cdots & -\frac{1}{n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-\frac{1}{n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & 1 - \frac{1}{n}
\end{pmatrix}.
\]

By proposition 3.9 \( Y \) is multivariate normally distributed with \( \mu_B = B\mu \) and \( C_B = BCB^T \). Since \( \mu \) and \( C \) are the mean vector and covariance matrix of \( X \), by the i.i.d. property of definition 3.24, \( \mu = (\mu, \ldots, \mu) \) and \( C = \sigma^2 I \) with \( I \) the identity matrix. Hence \( \mu_B = (\mu, 0, \ldots, 0) \) and \( C_B = \sigma^2 BB^T \). Finally, adding the first row of \( B \) to all the others does not change \( \det B \), and one obtains that this matrix has determinant \( 1/n \) and is thus invertible.
**3.2 INTERESTING PROPERTIES**

By 3.3:

\[ f_Y(y) = (2\pi)^{-n/2} |\det C_B|^{-1/2} \exp \left[ -\frac{1}{2} (y - \mu_B)^T C_B^{-1} (y - \mu_B) \right]. \]

Our goal is to prove that the exponent in the exponential function is a sum of \(y_i^2\) and an expression involving \((y_2, ..., y_n)\) since this then assures that \(f_Y(y) = f(y_1) f(y_2, ..., y_n)\) and the independence proof will be complete. To this end, recalling that \((AB)^{-1} = B^{-1} A^{-1}\) and \((A^T)^{-1} = (A^{-1})^T\):

\[
(y - \mu_B)^T C_B^{-1} (y - \mu_B) = \sigma^2 (y - \mu_B)^T (B^{-1})^T B^{-1} (y - \mu_B) \\
= \sigma^2 [B^{-1} (y - \mu_B)]^T B^{-1} (y - \mu_B) \\
= \sigma^2 |B^{-1} (y - \mu_B)|^2,
\]

recalling that for a vector \(x\), \(x^T x = |x|^2 \equiv \sum_{j=1}^n x_j^2\), the squared length of \(x\).

The final step is to evaluate this squared length, for which \(B^{-1}\) is needed. Using row reduction one obtains (or at least confirms):

\[
B^{-1} = \begin{pmatrix}
1 & -1 & -1 & -1 & \cdots & -1 \\
1 & 1 & 0 & 0 & \cdots & 0 \\
1 & 0 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & 0 & \cdots & 1
\end{pmatrix}.
\]

Letting \(z \equiv y - \mu_B\) for notational convenience:

\[
B^{-1} z = (z_1 - \sum_{j=2}^n z_j, z_1 + z_2, ..., z_1 + z_n),
\]

so

\[
|B^{-1} z|^2 = \left(z_1 - \sum_{j=2}^n z_j\right)^2 + \sum_{j=2}^n (z_1 + z_j)^2 \\
= n z_1^2 + \left(\sum_{j=2}^n z_j\right)^2 + \sum_{j=2}^n z_j^2 \\
= n (y_1 - \mu)^2 + \left(\sum_{j=2}^n y_j\right)^2 + \sum_{j=2}^n y_j^2.
\]

Thus \(f_Y(y) = f(y_1) f(y_2, ..., y_n)\) and the proof is complete. ■

The final detail is for the distributions of \(m\) and \(s^2\).
Proposition 3.30 If \( \{X_j\}_{j=1}^n \) is a random sample from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), then:

1. \( m \) has a **normal distribution** with mean \( \mu \) and variance \( \sigma^2 / n \).

2. \( ns^2 / \sigma^2 = (n - 1)s^2 / \sigma^2 \) has a **chi-squared distribution** with \( n - 1 \) degrees of freedom.

**Proof.** The distributional result in 1 is assigned as an exercise. For 2, example 1.14 and notation 1.15 of book 4 obtain that \( (X_j - \mu)^2 / \sigma^2 \) has a gamma distribution defined in 1.22 of that book with parameters \( \lambda = \alpha = 1/2 \), which is by definition the chi-squared distribution with 1 degree of freedom. Thus a summation of \( n \)-independent such gammas is by example 3.59 of book 4 a gamma with \( \lambda = 1/2 \) and \( \alpha = n/2 \), which by definition is the chi-squared distribution with \( n \) degrees of freedom, denoted \( \chi^2_{n, \text{d.f.}} \). The discrepancy of 1 degree of freedom will be accommodated by switching from \( (X_j - \mu)^2 / \sigma^2 \) to \( (X_j - m)^2 / \sigma^2 \) as follows.

First:

\[
\sum_{k=1}^n \left( \frac{X_k - \mu}{\sigma} \right)^2 = \sum_{k=1}^n \left( \frac{X_k - m}{\sigma} + \frac{m - \mu}{\sigma} \right)^2 = \sum_{k=1}^n \left( \frac{X_k - m}{\sigma} \right)^2 + n \left( \frac{m - \mu}{\sigma} \right)^2 ,
\]

by (*) of the preceding proof. Thus:

\[
\sum_{k=1}^n \left( \frac{X_k - \mu}{\sigma} \right)^2 = \frac{ns^2}{\sigma^2} + \left( \frac{m - \mu}{\sigma / \sqrt{n}} \right)^2 .
\]

From the above discussion the term on the left is \( \chi^2_{n, \text{d.f.}} \); the second term on the right is \( \chi^2_{1, \text{d.f.}} \) by 1, since \( \frac{m - \mu}{\sigma / \sqrt{n}} \) is standard normal. By independence of \( m \) and \( s^2 \), this obtains independence of the terms on the right by proposition 3.56 of book 2. Evaluating the moment generating function of both sides obtains by independence and 3.59 of book 4 that for \( t < \lambda = 1/2 \):

\[
(1 - 2t)^{-n/2} = M_{ns^2 / \sigma^2}(t)(1 - 2t)^{-1/2}.
\]

Thus \( M_{ns^2 / \sigma^2}(t) = (1 - 2t)^{-(n-1)/2} \) and \( ns^2 / \sigma^2 \) is gamma with \( \lambda = 1/2 \) and \( \alpha = (n - 1)/2 \) by corollary 3.58 of book 4. Again by definition, \( ns^2 / \sigma^2 \) is \( \chi^2_{n-1, \text{d.f.}} \).  ■
Exercise 3.31 If \( \{X_j\}_{j=1}^n \) is a random sample from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \), use proposition 3.9 to show that \( m \) has a normal distribution with mean \( \mu \) and variance \( \sigma^2/n \). Hint: \( B = (1/n, 1/n, \ldots, 1/n) \).

3.3 Existence of Multivariate Normal Densities

Generalizing the result of proposition 3.1, let \( X \equiv (X_1, X_2, \ldots, X_n) \) be \( n \) independent unit normals with density function as in 3.2, and let \( Y \equiv (Y_1, Y_2, \ldots, Y_m) \) be defined by \( Y = AX + \mu \) where \( A \) is an \( m \times n \) matrix, so \( A : \mathbb{R}^n \to \mathbb{R}^m \), and \( \mu \equiv (\mu_1, \mu_2, \ldots, \mu_m) \) a constant vector. If \( n = m \) and \( A \) is invertible then this is the set-up of proposition 3.1. The goal of this section is to investigate cases with \( n \neq m \), or, \( n = m \) and \( A \) is singular, meaning not invertible.

In all of these cases proposition 3.9 still applies, using \( A \) here as \( B \) in that result. Since \( \mu_X = 0 \) the zero vector, and \( C_X = I \) the \( n \times n \) identity matrix, it follows from proposition 3.9 that \( Y \) has a multivariate normal distribution with parameters given by:

\[
\mu_Y = \mu, \quad C_Y = AA^T.
\]

Thus \( C_Y \) is an \( m \times m \) matrix.

The moment generating function of \( Y \) is defined on \( t \equiv (t_1, \ldots, t_m) \) and given by 3.8:

\[
M_Y(t) = \exp \left[ \mu \cdot t + \frac{1}{2} \sigma^2 t^T (AA^T) t \right].
\]

Comparing this result to 3.7 it follows that in the special case where \( C \equiv \sigma^2 AA^T \) is an invertible \( m \times m \) matrix, the multivariate normal distribution of \( Y \) has an associated density function defined in 3.3. This follows by the uniqueness theorem discussed in remark 3.5. In general however, \( C \) need not be invertible and thus \( Y \) will have no density function.

Some examples will motivate the general results.

Example 3.32 1. \( n < m \): Let \( A : \mathbb{R}^2 \to \mathbb{R}^3 \) be defined by \( A : (X_1, X_2) \to (X_1, X_2, X_1 + X_2) \). Then

\[
A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad AA^T = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{pmatrix}.
\]
and \( \det(AA^T) = 0 \).

2. \( n > m \): Let \( A : \mathbb{R}^3 \to \mathbb{R}^2 \) be defined by \( A : (X_1, X_2, X_3) \to (X_1 + X_2, X_2 + X_3) \). Then

\[
A = \begin{pmatrix}
1 & 1 & 0 \\
0 & 1 & 1
\end{pmatrix}, \quad AA^T = \begin{pmatrix}
2 & 1 \\
1 & 2
\end{pmatrix},
\]

and \( AA^T \) is invertible.

On the other hand, defining \( A : (X_1, X_2, X_3) \to (X_1 + X_2 + X_3, X_1 + X_2 + X_3) \) leads to the conclusion that \( \det(AA^T) = 0 \).

3. \( n = m \): Let \( A : \mathbb{R}^3 \to \mathbb{R}^3 \) be defined by \( A : (X_1, X_2, X_3) \to (X_1, X_2, X_1 + X_2) \). Then

\[
A = \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
1 & 1 & 0
\end{pmatrix}, \quad AA^T = \begin{pmatrix}
1 & 0 & 1 \\
0 & 1 & 1 \\
1 & 1 & 2
\end{pmatrix},
\]

and \( \det(AA^T) = 0 \).

These examples illustrate some general results from linear algebra for a matrix \( A : \mathbb{R}^n \to \mathbb{R}^m \) where the elements of \( A \) are real. We summarize the needed results below. For more detail on these and related results the reader is referred to Strang (2009).

1. Define **column rank** \( \text{column rank}(A) \) as the dimension of the subspace of \( \mathbb{R}^m \) spanned by the \( n \) column vectors of \( A \), and analogously, define **row rank** \( \text{row rank}(A) \) as the dimension of the subspace of \( \mathbb{R}^n \) spanned by the \( m \) row vectors of \( A \). It is a remarkable result that:

\[
\text{column rank}(A) = \text{row rank}(A), \tag{3.27}
\]

a result known as the **rank theorem**.

2. **Rank** \( (A) \) is defined as the dimension of the range of \( A \). However the range of \( A \) is spanned by the columns of this matrix since \( Ax = \sum_{j=1}^n A^{(j)}x_j \) where \( A^{(j)} \) denotes the \( j \)th column of \( A \). Hence:

\[
\text{rank}(A) = \text{column rank}(A).
\]
Similarly, since the columns of $A^T$ are the rows of $A$,
\[ \text{rank}(A^T) = \text{row rank}(A), \]
and thus by the rank theorem:
\[ \text{rank}(A) = \text{rank}(A^T). \tag{3.28} \]
A consequence of this is if $A$ is $n \times n$, then $A$ is invertible if and only if $A^T$ is invertible. In this case, $(A^T)^{-1} = (A^{-1})^T$.

3. As a subspace of $\mathbb{R}^m$, column rank $(A) \leq m$, and similarly row rank $(A) \leq n$, and thus:
\[ \text{rank}(A) \leq \min(n, m). \tag{3.29} \]

4. The final result is that:
\[ \text{rank}(A) = \text{rank}(AA^T) = \text{rank}(A^TA). \tag{3.30} \]

**Proposition 3.33** Let $X \equiv (X_1, X_2, \ldots, X_n)$ be $n$ independent unit normals with density function as in 3.2, and let $Y \equiv (Y_1, Y_2, \ldots, Y_m)$ be defined by $Y = AX + \mu$ where $A$ is an $m \times n$ matrix, so $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$, and $\mu \equiv (\mu_1, \mu_2, \ldots, \mu_m)$ a constant vector. Then $Y$ has a multivariate normal density function if and only if $AA^T$ is invertible, and in this case the density function of $Y$ is given in 3.3 with parameters $\mu$ and $C = AA^T$.

More generally, if $Y \equiv (Y_1, Y_2, \ldots, Y_m)$ has a multivariate normal distribution by definition 3.6 with parameters $\mu$ and $C$ and $B : \mathbb{R}^m \rightarrow \mathbb{R}^p$ is a given $p \times m$ matrix, then $W \equiv BY$ has a multivariate normal density function if and only if $BCB^T$ is invertible. In this case the density function of $W$ is given in 3.3 with parameters $\mu_W = B\mu$ and $C_W = BCB^T$.

**Proof.** We prove the first statement and leave the details of the second as an exercise. If $Y$ has a multivariate normal density function given in 3.3 with parameters $\mu_Y$ and $C_Y$, where $C_Y$ is invertible, then the moment generating function is given in 3.7 with these parameters. On the other hand by proposition 3.9,
\[ M_{AX+\mu}(t) = \exp \left[ \mu \cdot t + \frac{1}{2} t^T AA^T t \right]. \]

Thus by uniqueness (remark 3.5), $\mu_Y = \mu$, $C_Y = AA^T$, and $AA^T$ is invertible.
Conversely, if $A A^T$ is invertible and $Y = AX + \mu$, then by exercise 3.4 and proposition 3.9:

$$M_{AX+\mu}(t) = \exp[\mu \cdot t] M_{AX}(t)$$

$$\quad = \exp \left[ \mu \cdot t + \frac{1}{2} t^T A A^T t \right].$$

Now define a multivariate normal density function $f(y)$ as in 3.3 with parameters $\mu$ and $C \equiv AA^T$. Then the associated moment generating function is given in 3.7 of exercise 3.4. By uniqueness, $f(y)$ is the density function of $Y$. \qed
Chapter 4

Weak Convergence of Measures

With the aid of the integration theory of book 5, we now continue our investigation into the weak convergence of probability measures, first extending the book 2 results for probability measures on $\mathbb{R}$, then generalizing results to such measures on $\mathbb{R}^n$. As stated in the book 2 definition 8.2 and confirmed in exercise 8.4, weak convergence of probability measures can be equivalently defined in terms of weak convergence of the associated distribution functions. The integrals in the propositions below can then be expressed in terms of probability measures or the associated Lebesgue-Stieltjes integrals, and the latter approach is the version often applied in probability theory.

4.1 Measures on $\mathbb{R}$

4.1.1 Portmanteau Theorem on $\mathbb{R}$

In this section we present two important implications of the weak convergence of a sequence of measures on $\mathbb{R}$. Recall that in definition 8.2 of book 2, a sequence of probability measures $\{\mu_n\}$ on $\mathbb{R}$ is said to converge weakly to a probability measure $\mu$, denoted $\mu_n \Rightarrow \mu$, if $\mu_n((-\infty, x])$ converges to $\mu((-\infty, x])$ for all $x$ for which $\mu\{x\} = 0$. It was then seen in corollary 8.8 that this generalized to the conclusion that $\mu_n(I) \to \mu(I)$ for any interval $I = (a, b)$, whether open, closed or semi-closed, such that $\mu\{(a, b)\} = 0$, where $\{a, b\}$ denotes the set of two points that defines the "boundary" of the interval $I$. By example 8.10 of that book, however, it
was shown that weak convergence did not imply that $\mu_n(A) \to \mu(A)$ for all Borel measurable sets. In that example, $A$ was the set of rational numbers in $[0,1]$, and $\mu = m$, Lebesgue measure.

In the proposition below it will be seen that the implication for convergence of set measures, $\mu_n(A) \to \mu(A)$, can be expanded to any measurable set $A$ if the "boundary" of $A$ has $\mu$-measure 0. Such a measurable set $A$ is sometimes called a $\mu$-continuity set. This proposition will also document an important result in property 2 for the convergence of integrals defined by these measures. In fact, it is not uncommon to see weak convergence of measures defined in terms of property 2 of this proposition. Since this is an equivalent definition, in the end it matters little which is introduced as the definition. The approach taken here and elsewhere has the advantage that the basic properties of weak convergence can be introduced and applied before an integration theory is developed.

Remark 4.1 The result below is often referred to as the Portmanteau theorem, or the portmanteau theorem since it is not named for a mathematician and thus is often not capitalized. It was first proved in 1940 by Aleksandr Aleksandrov (1912 – 1999), and in some references is identified with this name. The portmanteau label derives from an altogether different reference.

The word "portmanteau" is used in linguistics, sometimes qualified as a "portmanteau word," to describe words that are formed by packing two other words together. Examples include "smog," which derives from smoke and fog, as well as "guesstimate" from guess and estimate, though there are literally hundreds of examples. This use was coined by Lewis Carroll, the pen name of Charles Lutwidge Dodgson (1832 – 1898), in Through the Looking-Glass.

Returning to Aleksandrov’s theorem, it is commonly referred to by the name of portmanteau because it combines together many seemingly different ways to think about weak convergence of measures, and proves them to be equivalent.

We begin with a definition.

Definition 4.2 Given a set $A$, the boundary of $A$, denoted $\partial A$, is the set of points that are simultaneously the limit of a sequence of points in $A$ and a sequence of points in $\overline{A} = A^c$, the complement of $A$. Equivalently,

$$\partial A = \overline{A} \cap \overline{A}^c,$$  \hspace{1cm} (4.1)
where \( \overline{A} \), respectively \( \overline{A^c} \), is defined as the collection of limit points of \( A \), respectively, \( A^c \). See also 4.2.

Given a measure \( \mu \), a measurable set \( A \) is said to be a \( \mu \)-continuity set if \( \mu(\partial A) = 0 \).

**Example 4.3** For any bounded interval, \( I = (a, b) \), whether open, closed or semi-closed, it is apparent that \( \partial A = \{a, b\} \) since one can construct sequences \( \{a + 1/n\} \) and \( \{b + 1/n\} \) as required by this definition. Alternatively, \( \overline{A} = [a, b] \), \( \overline{A^c} = (-\infty, a] \cup [b, \infty) \) and so \( \partial A = \{a, b\} \). Hence if \( \mu(\{a, b\}) = 0 \) then \( I \) is a \( \mu \)-continuity set.

If \( A = [0, 1] \cap \mathbb{Q} \), the rationals in \( [0, 1] \), then \( \partial A = [0, 1] \) since any point in this interval is the limit point of rational numbers or irrational numbers. Alternatively, \( \overline{A} = [0, 1] = \overline{A^c} \). If \( \mu \equiv m \), Lebesgue measure, then \( A \) is not an \( m \)-continuity set.

We now state a simple version of the portmanteau theorem. There are in fact other characterizations beyond those identified below. See the portmanteau theorem for \( \mathbb{R}^j \) for details.

**Proposition 4.4 (portmanteau theorem on \( \mathbb{R} \))** Let \( \{\mu_n\} \), \( \mu \) be probability measures on \( \mathbb{R} \). Then the following are equivalent:

1. \( \mu_n \Rightarrow \mu \).
2. \( \int g(x)d\mu_n \to \int g(x)d\mu \) for every bounded, continuous real-valued function \( g \).
3. \( \mu_n(A) \to \mu(A) \) for every \( \mu \)-continuity set \( A \).

**Proof.**

\( a. 1 \Rightarrow 2 : \) Assume \( \mu_n \Rightarrow \mu \) and let \( \{X_n, X\} \) be the random variables defined in Skorokhod’s theorem (proposition 8.30, book 2) with probability measures \( \{\mu_n, \mu\} \). Specifically, \( X_n : ((0, 1), \mathcal{B}((0, 1)), m) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_n) \), \( X : ((0, 1), \mathcal{B}((0, 1)), m) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu) \) and \( X_n(s) \to X(s) \) for all \( s \). Now by the Skorokhod construction, \( X_n \) has distribution function \( F_n \), the distribution function induced by the probability measure \( \mu_n \). Thus \( \mu_n \) is the measure on \( \mathcal{B}(\mathbb{R}) \) induced by \( X_n \) in the sense of definition 3.9 of book 5, meaning for all \( A \in \mathcal{B}(\mathbb{R}) \):

\[
\mu_n(A) = m(X_n^{-1}(A)).
\]
The same conclusion follows that $\mu$ is the measure on $B(\mathbb{R})$ induced by $X$. By the change of variable formula in proposition 3.14 (book 5):

$$\int g(x)d\mu_n = \int_{(0,1)} g(X_n(s))dm,$$

$$\int g(x)d\mu = \int_{(0,1)} g(X(s))dm.$$

Now $\mu(D_g) = 0$ where $D_g$ is the set of discontinuities of $g$, and we have from the mapping theorem (proposition 8.37, book 2) that $g(X_n) \to g(X)$ m-a.e. Then by the bounded convergence theorem of proposition 2.37 (book 3), since $g$ is bounded, $\int_{(0,1)} g(X_n(s))dm \to \int_{(0,1)} g(X(s))dm$ and the result follows.

b. $2 \implies 1$: Given $y$ and $\epsilon > 0$, define:

$$g^+(x) = \begin{cases} 1, & x \leq y, \\ (y + \epsilon - x)/\epsilon, & y \leq x \leq y + \epsilon, \\ 0, & x \geq y + \epsilon. \end{cases}$$

Then $F_n(y) \leq \int g^+(x)d\mu_n$ and $\int g^+(x)d\mu \leq F(y + \epsilon)$, where $F, F_n$ are the distribution functions associated with $\mu, \mu_n$. By property 2 $\int g^+(x)d\mu_n \to \int g^+(x)d\mu$ and thus $\limsup_n F_n(y) \leq F(y + \epsilon)$ for all $\epsilon > 0$, and then by right continuity of $F$, $\limsup_n F_n(y) \leq F(y)$.

Similarly, define

$$g^-(x) = \begin{cases} 1, & x \leq y - \epsilon, \\ (y - x)/\epsilon, & y - \epsilon \leq x \leq y, \\ 0, & x \geq y. \end{cases}$$

Then $F(y - \epsilon) \leq \int g^-(x)d\mu$ and $\int g^-(x)d\mu_n \leq F_n(y)$, so $F(y - \epsilon) \leq \liminf_n F_n(y)$ for all $\epsilon > 0$, which implies that $F(y^-) \leq \liminf_n F_n(y)$.

Combining:

$$F(y^-) \leq \liminf_n F_n(y) \leq \limsup_n F_n(y) \leq F(y).$$

If $y$ is a continuity point of $F$, then $F(y^-) = F(y)$ and thus $F_n(y) \to F(y)$. By exercise 8.4 of book 2, this is equivalent to $\mu_n \Rightarrow \mu$. 
4.1 MEASURES ON $\mathbb{R}$

c. $1 \Rightarrow 3$: Let $g = \chi_A$, then $D_g$ the set of discontinuities of $g$, satisfies $D_g = \partial A$, and so $\mu(\partial A) = 0$ implies that $\mu(D_g) = 0$. Thus we have as in part a that $g(X_n) \to g(X)$ m-a.e., and by the change of variable formula and the bounded convergence theorem conclude that

$$
\mu_n(A) = \int_{(0,1)} g(X_n(s))dm, \to \int_{(0,1)} g(X(s))dm = \mu(A).
$$

d. $3 \Rightarrow 1$: Let $A = (-\infty, x]$. Then $\partial A = \{x\}$ and $\mu_n(A) \to \mu(A)$ for all $x$ with $\mu(\{x\}) = 0$, and so $\mu_n \Rightarrow \mu$ by definition.

Remark 4.5 (On Expectations) On quick reading it is tempting to view 2 of the above proposition to state the following:

If $\mu_n \Rightarrow \mu$, then since $E[g(X_n)] = \int g(x)d\mu_n$ (section 3.1.2 of book 4) and similarly $E[g(X)] = \int g(x)d\mu$, 2 implies that the moments of $\mu_n$ converge to the moments of $\mu$.

But this is an incorrect reading. First off, the moments of a distribution are defined by $g(x) = x^m$, which while continuous are not bounded. Also recall the section 3.2.8 of book 4 on Method of Moments, where it was demonstrated by example that weak convergence does not assure convergence of moments, not even when the sequence of measures is tight. What was needed for the desired result was the extra condition of boundedness of the collection of even moments of the $\mu_n$ measures, which then allowed a proof using an uniform integrability argument.

4.1.2 Applications

Some of the most important results relating to weak convergence have already been developed in book 2. Specifically, book 2 presents the standard version of Slutsky’s theorem in chapter 5, and then in chapter 8 one finds Prokhorov’s theorem, the mapping theorem (also known as the continuous mapping theorem) and the delta method. With the results of the portmanteau theorem, some of these earlier results could have been derived with different tools. This will be seen in the next section when these earlier results are generalized from distributions on $\mathbb{R}$ to distributions on $\mathbb{R}^d$.

One application of the portmanteau conclusion in 2 was earlier seen in book 5, proposition 6.54, in the proof of the continuity theorem for the Fourier transform.
4.2 Measures on \( \mathbb{R}^j \)

4.2.1 Portmanteau Theorem on \( \mathbb{R}^j \)

The primary goals of this section are to generalize the portmanteau theorem to measures on \( \mathbb{R}^j \), and to derive a result on measures induced by measurable transformations which generalize one-dimensional results from chapter 8 of book 2. To generalize proposition 4.4 first requires a definition of weak convergence since definitions 5.19 and 8.2 of book 2 pertained to random variables \( X \), their distribution functions \( F(x) \), and the associated probability measures \( \mu \). This definition will not surprise.

**Definition 4.6** A sequence of distribution functions on \( \mathbb{R}^j \), \( \{F_n(x)\} \), will be said to **converge weakly** to a distribution function \( F(x) \), denoted \( F_n \Rightarrow F \), if \( F_n(x) \rightarrow F(x) \) for every continuity point of \( F(x) \).

A sequence of probability measures on \( \mathbb{R}^j \), \( \{\mu_n\} \), will be said to **converge weakly** to a probability measure \( \mu \), denoted \( \mu_n \Rightarrow \mu \), if \( F_n \Rightarrow F \) for the associated distribution functions (recall 8.2 of book 1).

A sequence of random vectors \( \{X_n\} \) **converges in distribution** or **converges in law** to a random vector \( X \), denoted \( X_n \Rightarrow X \), if \( F_n \Rightarrow F \) for the associated distribution functions.

We will also require an alternative formulation for the **boundary** \( \partial A \) of a set \( A \subset \mathbb{R}^j \). By definition 4.2, the boundary of a set \( A \) is the collection of points that are limit points both of sequences in \( A \) and sequences in \( A^c \equiv \mathbb{R}^j - A \), the complement of \( A \). Since \( \overline{A} \), the closure of \( A \), is by definition the collection of limit points of sequences in \( A \), it follows by definition that \( \partial A \subset \overline{A} \). To characterize \( \partial A \) now requires that we remove from \( \overline{A} \) the collection all points which are not limit points of \( A^c \). The **interior** of \( A \), denoted \( \mathring{A} \), is defined to be this set. Thus since \( \overline{A^c} \) is the collection of limit points of \( A^c \) it follows that \( \mathring{A} \equiv \mathbb{R}^j - \overline{A^c} \). Then from 4.1,

\[
\partial A \equiv \overline{A} \cap \overline{A^c} = \overline{A} - (\overline{A^c})^c
\]

and so:

\[
\partial A \equiv \overline{A} - \mathring{A}.
\] (4.2)

Note that \( \mathring{A} \) is an open set and \( \mathring{A} \subset A \). The notion of a **\( \mu \)-continuity set** \( A \) in \( \mathbb{R}^j \) is then given as in definition 4.2.

We next recall the notion of **Lipschitz continuity**, named for Rudolf Lipschitz (1832 – 1903).
4.2 MEASURES ON $\mathbb{R}^j$

**Definition 4.7** A function $f : \mathbb{R}^j \to \mathbb{R}$ is said to be **Lipschitz continuous** if there exists $L > 0$ so that for all $x, y \in \mathbb{R}^j$:

$$|f(x) - f(y)| \leq L |x - y|,$$

where $|x - y|$ denotes the standard norm in $\mathbb{R}^j$, meaning $|x - y|^2 = \sum_{i=1}^{j} (x_i - y_i)^2$.

Lipschitz continuous functions are certainly continuous, since continuity only requires that

$$\lim_{y \to x} f(y) = f(x),$$

while 4.3 actually specifies the rate of convergence. Conversely, the continuous function $f(x) = \sqrt{x}$ on $[0, \infty)$ is not Lipschitz at $x = 0$ and so Lipschitz continuity is indeed more restrictive than continuity. It is also the case that continuously differentiable functions are Lipschitz continuous, but again not conversely as $f(x) = |x|$ exemplifies.

The following example is needed for the next proof.

**Example 4.8** Let $F \subset \mathbb{R}^j$ be a closed set and define $d(x, F)$, the **distance to** $F$, by:

$$d(x, F) = \inf \{|x - y| \mid y \in F\}.$$  \hspace{1cm} (4.4)

It is an exercise to check that for any $x$ there is a $y \in F$ so that $d(x, F) = |x - y|$, and a second exercise to show by example that in general such $y$ is not unique.

To show that $d(x, F)$ is Lipschitz continuous with $L = 1$, let $x_1, x_2$ be given. Then by the triangle inequality, for all $y \in F$:

$$d(x_1, F) \leq |x_1 - y| \leq |x_1 - x_2| + |x_2 - y|,$$

and taking an infimum, $d(x_1, F) \leq |x_1 - x_2| + d(x_2, F)$. Interchanging variates obtains $|d(x_1, F) - d(x_2, F)| \leq |x_1 - x_2|$.

Similarly, $g_k(x) \equiv \max(1 - kd(x, F), 0)$ is Lipschitz continuous with $L = k$. If $x_1, x_2$ are given with both $d(x_i, F) \geq 1/k$ then both $g_k(x_i) = 0$ and the conclusion follows. If both $d(x_1, F) < 1/k$ then:

$$g_k(x_2) - g_k(x_1) = k [d(x_1, F) - d(x_2, F)],$$

and the result follows from Lipschitz continuity of $d(x, F)$. Finally, if $d(x_1, F) \geq 1/k$ and $d(x_2, F) < 1/k$ then:

$$0 \leq g_k(x_2) - g_k(x_1) = 1 - kd(x_2, F) \leq k |d(x_1, F) - d(x_2, F)|.$$

Absolute values change nothing, and we are done.
Remark 4.9 There is a technical detail in the proof of part a below which we address here. Given a countable collection of reals \( \{r_j\}_{j=1}^{\infty} \), it is apparent that \( \mathbb{R} - \bigcup_{j=1}^{\infty} r_j \) is uncountable, but we want to conclude that this set is dense in \( \mathbb{R} \). Assume otherwise, that there exists \( x \in \mathbb{R} \) so that for some \( \varepsilon > 0 \):
\[
\left( \mathbb{R} - \bigcup_{j=1}^{\infty} r_j \right) \cap (x - \varepsilon, x + \varepsilon) = \emptyset. \tag{(*)}
\]

But letting \( \tilde{B} \) denote the complement of a set \( B \):
\[
\mathbb{R} - \bigcup_{j=1}^{\infty} r_j \equiv \mathbb{R} \cap \left( \bigcup_{j=1}^{\infty} r_j \right),
\]
and thus (*) implies that \( (x - \varepsilon, x + \varepsilon) \cap \left( \bigcup_{j=1}^{\infty} r_j \right) = \emptyset \). This obtains \( (x - \varepsilon, x + \varepsilon) \setminus \bigcup_{j=1}^{\infty} r_j = \emptyset \), a contradiction.

For the proof below, the collection \( \{r_j\}_{j=1}^{\infty} \) is the union of \( k \) countable sets, one for each component.

Exercise 4.10 Develop another proof using Baire’s category theorem of proposition 4.44 of book 3. Hint: Let \( A_j \equiv \mathbb{R} - r_j \), then note that \( A_j \) is open and dense in \( \mathbb{R} \).

Proposition 4.11 (portmanteau theorem on \( \mathbb{R}^j \)) Let \( \{\mu_n\}, \mu \) be probability measures on \( \mathbb{R}^j \). Then the following are equivalent:

1. \( \mu_n \Rightarrow \mu \).
2. \( \liminf_n \mu_n(G) \geq \mu(G) \) for all open sets \( G \).
3. \( \limsup_n \mu_n(F) \leq \mu(F) \) for all closed sets \( F \).
4. \( \lim_n \mu_n(A) = \mu(A) \) for all \( \mu \)-continuity sets \( A \).
5. \( \int g(x) d\mu_n \to \int g(x) d\mu \), for every bounded, continuous real-valued function \( g \) defined on \( \mathbb{R}^j \).
6. \( \int g(x) d\mu_n \to \int g(x) d\mu \), for every bounded, Lipschitz continuous real-valued function \( g \) defined on \( \mathbb{R}^j \).

Proof. We first show that 1 – 4 are equivalent, then complete the proof with 4 \( \Rightarrow \) 5 \( \Rightarrow \) 6 and 6 \( \Rightarrow \) 3.
4.2 MEASURES ON $\mathbb{R}^j$

a. $1 \Rightarrow 2$: For any component index $k$, the hyperplane defined by $\{x \in \mathbb{R}^j | x_k = r\}$ has non-zero $\mu$-measure for at most countably many reals $r$, since otherwise $\mu$ could not be a finite measure. Hence there is a dense and uncountable set of reals denoted $D$, so that for any $k$, $\mu\left(\{x \in \mathbb{R}^j | x_k = r\}\right) = 0$ for any $r \in D$. Define a class of rectangles denoted $\mathcal{R}$, by $\mathcal{R} = \left\{ \prod_{i=1}^j (a_i, b_i) | a_i, b_i \in D \right\}$, and we now prove that every vertex of such a rectangle is a continuity point of the distribution function $F$ induced by $\mu$. If $y$ denotes such a vertex then by 8.2 of book 1:

$$F(y) = \mu\left(\prod_{i=1}^j (-\infty, y_i]\right) = \mu\left(\prod_{i=1}^j (-\infty, y_i)\right), \quad ((*)$$

Since the bounding hyperplanes $\{x \in \mathbb{R}^j | x_k = y_k\}$ have $\mu$-measure 0. This now assures that $F$ is continuous from below at such $y$. If $y^{(k)} < y$ and $y^{(k)} \to y$, where both statements are to be interpreted componentwise, then

$$\bigcup_{k=1}^\infty \left[ \prod_{i=1}^j (-\infty, y_i^{(k)}) \right] = \prod_{i=1}^j (-\infty, y_i).$$

The union on the left can be made into a nested union, and thus continuity from below of the measure $\mu$ (proposition 5.26, book 1) assures that $F(y^{(k)}) \to F(y)$. As any distribution function $F$ is continuous from above and increasing in each variable, we can now show that $F$ is in fact continuous at each $y$ as follows. Given $z$, define $z^{\text{max}}$ by $z_i^{\text{max}} = y_i + \max \{z_k - y_k\}$, then:

$$|F(y) - F(z)| \leq |F(y) - F(z^{\text{max}})| + |F(z^{\text{max}}) - F(z)|.$$

If $z^{(m)} \to y$, then $z^{(m)} \to y$ and $z^{(m)}$ approaches $y$ from above or below depending on the sign of $\max \{z_k^{(m)} - y_k\}$. In the most general case the sign will vary with $m$ and then we have two subsequences, one approaching from below and one from above. In all cases $|F(y) - F(z^{(m)}^{\text{max}})| \to 0$ since $F$ is continuous from above and below. For the second term, since $z^{(m)} \leq z^{(m)}$, note that by (*):

$$\{x | x \leq z^{(m)}\} - \{x | x \leq z^{(m)}\} \subset \bigcup_{i=1}^m \{x | x_i^{(m)} < x_i < z_i^{(m)} + c_i^{(m)}\},$$

where $c_i^{(m)} = \max \{z_k^{(m)} - y_k^{(m)}\} - (z_i^{(m)} - y_i^{(m)})$. Thus for each $m$, $F(z^{(m)}^{\text{max}}) - F(z^{(m)})$ is bounded by the sum of the $\mu$-measures of the sets on the right. Since all $c_i^{(m)} \to 0$ as $m \to \infty$, $|F(z^{\text{max}}) - F(z)| \to 0$
and continuity of $F$ follows for any $y$ that is a vertex of a rectangle in $\mathcal{R}$.

Now by 1 and definition 4.6, $F_n(y) \to F(y)$ for any vertex $y$ of a rectangle $A \in \mathcal{R}$. By proposition 8.9 of book 1 it then follows that $\mu_n(A) \to \mu(A)$ if $A \in \mathcal{R}$, and by the inclusion-exclusion formula of that book's proposition 8.8, $\mu_n(B) \to \mu(B)$ for any finite union of elements of $\mathcal{R}$. If $G$ is open, then by definition $G$ is a countable union of open rectangles, each of which is a countable union of $\mathcal{R}$-sets since $D$ is dense. Hence $G = \bigcup_m A_m$, a countable union of $\mathcal{R}$-sets, and for any $M$:

$$\mu \left[ \bigcup_{m \leq M} A_m \right] = \lim_{n \to \infty} \mu_n \left[ \bigcup_{m \leq M} A_m \right] \leq \liminf_n \mu_n(G).$$

By continuity of $\mu$ from below, $\mu \left[ \bigcup_{m \leq M} A_m \right] \to \mu(G)$ as $M \to \infty$, proving 2.

b. $2 \iff 3$: Given closed $F$, let $G = \mathbb{R}^j - F$. Then by 2,

$$\limsup_n \mu_n \left( \mathbb{R}^j - F \right) \geq \liminf_n \mu_n \left( \mathbb{R}^j - F \right) \geq \mu \left( \mathbb{R}^j - F \right).$$

By finite additivity:

$$\mu_n \left( \mathbb{R}^j - F \right) + \mu_n \left( F \right) = \mu_n \left( \mathbb{R}^j \right) \equiv 1,$$

and similarly for $\mu$, and thus

$$\limsup_n \left[ 1 - \mu_n \left( F \right) \right] \geq 1 - \mu \left( F \right).$$

This obtains $\limsup_n \mu_n \left( F \right) \leq \mu \left( F \right)$. Similarly, given $G$ define $F = \mathbb{R}^j - G$.

c. $2, 3 \Rightarrow 4$: For any $A$, since $\mathring{A}$ is an open set and $\overline{A}$ is closed, it follows by 2, 3 and monotonicity of measures:

$$\mu \left( \mathring{A} \right) \leq \liminf_n \mu_n \left( \mathring{A} \right) \leq \liminf_n \mu_n \left( A \right) \leq \limsup_n \mu_n \left( A \right) \leq \limsup_n \mu_n \left( \overline{A} \right) \leq \mu \left( \overline{A} \right).$$

If $A$ is a $\mu$-continuity set then $\mu [\partial A] = 0$, and so by 4.2 and finite additivity, $\mu \left( \mathring{A} \right) = \mu \left( \overline{A} \right)$ and 4 follows.
4.2 MEASURES ON $\mathbb{R}^J$

d. $4 \Rightarrow 1$: Let $F(x)$ and $F_n(x)$ denote as above the distribution functions associated with $\mu$ and $\mu_n$. If $x$ is a continuity point of $F$ then $\lim_{y \to x^-} F(y) = F(x)$, where this notation implies that $y_j < x_j$ for all $j$. By continuity of $\mu$ from below:

$$\lim_{y \to x^-} F(y) = \mu\left[ \hat{A}_x \right] = \mu \left[ \prod_{i=1}^j (-\infty, x_i) \right].$$

It now follows that $F(x) = \mu\left[ \hat{A}_x \right]$ and thus $\mu [ \partial A_x ] = 0$ by 4.2. This implies that $A_x$ is a $\mu$-continuity set and hence by part 4, $\mu_n [A_x] \to \mu [A_x]$. In other words, $F_n(x) \to F(x)$ for all continuity points of $F$ and thus $\mu_n \Rightarrow \mu$.

e. $4 \Rightarrow 5$: Given measurable $g$, since $\{g^{-1}(y)\}$ are disjoint for different $y \in \mathbb{R}$, it follows that $\mu [g^{-1}(y)] \neq 0$ for all but at most countably many $y$ and as in the proof of 1, the collection of $y$ with $\mu [g^{-1}(y)] = 0$ is dense in $\mathbb{R}$. Thus if $g$ is continuous and bounded with $|g| \leq C$, for any $\epsilon > 0$ we can choose increasing $\{y_k\}_{k=0}^N$ with $\mu [g^{-1}(y_k)] = 0$, $y_0 < -C$, $y_N > C$, and $y_{k+1} - y_k < \epsilon$. Define $A_k \subseteq \mathbb{R}^j$ by $A_k = \{x \mid y_{k-1} < g(x) \leq y_k\}$. Then since $g$ is continuous and hence sequentially continuous, $A_k \subseteq \{x \mid y_{k-1} \leq g(x) \leq y_k\}$. By the same argument with $A^c_k$ the complement of $A_k$, one obtains $A^c_k = \{x \mid g(x) \leq y_{k-1} \text{ or } g(x) \geq y_k\}$ and since $A_k \equiv (A^c_k)^c$ it follows that $\{x \mid y_{k-1} < g(x) < y_k\} \subseteq A_k$. By the definition of boundary,

$$\partial A_k = \overline{A}_k - \overline{A}_k \subseteq \{x \mid g(x) = y_{k-1} \text{ or } g(x) = y_k\},$$

and so by the selection of $y_k$ we conclude that $\mu [\partial A_k] = 0$ for all $k$ and thus each $A_k$ is a $\mu$-continuity set. By part 4:

$$\sum_{k=1}^N y_k \mu_n [A_k] \to \sum_{k=1}^N y_k \mu [A_k]$$

as $n \to \infty$, and so given the above $\epsilon$ there exists $M(\epsilon)$ so that for $n \geq M(\epsilon)$:

$$\left| \sum_{k=1}^N y_k \mu_n [A_k] - \sum_{k=1}^N y_k \mu [A_k] \right| < \epsilon.$$

By construction of the $A_k$-sets, for either $\nu = \mu_n$ or $\nu = \mu$:

$$\left| \int g(x) d\nu - \sum_{k=1}^N y_k \nu [A_k] \right| < \epsilon.$$
and thus for \( n \geq M(\epsilon) \):

\[
\left| \int g(x) d\mu_n - \int g(x) d\mu \right| \leq 3\epsilon.
\]

This proves 5.

f. 5 \( \Rightarrow \) 6: True by definition.

g. 6 \( \Rightarrow \) 3: Given closed \( F \subset \mathbb{R}^j \), define the function \( d(x, F) \) as in 4.4, and define \( g_k(x) = \max(1-kd(x,F),0) \) as in example 4.8. Then \( g_k(x) \) is bounded and Lipschitz continuous, monotonically decreasing in \( k \) for each \( x \), and since \( F \) is closed, \( g_k(x) \to \chi_F(x) \) pointwise. Recall that \( \chi_F(x) \) is the characteristic function of \( F \), defined to be 1 for \( x \in F \) and 0 otherwise. Hence for any \( k \),

\[
\mu_n(F) = \int \chi_F(x) d\mu_n \leq \int g_k(x) d\mu_n.
\]

Taking limit superior and applying 6:

\[
\limsup_n \mu_n(F) \leq \lim \int g_k(x) d\mu_n = \int g_k(x) d\mu.
\]

This is true for all \( k \), and by the bounded convergence theorem \( \lim_k \int g_k(x) d\mu \to \mu(F) \), and part 3 follows.

Example 4.12 That statements 5, 6 cannot in general be extended beyond bounded functions is not difficult to illustrate. Define \( \mu_n \) on \( \mathbb{R} \) by \( \mu_n \{n\} = 1/n \) and \( \mu_n \{0\} = 1 - 1/n \), and \( \mu \) on \( \mathbb{R} \) by \( \mu \{0\} = 1 \). Then \( \mu_n \Rightarrow \mu \) (and in fact \( \mu_n \) also converges to \( \mu \) strongly and in total variation, see Scheffé’s theorem of proposition 4.36 below). But with \( g(x) \equiv x \),

\[
1 = \int g(x) d\mu_n \Rightarrow \int g(x) d\mu = 0.
\]

4.2.2 Applications

In this section we generalize many of the results noted above that were developed in book 2 from distributions on \( \mathbb{R} \) to distributions on \( \mathbb{R}^j \). In addition we introduce part one of a new result called the Cramér-Wold device, and provide an easy but very useful application that was already called into service in chapter 8 of book 2 for the development of multivariate extreme value theory. Part two of Cramér-Wold is of necessity deferred to chapter 6.
4.2 MEASURES ON $\mathbb{R}^J$

(Continuous) Mapping Theorem

Proposition 4.11 allows a relatively easy generalization of two mapping theorems earlier developed on $\mathbb{R}$ in propositions 8.35 and 8.37 of book 2. The first addresses weak convergence of measures induced by transformations, for which the book 2 result required Skorokhod’s Representation theorem for its proof. The second is essentially a corollary of the first and provides an application of this result to various modes of convergence of transformed random vectors. We will then turn to the Cramér-Wold device.

Recall that the notion of a measure induced by a transformation was introduced in definition 3.9 of book 5. Specifically, changing notation from a mathematics to a probability context:

**Definition 4.13** Given measure spaces $(S, \mathcal{E}, \lambda)$ and $(S', \mathcal{E}', \lambda')$, a transformation $T : S \to S'$ is said to be measurable, or more specifically $\mathcal{E}/\mathcal{E}'$-measurable, if $T^{-1}(\mathcal{E}') \subset \mathcal{E}$. That is:

$$T^{-1}(A') \in \mathcal{E} \text{ for all } A' \in \mathcal{E}'.$$ 

A measurable transformation then induces a new measure on the range space $S'$, denoted $(S', \mathcal{E}', \lambda_T)$, where $\lambda_T$ is called the measure induced by $T$ and defined on $A' \in \mathcal{E}'$ by:

$$\lambda_T(A') = \lambda[T^{-1}(A')] \quad (4.5)$$

**Example 4.14** The classic example of a measurable transformation is a random variable or random vector defined on $(S, \mathcal{E}, \lambda)$ with range space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), m)$ or $(\mathbb{R}^j, \mathcal{B}(\mathbb{R}^j), m^j)$. In either case the measure induced by $X$ is precisely the probability measure $\mu \equiv \mu_X$ defined as in 4.5 on a Borel set $A$:

$$\mu(A) = \lambda[X^{-1}(A)] \quad (4.6)$$

For the current investigation we denote $T$ by $h : \mathbb{R}^j \to \mathbb{R}^k$ since this is a more natural notation for such transformations. Assume that $h$ is Borel measurable, which is to say $\mathcal{B}(\mathbb{R}^j)/\mathcal{B}(\mathbb{R}^k)$-measurable, and let $D_h$ denote the collection of discontinuities of $h$. First note that for any such $h$ that $D_h \in \mathcal{B}(\mathbb{R}^j)$ and is thus a measurable set.

**Exercise 4.15** Prove that $D_h$ is measurable even if $h$ is not measurable. Hint: Define the set $A(\epsilon, \delta)$ for arbitrary $\epsilon, \delta > 0$ by:

$$A(\epsilon, \delta) = \{x | \text{there exists } y, z \in B_\delta(x) \text{ and } |h(y) - h(z)| > \epsilon\},$$
where $B_\delta(x) = \{y \mid |y - x| < \delta\}$. Prove that $A(\epsilon, \delta)$ is open, so $x \in A(\epsilon, \delta)$ implies $B_{\delta'}(x) \subset A(\epsilon, \delta)$ for some $\delta' > 0$. Then prove that:

$$D_h = \bigcup_{\epsilon} \bigcap_{\delta} A(\epsilon, \delta),$$

where these set functions are over all positive rational $\epsilon, \delta$.

Thus not only is $D_h$ Borel measurable, but it is a $G_\delta$-set as in notation 2.16 of book 1, which means it is a countable union (the "\(\sigma\)" of sets, each of which is a countable intersection (the "\(\delta\)" of open (the "\(G\)" sets.

Generalizing proposition 8.35 of book 2, we now show that if $\mu(D_h) = 0$, then weak convergence $\mu_n \Rightarrow \mu$ is preserved in the induced measures. This is sometimes called the continuous mapping theorem, and not only because $h$ is sometimes assumed to be continuous. More generally, the term continuity here also implies "convergence preserving," since continuous functions have this property, and the general theorem states that something less than continuity of $h$ is sufficient to preserve convergence of measures.

**Exercise 4.16** Prove that if $x_n \to x$ and $f$ is continuous at $x$, then $f(x_n) \to f(x)$. Note: Here $x_n, x \in \mathbb{R}^j$ and $f : \mathbb{R}^j \to \mathbb{R}^k$.

**Proposition 4.17 (Mapping theorem on $\mathbb{R}^j / \mathbb{R}^k$)** Let $\{\mu_n, \mu\}$ be a collection of probability measures on $(\mathbb{R}^j, B(\mathbb{R}^j))$ with $\mu_n \Rightarrow \mu$, and $h : (\mathbb{R}^j, B(\mathbb{R}^j)) \to (\mathbb{R}^k, B(\mathbb{R}^k))$ a Borel measurable transformation such that $\mu(D_h) = 0$. Then $(\mu_n)_h \Rightarrow \mu_h$.

**Proof.** For closed $F \in B(\mathbb{R}^k)$, note that as a closed set that $h^{-1}(F) \in B(\mathbb{R}^j)$. If $x \in h^{-1}(F)$ there exists $\{x_m\} \subset h^{-1}(F)$ with $x_m \to x$. If $x$ is a continuity point of $h$ then $h(x_m) \to h(x)$ as noted above, and since $F$ is closed it follows that $h(x) \in F$ and thus $x \in h^{-1}(F)$. Hence, $\overline{h^{-1}(F)} \subset h^{-1}(F) \cup D_h$.

By monotonicity and subadditivity of measures and part 3 of the proposition 4.11:

$$\limsup \mu_n [h^{-1}(F)] \leq \limsup \mu_n [h^{-1}(F)] \leq \mu [h^{-1}(F)] \leq \mu [h^{-1}(F)] + \mu [D_h].$$

Thus if $\mu [D_h] = 0$, then by definition 4.13:

$$\limsup \mu_n (\mu_n)_h (F) \leq \mu_h (F),$$

and by part 3 again, $(\mu_n)_h \Rightarrow \mu_h$. ■

**Remark 4.18** Note that the conclusion of this proposition is always valid for continuous $h$. 
There is another version of the continuous mapping theorem that applies to random variable sequences that converge in one of several ways. The one-variable version of this result is proposition 8.37 of book 2. For its statement, we need to recall some definitions.

**Definition 4.19** Given a probability space \((\mathcal{S}, \mathcal{F}, \lambda)\) and random vectors \(X\) and \(\{X_n\}_{n=1}^{\infty}\) with range in \(\mathbb{R}^J\):

1. **\(X_n\) converges to \(X\) in probability**, denoted \(X_n \rightarrow_p X\), if for every \(\epsilon > 0\):
   \[
   \lim_{n \to \infty} \lambda(\{|X_n - X| \geq \epsilon\}) = 0, \quad (4.7)
   \]
   where \(|X_n - X|\) in 4.7 is interpreted in terms of the standard norm on \(\mathbb{R}^J\).

2. **\(X_n\) converges to \(X\) with probability 1**, or **\(X_n\) converges to \(X\) almost everywhere (almost surely)**, denoted \(X_n \to_1 X\) or \(X_n \to_{a.e.} X\) or \(X_n \to_{a.s.} X\), if:
   \[
   \lambda(\{\lim_{n \to \infty} X_n = X\}) = 1, \quad (4.8)
   \]
   where \(\{\lim_{n \to \infty} X_n = X\}\) in 4.8 is interpreted in terms of the standard norm on \(\mathbb{R}^J\).

3. **\(X_n\) converges in distribution to \(X\)**, or **converges in law**, denoted \(X_n \to_d X\), if \(F_n \Rightarrow F\), meaning that \(F_n(x) \to F(x)\) for every continuity point of \(F\). Equivalently, \(X_n \to_d X\) if \(\mu_n \Rightarrow \mu\) for the associated probability measures.

**Remark 4.20** Note that for the definition of \(X_n \to_p X\), it is sometime convenient to use an alternative norm on \(\mathbb{R}^J\). Recalling the discussion on Banach spaces in section 4.1 of book 5, the standard norm is also called the \(l_2\)-norm, while the general \(l_p\)-norm is defined for \(1 \leq p < \infty\) by:

\[
\|x\|_p = \left(\sum_{i=1}^{J} |x_i|^p\right)^{1/p}, \quad (4.9)
\]

and for \(p = \infty\) by:

\[
\|x\|_\infty = \max\{|x_i|\}. \quad (4.10)
\]

Any one of these norms can be used in the definition of convergence in probability since they are all equivalent norms. That is, given \(p, p'\) there exists positive \(c_{p,p'}\) and \(C_{p,p'}\) so that for all \(x\):

\[
c_{p,p'} \|x\|_{p'} \leq \|x\|_p \leq C_{p,p'} \|x\|_{p'}. \quad (4.11)
\]

Thus if 4.7 is true for \(p = 2\) it is true for any \(p\), and conversely.
CHAPTER 4 WEAK CONVERGENCE OF MEASURES

The following version of the mapping theorem is sometimes called the Mann-Wald theorem, named for Henry Mann (1905 – 2000) and Abraham Wald (1902 – 1950), who published it in 1943. For its proof we require various results from book 2 on relationships between these modes of convergence. For example, proposition 5.21 states that $X_n \rightarrow_{a.e.} X$ implies $X_n \rightarrow_P X$, while proposition 5.25 states that if $X_n \rightarrow_P X$ then there is a subsequence such that $X_{n_m} \rightarrow_{a.e.} X$. It will be seen that the proofs of such results are dimensionless relative to the random variables if we simply interpret expressions such as $|X_n - X|$ in the appropriate $\mathbb{R}^j$-norm.

Note that by 4.6, the condition below on $D_h$ is comparable to that in proposition 4.17, that $\mu_X(D_h) = 0$ where $\mu_X$ is the probability measure on $\mathbb{R}^j$ induced by $X$. The proof of the following result is very similar to that of the random variable counterpart, proposition 8.37 of book 2, with one exception. Part 3 below has a simpler proof that just applies proposition 4.17.

**Proposition 4.21 (Mann-Wald theorem)** Let $\{X_n, X\}$ be random vectors defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with range in $\mathbb{R}^j$, and $h$ a Borel measurable function $h : \mathbb{R}^j \rightarrow \mathbb{R}^k$ with discontinuity set $D_h$. If $\lambda\{X^{-1}(D_h)\} = 0$ then:

1. If $X_n \rightarrow_{a.e.} X$ then $h(X_n) \rightarrow_{a.e.} h(X)$.
2. If $X_n \rightarrow_P X$ then $h(X_n) \rightarrow_P h(X)$.
3. If $X_n \rightarrow_d X$ then $h(X_n) \rightarrow_d h(X)$.

In particular, if $h$ is a continuous function then $X_n \rightarrow_{\ast} X$ implies that $h(X_n) \rightarrow_{\ast} h(X)$ for all three notions of convergence.

**Proof.** 1. If $X_n \rightarrow_{a.e.} X$ then there is an exceptional set $E \in \mathcal{E}$ with $\lambda(E) = 0$ and $X_n(s) \rightarrow X(s)$ for $s \in \bar{E}$. Let $D = \{X^{-1}(D_h)\}$, and note that for $s \in (D \cup E)^c$ that $X_n(s) \rightarrow X(s)$ and $X(s)$ is a continuity point of $h$. Thus $h(X_n(s)) \rightarrow h(X(s))$ for $s \in (D \cup E)^c$ by exercise 4.16, and since $\lambda(D \cup E) = 0$ this obtains $h(X_n) \rightarrow_{a.e.} h(X)$.

2. Arguing by contradiction, if $X_n \rightarrow_P X$ and $h(X_n) \rightarrow_P h(X)$ then given $\delta > 0$ and $\epsilon > 0$ there is a subsequence $\{X_{n_m}\}$ so that for all $m$:

$$\lambda\{|h(X_{n_m}) - h(X)| > \epsilon\} > \delta.$$ 

Further, because $\lambda\{X^{-1}(D_h)\} = 0$ it can be assumed that the sets $\{|h(X_{n_m}) - h(X)| > \epsilon\}$ include only continuity points of $h$. Since it is now also true that $X_{n_m} \rightarrow_P X$, apply proposition 5.25 of book 2 to identify a subsequence of $\{X_{n_m}\}$ so
4.2 MEASURES ON $\mathbb{R}^J$

that $X_{nm_k} \rightarrow_{a.e.} X$, and thus by part 1 it follows that $h \left( X_{nm_k} \right) \rightarrow_{a.e.} h(X)$.

This new subsequence retains the property that $h \left( X_{nm_k} \right) \rightarrow_p h(X)$ and this contradicts proposition 5.21 of book 2, that convergence with probability 1 assures convergence in probability.

3. By definition, $X_n \rightarrow_d X$ means that $\mu_n \Rightarrow \mu$ for the associated probability measures on $(\mathbb{R}^j, \mathcal{B}(\mathbb{R}^j))$. Since $\mu(D_h) = \lambda \left( X^{-1}(D_h) \right) = 0$, proposition 4.17 states that $(\mu_n)_h \Rightarrow \mu_h$. Letting $F_n$ and $F$ denote the distribution functions associated with $(\mu_n)_h$ and $\mu_h$, this implies that $F_n(x) \rightarrow F(x)$ for every continuity point of $F$. Now with $A_x \equiv \prod_{i=1}^j (-\infty, x_i]$, applying definitions:

$$F_n(x) \equiv (\mu_n)_h[A_x] \equiv \mu_n \left[ h^{-1}(A_x) \right] \equiv \lambda \left[ X_n^{-1} \left( h^{-1}(A_x) \right) \right] = \lambda \left( h(X_n) \right)^{-1} \left( A_x \right).$$

In other words, $F_n(x)$ is the distribution function of $h(X_n)$, and similarly $F(x)$ is the distribution function of $h(X)$. Thus $F_n(x) \rightarrow F(x)$ for every continuity point of $F$ obtains by definition that $h(X_n) \rightarrow_d h(X)$. ■

Cramér-Wold Device - Part 1

In remark 9.51 of the book 2 discussion on multivariate extreme value theory, the above mapping theorem was referenced as the key ingredient for the following result. Given the weak convergence of joint distribution functions $F_n \Rightarrow F$, it was necessary to know that the marginal distribution functions of $\{F_n\}$ also converge weakly to the respective marginal distribution functions of $F$. We now demonstrate this result as a simple corollary to the next result, which we call Part 1 of the Cramér-Wold theorem.

A quite remarkable result, the Cramér-Wold device or the Cramér-Wold theorem, named for Harald Cramér (1893 – 1985) and Herman Wold (1908 – 1992), states that $X_n \Rightarrow X$ for random vectors if and only if $t \cdot X_n \Rightarrow t \cdot X$ for all $t$. Recall that the dot product of $j$-vectors is defined by:

$$x \cdot y \equiv \sum_{k=1}^j x_k y_k. \quad (4.12)$$

What we call Part 1 is the easier part, that $X_n \Rightarrow X$ for random vectors only if $t \cdot X_n \Rightarrow t \cdot X$ for all $t$.

Given $t \in \mathbb{R}^j$ and $\alpha \in \mathbb{R}$, note that $H_t(\alpha) \equiv \{ x \in \mathbb{R}^j : t \cdot x \leq \alpha \}$ is a half-space in $\mathbb{R}^j$ with bounding hyperplane $\{ x \in \mathbb{R}^j : t \cdot x = \alpha \}$. The distribution function associated with $X' \equiv t \cdot X$ is then given by:

$$F_{X'}(\alpha) = \mu[H_t(\alpha)] ,$$
where $\mu$ is the probability measure on $\mathbb{R}^j$ induced by the random variable $X$. Thus Part 1 states that if $F_n \Rightarrow F$, which is the distribution function equivalent of $X_n \Rightarrow X$, then for any $t \in \mathbb{R}^j$,

$$F_{X_n^t} \Rightarrow F_{X^t}.$$  

Part 2 of this result states that by demonstrating $F_{X_n^t} \Rightarrow F_{X^t}$ for all $t$, meaning that all of the one-dimensional distributions defined relative to half-spaces converge weakly for all $t$, this assures the weak convergence of the joint distributions. The proof of this will require the tools of characteristic functions, and in particular a continuity theorem on $\mathbb{R}^n$ extending proposition 6.54 of book 5. This result is deferred to chapter 6.

**Proposition 4.22 (Cramér-Wold theorem - Part 1)** Let $\{X_n\}_{n=1}^\infty, X,$ be random vectors defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with range in $\mathbb{R}^j$ and assume that $X_n \Rightarrow X$. Then for any $t \in \mathbb{R}^j$,

$$t \cdot X_n \Rightarrow t \cdot X. \quad (4.13)$$

**Proof.** Recall that $X_n \Rightarrow X$ by definition means that $\mu_n \Rightarrow \mu$ for the associated probability measures. Given $t \in \mathbb{R}^j$, define $h : \mathbb{R}^j \to \mathbb{R}$ by $h(x) = t \cdot x$. Then $h$ is continuous and thus measurable, $\mu(D_h) = 0$, and so by proposition 4.17 $(\mu_n)_h \Rightarrow \mu_h$. Now by 4.5, $\mu_h$ is the probability measure associated with the random variable $h(X)$. If $A \in \mathcal{B}(\mathbb{R})$:

$$\mu_h[A] \equiv \mu[h^{-1}(A)] \equiv \lambda[h(X)^{-1}(A)],$$

and similarly for $(\mu_n)_h$. Thus $(\mu_n)_h \Rightarrow \mu_h$ means by definition that $t \cdot X_n \Rightarrow t \cdot X$. $\blacksquare$

We now turn to the result on weak convergence of marginal distribution functions noted above and needed in the book 2 discussion on multivariate extreme value theory.

**Corollary 4.23 (Cramér-Wold theorem - Part 1)** Let $\{X_n\}_{n=1}^\infty, X,$ be random vectors defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with range in $\mathbb{R}^j$ with $X_n \Rightarrow X$ and let $F_n, F$ denote the associated distribution functions. If $F_{n,k}, F_k$ denote the marginal distribution functions of the $k$th component variates $X_{n,k}$ and $X_k$, where $X_n = (X_{n,1}, \ldots, X_{n,j})$ and $X = (X_1, \ldots, X_j)$, then $F_{n,k} \Rightarrow F_k$.

**Proof.** Define $t \in \mathbb{R}^j$ by $t_k = 1$ and $t_i = 0$ otherwise. Then $t \cdot X_n = X_{n,k}$ and $t \cdot X = X_k$, and this is a restatement of 4.13. $\blacksquare$

**Exercise 4.24** Prove that the rest of the $2^j - 2$ proper marginal distribution functions of the sequence $\{F_n\}_{n=1}^\infty$, as defined in definition 3.34 of book 2, also converge weakly to the respective marginal distributions of $F$. 


4.2 MEASURES ON $\mathbb{R}^J$

Slutsky’s Theorem

Slutsky’s theorem, named for Evgeny "Eugen" Slutsky (1880 – 1948), was introduced in proposition 5.29 of book 2, and addressed the following question. If $X_n \to_d X$ and $Y_n \to Y$ where "$\to$" denotes convergence in some manner, does $X_n + Y_n \to_d X + Y$, or $X_nY_n \to_d XY$, etc.? The earlier version of Slutsky’s theorem and associated exercise 5.30 provided affirmative results. If $X_n \to_d X; Y_n \to_P a$ and $Z_n \to_P b$, then:

1. $X_n + Y_n \to_d X + a$,
2. $X_nY_n \to_d aX$,
3. $X_n/Y_n \to_d X/a$ if $a \neq 0$,
4. $X_nY_n + Z_n \to_d aX + b$.

Example 5.31 then demonstrated that if $X_n \to_d X$ and $Y_n \to_d Y$ then it need not be the case that $X_n + Y_n \to_d X + Y$, nor that $X_nY_n \to_d XY$.

The earlier result can now be extended with the aid of the portmanteau theorem and the mapping theorem.

**Proposition 4.25 (Slutsky’s theorem)** Given sequences of random variables $X_n : \mathcal{S} \to \mathbb{R}^j$ and $Y_n : \mathcal{S} \to \mathbb{R}^k$ defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with $X_n \to_d X$ and $Y_n \to_P a$, where $a \in \mathbb{R}^k$ is a constant. If $h : \mathbb{R}^{j+k} \to \mathbb{R}$ is continuous, then:

$$h(X_n, Y_n) \to_d h(X, a).$$  \hfill (4.14)

**Proof.** By the Mann-Wald theorem of proposition 4.21 we only need to prove that $(X_n, Y_n) \to_d (X, a)$. To this end, let $\mu_n$ denote the probability measure on $\mathbb{R}^{j+k}$ induced by $(X_n, Y_n)$, and $\mu$ analogously defined for $(X, a)$. So for $A \in \mathcal{B}(\mathbb{R}^{j+k})$:

$$\mu_n(A) = \lambda \left[ (X_n, Y_n)^{-1}(A) \right], \quad \mu(A) = \lambda \left[ (X, a)^{-1}(A) \right].$$

Note that $\mu$ is well defined with a constant since $(X, a)^{-1}(A) = \{ s \mid (X(s), a) \in A \}$. To prove that $(X_n, Y_n) \to_d (X, a)$ is by definition to prove that $\mu_n \Rightarrow \mu$, and we do this with part 6 of the portmanteau theorem.

Let $g : \mathbb{R}^{j+k} \to \mathbb{R}$ be bounded by $K > 0$ and Lipschitz continuous with constant $L$. By definition of $Y_n \to_P a$, given $\epsilon > 0, \delta > 0$ there exists $N$ so that $\mu_n \left[ Y_n - a \right] \leq \epsilon$ for $n \geq N$. Let $A_{\epsilon, \delta} \in \mathbb{R}^{j+k}$ be defined by $A_{\epsilon, \delta} = \{ y \mid \|y - a\| \leq \delta \}$, and let $A_{\epsilon, \delta}^c = \{ (x, y) \mid \|y - a\| > \delta \}$. For notational
convenience we denote the respective characteristic function of these sets as $\chi_1(x,y)$ and $\chi_2(x,y)$. Then:

$$\left| \int g(x,y)d\mu_n - \int g(x,a)d\mu \right| \leq \left| \int [g(x,y) - g(x,a)] \chi_1(x,y)d\mu_n \right|$$

$$+ \left| \int [g(x,y) - g(x,a)] \chi_2(x,y)d\mu_n \right|$$

$$+ \left| \int g(x,a)d\mu_n - \int g(x,a)d\mu \right|$$

$$+ \left| \int [g(x,a) - g(x,y)]d\mu \right|.$$

Taking these bounding integrals in turn, note that for the first that the Lipschitz condition and definition of $\chi_1$ obtain for $n \geq N$:

$$\left| \int [g(x,y) - g(x,a)] \chi_1(x,y)d\mu_n \right| \leq L \int |y - a| \chi_1(x,y)d\mu_n \leq L\delta.$$

For the second integral, recall $g$ is bounded by $K$, and then by the definition of $\chi_2$ obtains for $n \geq N$:

$$\left| \int [g(x,y) - g(x,a)] \chi_2(x,y)d\mu_n \right| \leq 2K \int \chi_2(x,y)d\mu_n \leq 2K\epsilon.$$

For the third, let $\nu_n$ and $\nu$ be the probability measures on $\mathbb{R}^j$ induced by $X_n$ and $X$. If $F_n(x,y)$ is the distribution function associated with $\mu_n$ and $G_n(x)$ the distribution function associated with $\nu_n$, then by proposition 3.36 and remark 3.37 of book 2, $G_n(x)$ is the marginal distribution of $F_n(x,y)$. Using Riemann-Stieltjes notation:

$$\int g(x,a)d\mu_n = \int g(x,a)dF_n(x,y) = \int g(x,a)dG_n(x) = \int g(x,a)d\nu_n.$$

Using the same logic for the $d\nu$ integral:

$$\left| \int g(x,a)d\mu_n - \int g(x,a)d\nu \right| = \left| \int g(x,a)d\nu_n - \int g(x,a)d\nu \right|.$$

Now the assumption that $X_n \rightarrow_d X$ means by definition that $\nu_n \Rightarrow \nu$. Since $f(x) \equiv g(x,a)$ is Lipschitz continuous on $\mathbb{R}^j$, statement 6 of the portmanteau theorem assures that this expression converges to zero as $n \rightarrow \infty$. That is, for $\epsilon$ as above and $n \geq N'$:

$$\left| \int g(x,a)d\mu_n - \int g(x,a)d\mu \right| \leq \epsilon.$$
Finally, for the last integral let \( \chi_1(x,y) \) and \( \chi_2(x,y) \) be defined as above. Using the Lipschitz continuity of \( g \):

\[
\left| \int [g(x,a) - g(x,y)] \, d\mu \right| \leq L \int |y - a| \, d\mu = L \int |y - a| \chi_1(x,y) \, d\mu,
\]

since by definition of \( \mu \) above, \( \mu(A_{x,a}^c) = 0 \). Thus:

\[
\left| \int [g(x,a) - g(x,y)] \, d\mu \right| \leq L\delta.
\]

Combining, for \( n \geq \max(N, N') \):

\[
\left| \int g(x,y) \, d\mu_n - \int g(x,y) \, d\mu \right| \leq 2L\delta + (2K + 1)\epsilon.
\]

As \( \epsilon, \delta \) were arbitrary, this proves that for all bounded, Lipschitz continuous functions \( g \),

\[
\int g(x,y) \, d\mu_n \to \int g(x,y) \, d\mu.
\]

The proof that \( (X_n, Y_n) \to_d (X, a) \) is complete by part 6 of the portmanteau theorem.

**Remark 4.26** The assumption that \( h \) is continuous can be further generalized to \( h \) Borel measurable and \( \mu(D_h) \equiv \lambda \left[ \{(X, a)^{-1}(D_h)\} \right] = 0 \). Nothing in the proof changes since the proof that \( (X_n, Y_n) \to_d (X, a) \) is independent of \( h \), and the mapping theorem remains applicable.

**The Delta Method**

The **Delta method**, also written as the \( \Delta \)-method, was introduced in proposition 8.40 of book 2 in the one variable context. It addressed transformations of normalized sequences of random variables. In this case, the result is:

**Proposition 8.40 (The \( \Delta \)-method, book 2)** Let \( \{X_n, X\} \) be random variables defined on a probability space \( (\mathcal{S}, \mathcal{E}, \lambda) \) and assume that there exists a positive sequence \( \{c_n\} \subset \mathbb{R} \) with \( c_n \to \infty \) as \( n \to \infty \), and a constant \( x_0 \) so that \( c_n(X_n - x_0) \to_d X \). If \( g \) is a function that is differentiable at \( x_0 \) then:

\[
c_n [g(X_n) - g(x_0)] \to_d g'(x_0)X. \tag{4.15}
\]
A common application of this result is in the case where $X$ has a standard normal distribution, and thus the delta method provides additional asymptotic results in cases where the central limit theorem (proposition 5.14 of book 4) applies.

In this section we address two generalizations. The first is to answer the question: What if $g(x_0) = 0$? The above proposition 8.40 then states that $c_n [g(X_n) - g(x_0)] \to_d 0$, a degenerate random variable, and by proposition 5.27 of book 2 this implies that $c_n [g(X_n) - g(x_0)] \to_P 0$. The answer to this question will only use the tools of book 2. The second generalization is to determine the multivariate analog of the earlier result, so that now $X_n, X$ are random vectors, and the proof uses the tools of this chapter.

For the first result, recall that if a function of a single variable $g(x)$ is $m$-times differentiable at $x_0$, then the $m$th-order Taylor polynomial of $g(x)$ about $x_0$, $T_m(x)$, is given by:

$$T_m(x) = \sum_{i=1}^{m} \frac{g^{(i)}(x_0)}{i!}(x - x_0)^i,$$

(4.16)

where $g^{(i)}(x_0)$ denotes the $i$th derivative of $g(x)$ evaluated at $x_0$. An important property of $T_m(x)$ is that

$$g(x) = T_m(x) + R_m(x),$$

with remainder term $R_m(x)$ given by:

$$R_m(x) = h_m(x)(x - x_0)^m,$$

where $h_m(x) \to 0$ as $x \to x_0$. This polynomial and the associated Taylor series $T_m(x)$ when it converges to $g(x)$, is named for Brook Taylor (1685 – 1731).

Assuming a little more of the function $g(x)$, this remainder term can be more usefully represented. If $g^{(m+1)}(x)$ is continuous in an open interval $I \equiv (x_0 - a, x_0 + a)$ about $x_0$, the Lagrange form of the remainder, named for Joseph-Louis Lagrange (1736 – 1813) states that for any $x \in I$:

$$R_m(x) = \frac{g^{(m+1)}(x')}{(m+1)!}(x - x_0)^{m+1},$$

(4.17)

where $x'$ is "between" $x$ and $x_0$. That is, either $x < x' < x_0$ or $x_0 < x' < x$, and this is often expressed $x' = tx + (1-t)x_0$ for $0 < t < 1.$
**Proposition 4.27 (The General ∆-Method)** Let \( \{X_n, X\} \) be random variables defined on a probability space \((\mathcal{S}, \mathcal{E}, \lambda)\) and assume that there exists a positive sequence \(\{c_n\} \subset \mathbb{R}\) with \(c_n \to \infty\) as \(n \to \infty\), and a constant \(x_0\) so that \(c_n (X_n - x_0) \to_d X\). Let \(g\) be a function that is \((m + 1)\)-times differentiable at \(x_0\) with \(g^{(m+1)}(x)\) continuous. If \(g'(x_0) = \ldots = g^{(m)}(x_0) = 0\), then:
\[
c_n^{m+1} [g(X_n) - g(x_0)] \to_d \frac{g^{(m+1)}(x_0)}{(m+1)!} X^{m+1}. \tag{4.18}
\]

**Proof.** By the above discussion:
\[
c_n^{m+1} [g(X_n) - g(x_0)] = \frac{g^{(m+1)}(X_n')}{(m+1)!} c_n^{m+1} (X_n - x_0)^{m+1}, \tag{(*)}
\]
where \(X_n' = tX_n + (1-t)x_0\) for \(0 < t < 1\). Now \(c_n (X_n - x_0) \to_d X\) and \(c_n \to \infty\) assure that \(X_n \to_p X_0\) by the last line of the proof of proposition 8.39 of book 2. This then obtains that \(X_n' \to_p x_0\) since \(X_n' - x_0 = t(X_n - x_0)\) for \(0 < t < 1\) assures that \(|X_n' - x_0| < |X_n - x_0|\) and hence:
\[
\lambda \left( \{|X_n' - x_0| \geq \epsilon\} \right) \leq \lambda \left( \{|X_n - x_0| \geq \epsilon\} \right).
\]
Then by the continuous mapping theorem of proposition 8.37 of book 2 with \(h(x) \equiv g^{(m+1)}(x) / (m+1)!\):
\[
\frac{g^{(m+1)}(X_n')}{(m+1)!} \to_p \frac{g^{(m+1)}(x_0)}{(m+1)!}.
\]
Similarly, since \(c_n (X_n - x_0) \to_d X\) the continuous mapping theorem with \(h(x) = x^{m+1}\) obtains that \(c_n^{m+1} (X_n - x_0)^{m+1} \to_d X^{m+1}\). Applying Slutsky’s theorem of proposition 5.29 of book 2 to the product in (*) yields 4.18. ■

**Example 4.28** Let \(\{Y_j\}_{j=1}^\infty\) be a sample for a random variable \(Y\) with mean \(\mu\) and variance \(\sigma^2\), and define \(X_n = \sum_{j=1}^n Y_j / n\). Then by the central limit theorem of proposition 5.14 of book 4, \(\frac{X_n - \mu}{\sigma} \to_d Z\), where \(Z\) has a standard normal distribution. Let \(g(x) = (x - \mu)^{m+1}\), then \(g(\mu) = g'(\mu) = \ldots = g^{(m)}(\mu) = 0\), and thus by 4.18:
\[
\left( \frac{\sqrt{n}}{\sigma} \right)^{m+1} (X_n - \mu)^{m+1} \to_d Z^{m+1}.
\]
In particular, when \(m = 1\):
\[
\frac{(X_n - \mu)^2}{\sigma^2/n} \to_d Z^2,
\]
recalling that \(Z^2\) is chi-squared with 1-degree of freedom, also denoted \(\chi^2_{1 \text{ d.f.}}\).
For the multivariate version of the delta method we again recall the Taylor series expansions. We will not need all of what follows but provide the extra details for completeness. Multivariate Taylor series have nearly identical properties to the one variable case because they can be derived from this model. Given \( g(x_1, \ldots, x_j) \), so \( g : \mathbb{R}^j \rightarrow \mathbb{R} \), we can represent the values of \( g \) in a given direction from \( x_0 \) in terms of a single parameter \( t \):
\[
g(t) = g(x_0 + t(x - x_0)),
\]
to which we can apply now 4.16 about \( t = 0 \). Each derivative of \( g(t) \) of order \( i \) will then equal a summation of partial derivatives of that order. For example:
\[
\begin{align*}
\tilde{g}^'(0) &= \sum_{i=1}^{j} \frac{\partial g(x_0)}{\partial x_i}, \\
\tilde{g}''(0) &= \sum_{j=1}^{j} \sum_{i=1}^{j} \frac{\partial^2 g(x_0)}{\partial x_i \partial x_j}.
\end{align*}
\]
A little thought and it becomes clear a more efficient notation is required.

Let \( \alpha \equiv (\alpha_1, \ldots, \alpha_j) \) denote a multi-index of nonnegative integers and define:
\[
|\alpha| = \sum_{i=1}^{j} \alpha_i, \quad \alpha! = \prod_{i=1}^{j} \alpha_i!, \quad x^\alpha = \prod_{i=1}^{j} x_i^{\alpha_i}.
\]
The partial derivative \( \partial^\alpha \) has order \( |\alpha| \) and is defined by:
\[
\partial^\alpha g = \frac{\partial^{|\alpha|} g}{\partial x_1^{\alpha_1} \cdots \partial x_j^{\alpha_j}}.
\]
If the function \( g(x) \) is \( m \)-times differentiable at \( x_0 \), meaning \( \partial^\alpha g \) exists for all \( \alpha \) with \( |\alpha| \leq m \), then the \( m \)th-order Taylor polynomial of \( g(x) \) about \( x_0 \), \( T_m(x) \), is given by the above process by:
\[
T_m(x) = \sum_{|\alpha| \leq m} \frac{\partial^\alpha g(x_0)}{\alpha!} (x - x_0)^\alpha, \tag{4.19}
\]
where \( \partial^\alpha g(x_0) \) denotes this derivative of \( g(x) \) evaluated at \( x_0 \). As above it follows that
\[
g(x) = T_m(x) + R_m(x),
\]
with remainder term \( R_m(x) \) given by:
\[
R_m(x) = \sum_{|\alpha| = m} h_\alpha(x)(x - x_0)^\alpha,
\]
4.2 MEASURES ON $\mathbb{R}^J$

where $h_\alpha(x) \to 0$ for all $\alpha$ as $x \to x_0$. If $\partial^\alpha g(x)$ is continuous in an open ball $B \equiv \{x | |x-x_0| < a\}$ about $x_0$ for all $|\alpha| = m + 1$, the Lagrange form of the remainder states that for any $x \in B$:

$$R_m(x) = \sum_{|\alpha|=m+1} \frac{\partial^\alpha g(x')}{\alpha!} (x - x_0)^\alpha,$$

where $x'$ is "between" $x$ and $x_0$, meaning $x' = tx + (1-t)x_0$ for $0 < t < 1$.

For the next proposition on the multivariate delta method we also require the notion of tightness of a collection of probability measures on $\mathbb{R}^J$, generalizing the one variable case in definition 8.16 of book 2.

**Definition 4.29** A sequence of probability measures $\{\mu_n\}$ on $\mathbb{R}^J$ will be said to be tight if for any $\epsilon > 0$ there is a bounded rectangle, $A = \prod_{i=1}^J (a_i, b_i]$, so that $\mu_n(A) > 1 - \epsilon$ for all $n$.

**Exercise 4.30** Generalize proposition 8.18 of book 2 to show that if $\{\mu_n\}_{n=1}^\infty$ is a sequence of probability measures on $\mathbb{R}^J$ with $\mu_n \Rightarrow \mu$ for a probability measure $\mu$, then $\{\mu_n\}$ is tight. Hint: First prove that given $A = \prod_{i=1}^J (a_i, b_i]$ and $A_M = \prod_{i=1}^J (a_i-M, b_i+M]$, that for any probability measure $\mu[A_M] \to 1$ as $M \to \infty$, recalling that measures are continuous from below. Now complete the steps of the proposition 8.18 proof.

**Exercise 4.31** Generalize proposition 8.39 of book 2. Prove that if $\{X_n, X\}$ are random vectors defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$ with range in $\mathbb{R}^J$, and there exists a positive sequence $\{c_n\} \subset \mathbb{R}$ with $c_n \to \infty$ as $n \to \infty$, and a constant vector $x_0 \in \mathbb{R}^J$ so that $c_n (X_n - x_0) \Rightarrow_d X$, then $X_n \Rightarrow_p x_0$. Hint: Recalling the notation of remark 4.20 above, $X_n \Rightarrow_p x_0$ means that for any $\epsilon > 0$:

$$\lim_{n \to \infty} \lambda(\|X_n - x_0\|_2 \geq \epsilon) = 0.$$

But by 4.11 there exists positive $c_{2,\infty}$ and $C_{2,\infty}$ so that

$$c_{2,\infty} \|X_n - x_0\|_\infty \leq \|X_n - x_0\|_2 \leq C_{2,\infty} \|X_n - x_0\|_\infty,$$

and thus

$$\{\|X_n - x_0\|_2 \geq \epsilon\} \subset \{\|X_n - x_0\|_\infty \geq \epsilon/C_{2,\infty}\}.$$  

Now note that the set on the right is a rectangle centered on $x_0$ and use the tightness result of exercise 4.30.

The multivariate delta method is summarized next in the context of proposition 8.40 of book 2. It is left as an exercise for the reader to generalize this as above to the case where $\partial^\alpha g(x_0) = 0$ for all $|\alpha| \leq m$. 

Proposition 4.32 (The Multivariate $\Delta$-Method) Let $\{X_n, X\}$ be random vectors defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$, so $X_n : \mathcal{S} \to \mathbb{R}^j$ and $X : \mathcal{S} \to \mathbb{R}^j$, and assume that there exists a positive sequence $(c_n) \subset \mathbb{R}$ with $c_n \to \infty$ as $n \to \infty$, and a constant vector $x_0 \in \mathbb{R}^j$ so that $c_n (X_n - x_0) \to_d X$. If $g : \mathbb{R}^j \to \mathbb{R}$ is a continuously differentiable function, then:

$$c_n [g(X_n) - g(x_0)] \to_d \nabla g(x_0) \cdot X. \quad (4.20)$$

where $\nabla g(x_0) \equiv \left( \frac{\partial g(x_0)}{\partial x_1}, \ldots, \frac{\partial g(x_0)}{\partial x_j} \right)$ is the gradient of $g$ at $x_0$, and $x \cdot y \equiv \sum_{i=1}^j x_i y_i$ is the dot product of $x$ and $y$.

**Proof.** By the above discussion:

$$c_n [g(X_n) - g(x_0)] = \nabla g(X'_n) \cdot c_n (X_n - x_0), \quad (*)$$

where $X'_n = tX_n + (1-t)x_0$ for $0 < t < 1$. Now $c_n (X_n - x_0) \to_d X$ and $c_n \to \infty$ assure that $X_n \to_P x_0$ by exercise 4.31. This again assures that $X'_n \to_P x_0$ since $X'_n - x_0 = t(X_n - x_0)$ for $0 < t < 1$ obtains that $|X'_n - x_0| < |X_n - x_0|$ and hence:

$$\lambda (\{ |X'_n - x_0| \geq \epsilon \}) \leq \lambda (\{ |X_n - x_0| \geq \epsilon \}).$$

Then by the Mann-Wald theorem of proposition 4.21 with $h(x) \equiv \nabla g(x) : \nabla g(X'_n) \to_P \nabla g(x_0)$.

Since $c_n (X_n - x_0) \to_d X$ by assumption, we can obtain 4.20 by applying Slutsky’s theorem of proposition 4.25 to the dot product in (*) by defining $h(U_n, V_n) \equiv U_n \cdot V_n$ on $\mathbb{R}^{2j}$. 

### 4.2.3 Scheffé’s Theorem

Scheffé’s theorem is named for Henry Scheffé (1907 – 1977) who investigated the implication of almost everywhere convergence of probability density functions: $f_n \to f$, $m^j$-a.e., where $m^j$ denotes Lebesgue measure on $\mathbb{R}^j$. Perhaps not surprisingly this implies that $\mu_n \Rightarrow \mu$ for the associated probability measures. But in fact the almost everywhere convergence of density functions assures something stronger than weak convergence, and thus this result is not listed among the results of the prior section. It is also the case that the proof of Scheffé’s theorem does not require the tools of the portmanteau theorem, only the integration theory of book 5. And with this theory, it is also possible to prove Scheffé’s result for density functions defined relative to Borel measures.
Generalizing somewhat the terminology of definition 1.8:

**Definition 4.33** Given a Borel measure space \((\mathbb{R}^j, \mathcal{B}(\mathbb{R}^j), \nu)\), a Borel measurable function \(f : \mathbb{R}^j \rightarrow \mathbb{R}\) is a **density function with respect to \(\nu\)** if it is nonnegative, and \(\int f(x) \, d\nu = 1\).

As noted in proposition 3.3 of book 5, given such a density function \(f\), the associated set function \(\mu\) defined on \(A \in \mathcal{B}(\mathbb{R}^j)\) by

\[
\mu(A) \equiv \int_A f(x) \, d\nu, \tag{4.21}
\]

is in fact a measure on \(\mathcal{B}(\mathbb{R}^j)\). Thus every density function with respect to \(\nu\) on \(\mathbb{R}^j\) induces a Borel measure \(\mu\) on \(\mathbb{R}^j\).

**Example 4.34** Recalling chapter 4 of book 5, every \(g \in L_1(\mathbb{R}^j, \mathcal{B}(\mathbb{R}^j), \nu)\) induces a density function \(f\) on \(L_1(\mathbb{R}^j, \mathcal{B}(\mathbb{R}^j), \nu)\) defined by \(f = |g|/\|g\|_1\), and an associated Borel measure \(\mu\) defined in 4.21.

Given a density \(f\) with respect to \(\nu\), let \(\mu\) be defined in 4.21 with \(A = A(x_1, \ldots, x_j)\). Then:

\[
F_\mu(x_1, \ldots, x_j) \equiv \mu(A(x_1, \ldots, x_j))
\]

is a **distribution function** on \(\mathbb{R}^j\) by proposition 8.10 of book 1. But note that in general \(f\) is **not a density function associated with** \(F_\mu\) as given in definition 1.8. However in the special case where \(\nu\) is absolutely continuous with respect to Lebesgue measure \(m^j\), so \(\nu \ll m^j\) in the notation of definition 7.3 of book 5, then a density function associated with \(F_\mu\) can be identified in terms of \(f\).

To this end recall that if \(\nu \ll m^j\), the Radon-Nikodym theorem of proposition 7.22 of book 5 identifies a Borel measurable function \(g_\nu(x)\) so that for all \(A \in \mathcal{B}(\mathbb{R}^j)\):

\[
\nu(A) = \int_A g_\nu(x) \, dm^j.
\]

Thus by proposition 3.6 of book 5:

\[
\mu(A) \equiv \int_A f(x) \, d\nu = \int_A f(x) g_\nu(x) \, dm^j \tag{4.22}
\]

for all \(A \in \mathcal{B}(\mathbb{R}^j)\). Thus in this special case, \(f(x) g_\nu(x)\) is then a density function associated with the distribution function \(F_\mu\) as in definition 1.8.

In the general case \(F_\mu\) need not have a density function in this sense as noted in example 4.35.
Example 4.35 Let \( \{y_i\}_{i=1}^N \subset \mathbb{R}^d \) where if \( N = \infty \) we assume that this collection has no cluster points, and let \( \{c_i\}_{i=1}^N \subset \mathbb{R}^+ \) with \( \sum_{i=1}^N c_i = 1 \). Define
\[
\nu(A) = \sum_{y_i \in A} c_i,
\]
then \( (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \nu) \) is a Borel measure space. If \( f \) is a density with respect to \( \nu \), an exercise in the definition of integral obtains:
\[
\int f(x) d\nu = \sum_{i=1}^N c_i f(y_i) = 1.
\]
The measure \( \mu \) induced by \( f \) in 4.21 is then defined:
\[
\mu(A) = \sum_{y_i \in A} c_i f(y_i),
\]
and the induced distribution function \( F_\mu(x_1, ..., x_j) \) is defined as above.

It is not the case that \( \nu \ll m^j \) since \( \nu(y_i) = c_i \) and \( m^j(y_i) = 0 \). In fact, since \( \nu \perp m^j \) in the notation of definition 7.3 of book 5, meaning that these measures are mutually singular. In this case, \( F_\mu \) has no density function in the sense of definition 1.8 by proposition 1.3 of book 4, as \( F_\mu \) is a saltus function.

Proposition 4.36 (Scheffé’s Theorem) Let \( \{f_n\}, f \) be density functions with respect to \( \nu \) on \( (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \nu) \) with \( f_n \rightarrow f, \nu\text{-a.e.} \). Then:
\[
\int |f_n - f| d\nu \rightarrow 0. \tag{4.23}
\]
Further, if \( \{\mu_n\}, \mu \) denote the associated Borel measures given in 4.21, then:
1. \( \mu_n \Rightarrow \mu \),
2. \( \mu_n(A) \rightarrow \mu(A) \) for all \( A \in \mathcal{B}(\mathbb{R}^d) \),
3. \( \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\mu_n(A) - \mu(A)| \rightarrow 0 \).

Remark 4.37 The condition in 2 is the definition of strong convergence of measures, while that in 3 is the definition of convergence of measures in total variation.

Proof. Certainly 3 \( \Rightarrow \) 2, while 2 \( \Rightarrow \) 1 by 4 of the portmanteau theorem, and so we proceed with two steps.
4.2 Measures on $\mathbb{R}^J$

1. To prove 4.23, note that

$$|f_n - f| = f_n - f + 2 \max (f - f_n, 0).$$

Since $f_n$ and $f$ are density functions with respect to $\nu$:

$$\int |f_n - f| \, d\nu = 2 \int \max (f - f_n, 0) \, d\nu.$$

Now $\max (f - f_n, 0) \to 0$ pointwise $\nu$-a.e., and $\max (f - f_n, 0) \leq f$. Thus by Lebesgue’s dominated convergence theorem of proposition 2.43 of book 5, 4.23 is proved.

2. To prove 3:

$$\sup_{A \in B(\mathbb{R}^j)} |\mu_n(A) - \mu(A)| = \sup_{A \in B(\mathbb{R}^j)} \left| \int_A (f_n - f) \, d\nu \right|$$

$$\leq \sup_{A \in B(\mathbb{R}^j)} \int_A |f_n - f| \, d\nu$$

$$\leq \int |f_n - f| \, d\nu.$$

Thus 4.23 obtains 3.

\[\Box\]

**Corollary 4.38** With the assumptions above:

$$\sup_{A \in B(\mathbb{R}^j)} |\mu_n(A) - \mu(A)| = \frac{1}{2} \int |f_n - f| \, d\nu.$$  \hspace{1cm} (4.24)

**Proof.** Let $g_n \equiv f_n - f$, and recalling definition 2.36 of book 5 define $g_n^+ \equiv \max (g_n, 0)$ and $g_n^- \equiv \max (-g_n, 0)$. Then $g_n = g_n^+ - g_n^-$ and since $\int g_n \, d\nu = 0$ it follows that $\int g_n^+ \, d\nu = \int g_n^- \, d\nu$. Now:

$$\sup_{A \in B(\mathbb{R}^j)} (\mu_n(A) - \mu(A)) = \sup_{A \in B(\mathbb{R}^j)} \int_A g_n \, d\nu$$

$$= \int_{f_n \leq f} g_n \, d\nu$$

$$= \int g_n^+ \, d\nu.$$

Similarly,

$$\sup_{A \in B(\mathbb{R}^j)} (\mu(A) - \mu_n(A)) = \int_{f \leq f_n} (-g_n) \, d\nu = \int g_n^- \, d\nu.$$
Since both supremums are nonnegative, these equal the respective absolute values and addition obtains:

\[ 2 \sup_{A \in B(\mathbb{R}^d)} |\mu_n(A) - \mu(A)| = \int g_n^+ dv + \int g_n^- dv. \]

But |g_n| = g_n^+ + g_n^- and thus:

\[ 2 \sup_{A \in B(\mathbb{R}^d)} |\mu_n(A) - \mu(A)| = \int |f_n - f| dv. \]

\[ \blacksquare \]

**Remark 4.39** Scheffé published his results in 1947, but it turns out that the density function convergence conclusion in 4.23 is a special case of a 1928 result by Frigyes Riesz (1880 – 1956) and sometimes known as Riesz’s lemma, although there are many results associated with this name (see for example propositions 4.16 and 4.22 of book 5). Riesz’s result is as follows, recalling the notion of the Banach space \(L_p(X, \sigma(X), \lambda)\) from chapter 4 of book 5.

**Riesz’s lemma:** If \(f_n, f \in L_p(X, \sigma(X), \lambda)\) for \(p \geq 1\) with \(f_n \to f\), \(\lambda\)-a.e. and \(\|f_n\|_p \to \|f\|_p\), then \(\|f_n - f\|_p \to 0\).

As density functions are elements of \(L_1(X, \sigma(X), \lambda)\) and all have \(\|f\|_1 = 1\), Riesz’s lemma states that \(f_n \to f\), \(\lambda\)-a.e. assures \(\|f_n - f\|_1 \to 0\) which is 4.23.

**Example 4.40** 1. **Student \(T\) ⇒ Standard Normal:** Recall the density function of the Student \(T\) distribution or Student’s \(T\) distribution with \(\nu > 0\) degrees of freedom introduced in remark 1.24 of book 4:

\[ f_T(t) = \frac{\Gamma((\nu + 1)/2)}{\sqrt{\pi \nu} \Gamma(\nu/2)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \quad (4.25) \]

where the gamma function \(\Gamma(\alpha)\) is defined by in 2.25. It is named for William Sealy Gosset (1876 – 1937) who published under the pen name of Student. The coefficient of this function is also expressed in terms of the beta function \(B(v, w)\) defined by:

\[ B(v, w) = \int_0^1 y^{v-1}(1 - y)^{w-1} dy. \quad (4.26) \]

First \(\Gamma(1/2) = \sqrt{\pi}\) as is verified by a substitution of \(u = \sqrt{2\pi} \) in 2.25 to obtain an integral related to the normal density, and then \(B(v, w) = \)
4.2 MEASURES ON $\mathbb{R}^J$

\[ f_T(t) = \frac{1}{B(\nu/2, 1/2)/\nu} \left( 1 + \frac{t^2}{\nu} \right)^{-\nu/2}. \]

Now $(1 + a/n)^{-n} \to e^{-a}$ as $n \to \infty$, so letting $\nu \to \infty$ obtains:

\[ \left( 1 + \frac{t^2}{\nu} \right)^{-\nu/2} \to e^{-t^2/2}. \]

Using identities for the gamma function obtains:

\[ \frac{\Gamma((\nu + 1)/2)}{\sqrt{\pi \nu \Gamma(\nu/2)}} \to \frac{1}{\sqrt{2\pi}}. \quad (\ast) \]

For example, using the identity $\Gamma(x+1) = x\Gamma(x)$ for general $x \geq 0$ and $\Gamma(1) = 1$ obtains $\Gamma(n) = (n-1)!$ for integer $n$, and this plus $\Gamma(1/2) = \sqrt{\pi}$ allows a verification of $(\ast)$ for $\nu$ restricted to odd integers, $\nu = 2m + 1$, or even integers, $\nu = 2m$.

Thus the Student $T$ density converges pointwise to the standard normal density in 3.2 as $\nu \to \infty$. By Scheffé’s theorem, the associated probability measures also converge in the various modes identified in proposition 4.36.

2. **Binomial $\Rightarrow$ Poisson:** The density function of the **binomial distribution** with parameters $0 < p < 1$ and $n \in \mathbb{N}$ is given:

\[ f_{B_n}(j) = \binom{n}{j} p^j (1-p)^{n-j}, \quad j = 0, 1, \ldots, n. \quad (4.27) \]

The density function of the **Poisson distribution**, named for Siméon-Denis Poisson (1781 – 1840), has parameter $\lambda > 0$ and is given:

\[ f_P(j) = e^{-\lambda} \lambda^j / j!, \quad j = 0, 1, 2, \ldots, \quad (4.28) \]

Fix $\lambda$ and consider the binomial density with parameters $p \equiv \lambda/n$ and $n$:

\[ f_{B_n}(j) = \binom{n}{j} \left( \frac{\lambda}{n} \right)^j \left( 1 - \frac{\lambda}{n} \right)^{n-j} \]

\[ = \frac{\lambda^j}{j!} \left( 1 - \frac{\lambda}{n} \right)^n \left( 1 - \frac{\lambda}{n} \right)^{-j} \prod_{k=0}^{j-1} \left( 1 + k/n \right). \]
As $n \to \infty$ we have $(1 - \frac{\lambda}{n})^n \to e^{-\lambda}$, while the third and fourth terms converge to 1. Thus the binomial density with $p \equiv \lambda/n$ converges pointwise to the Poisson density in 3.2 as $n \to \infty$. By Scheffé’s theorem, the associated probability measures also converge in the various modes identified in proposition 4.36.

4.2.4 Prokhorov’s theorem

The final investigation for this chapter is into the generalization of the results of section 8.2 of book 2 which collectively provided a special case of Prokhorov’s theorem, named for Yuri Vasilyevich Prokhorov (1929 – 2013) and summarized in that book’s proposition 8.24. For its statement, we recall the notion of tightness of a collection of measures introduced in definition 4.29 and the conclusion of exercise 4.30.

Proposition 4.41 (Prokhorov’s theorem) Given a tight sequence of probability measures $\{\mu_n\}_{n=1}^{\infty}$ on $\mathbb{R}^j$, there exists a subsequence $\{\mu_{n_k}\}_{k=1}^{\infty}$, and a probability measure $\mu$, so that $\mu_{n_k} \Rightarrow \mu$.

**Proof.** Let $\{F_n\}_{n=1}^{\infty}$ denote the associated sequence of distribution functions, so $F_n(x) \equiv \mu_n \left[ \bigcap_{i=1}^{j} (-\infty, x_i) \right]$. The first step is in essence the $j$-dimensional version of Helly's selection theorem of proposition 8.14 of book 2, named for Eduard Helly (1884 – 1943). The goal of this step is to identify a subsequence $\{F_{n_k}\}_{k=1}^{\infty}$ and candidate increasing function $F$.

The larger step below will be to prove that due to tightness, this function $F$ is a distribution function and that $F_{n_k}(x) \to F(x)$ for all continuity points of $F$. For notation, let $x \leq y$ or $x < y$ be interpreted as component-wise statements, so that $x_i \leq y_i$ or $x_i < y_i$ for all $i$, respectively.

Let $\{r_i\}_{i=1}^{\infty} \subset \mathbb{R}^j$ denote an enumeration of all points with rational components, a countable collection. Since $\{F_n(r_1)\}_{n=1}^{\infty}$ is a sequence bounded by 1, there is an accumulation point which we denote $A(r_1)$, and a subsequence $\{n_{1,k}\}_{k=1}^{\infty}$ so that $F_{n_{1,k}}(r_1) \to A(r_1)$. Next, since $\{F_{n_{1,k}}(r_2)\}_{k=1}^{\infty}$ is a bounded sequence, there is an accumulation point $A(r_2)$, and a subsequence $\{n_{2,k}\}_{k=1}^{\infty} \subset \{n_{1,k}\}_{k=1}^{\infty}$ so that $F_{n_{2,k}}(r_2) \to A(r_2)$. Continuing in this way, define sequences $\{n_{i+1,k}\}_{k=1}^{\infty} \subset \{n_{i,k}\}_{k=1}^{\infty}$ with $F_{n_{i+1,k}}(r_{i+1}) \to A(r_{i+1})$. Now define $n_k \equiv n_{k,k}$. By construction, $F_{n_k}(r_i) \to A(r_i)$ for all rational points in $\mathbb{R}^j$, and $0 \leq A(r_i) \leq 1$ for all $r_i$. Also, given $r_m$, $r_n$ with $r_m \leq r_n$, if $k > \max(n, m)$ then $F_{n_k}(r_m) \leq F_{n_k}(r_n)$ since each $F_{n_k}$ is increasing.

Hence:

$$A(r_m) = \lim_k F_{n_k}(r_m) \leq \lim_k F_{n_k}(r_n) = A(r_n).$$
Now define $F(x) = \inf_{r > x} A(r)$. Then $0 \leq F(x) \leq 1$ for all $x$, and $F(x)$ is an increasing function in the same way that $A(r_i)$ is increasing. Specifically, if $x \leq y$ then $F(x) \leq F(y)$.

To prove that $F$ is a distribution function we first prove that $F$ is continuous from above in the sense of 1.4, and satisfies the $n$-increasing condition in 1.5.

1. For continuity from above, let $x$ and $\epsilon$ be given. By definition of infimum there is an $r_i > x$ so that $A(r_i) < F(x) + \epsilon$. Also, for any $y$ with $x \leq y < r_i$, it follows by the above observations that $F(x) \leq F(y) \leq A(r_i) < F(x) + \epsilon$, and so if $y \to x+$, then

$$F(x) \leq \lim_{y \to x+} F(y) < F(x) + \epsilon.$$ 

As $\epsilon > 0$ is arbitrary, $F$ is continuous from above.

2. For the $n$-increasing condition, let $B = \prod_{i=1}^J (a_i, b_i]$ be a bounded rectangle and $\epsilon > 0$ be given. By continuity from above just proved, there exists $c = (\delta, \delta, ..., \delta)$ so that if $x$ denotes any one of the $2^j$ vertices of $B$, then $|A(r) - F(x)| < \epsilon/2^j$ for rational $r$ with $x < r < x + c$. Choosing rational points $q$ and $r$ with $a < q < a + c$ and $b < r < b + c$, define $B' = \prod_{i=1}^J (q_i, r_i]$, and denote by $s$ any one of the $2^j$ rational vertices of $B'$. Since $F_{n_k}(r) \to A(r)$ for all rational points in $\mathbb{R}^j$, and each $F_{n_k}$ satisfies 1.5:

$$\sum_s sgn(s)A(s) = \lim_{k \to \infty} \sum_s sgn(s)F_{n_k}(s) \geq 0.$$ 

Also, pairing the respective vertices of $B$ and $B'$, let $s_x$ denote the rational vertex of $B'$ associated with $B$-vertex $x$, then:

$$\left|\sum_s sgn(s)A(s) - \sum_x sgn(x)F(x)\right| \leq \sum_x |A(s_x) - F(x)| < \epsilon,$$

since this is a summation of $2^j$ terms and by above, $|A(r) - F(x)| < \epsilon/2^j$ for rational $r$ with $x < r < x + c$. Combining estimates it follows that $\sum_x sgn(x)F(x) \geq 0$ since $\epsilon$ was arbitrary.

Because $F$ is continuous from above and satisfies the $n$-increasing condition, there exists a Borel measure $\mu$ associated with $F$ by proposition 8.14 of book 1. We next prove that $\mu$ is a probability measure, and for this the tightness of $\{\mu_n\}_{n=1}^\infty$ is essential. By tightness, for any $\epsilon > 0$ there exists $T$ so that with apparent notation, the rectangle $(-T, T)^j \subset \mathbb{R}^j$ satisfies $\mu_n \left[(-T, T)^j \right] > 1 - \epsilon$ for all $n$. Choose $x$ with $x_i > T$ for all $i$
and then note that for any rational \( r > x \) that \( F_n(r) > 1 - \epsilon \) for all \( n \) since \( F_n(r) \equiv \mu_n \left[ \prod_{i=1}^{j} (-\infty, r_i] \right] \geq \mu_n \left[ (-T, T]^j \right] \). But then \( A(r) = \lim_{n \to \infty} F_n(r) \geq 1 - \epsilon \) for all \( r > x \) and thus \( F(x) \geq 1 - \epsilon \) for \( x \) with all \( x_i > T \). On the other hand, choose \( x \) with \( x_i < -T \) for some \( i \) and then note that for any rational \( r < x \) that \( F_n(r) < \epsilon \) for all \( n \) since:
\[
F_n(r) \equiv \mu_n \left[ \prod_{i=1}^{j} (-\infty, r_i] \right] \leq \mu_n \left[ \mathbb{R}^j - (-T, T]^j \right] \leq \epsilon.
\]
As before, this implies that \( A(r) \leq \epsilon \) for all \( r < x \) and thus \( F(x) \leq \epsilon \) for \( x \) with \( x_i < -T \) for some \( i \). Since \( \epsilon \) is arbitrary, this proves that the range of \( F(x) \) includes \((0,1)\).

Now let \( x \) be given and define a rectangle for \( s \) large by \( B_x(s) = \prod_{i=1}^{j} (-s, x_i] \). Then by the proposition 8.11 construction for \( \mu \) in book 1:
\[
\mu [B_x(s)] = \sum_{y} \text{sgn}(y) F(y),
\]
where the \( y \)-summation is over the \( 2^j \) vertices of \( B_x(s) \). If \( y \neq x \) and \( s \) is large, it follows that \( y_i < -T \) for some \( i \) and thus by the previous result \( F(y) \leq \epsilon \). Hence for \( s \) large,
\[
\mu [B_x(s)] - F(x) = \sum_{y \neq x} \text{sgn}(y) F(y),
\]
which obtains
\[
|F(x) - \mu [B_x(s)]| \leq 2^{j-1} \epsilon.
\]
Since \( \mu [B_x(s)] \to \mu \left[ \prod_{i=1}^{j} (-\infty, x_i] \right] \) as \( s \to \infty \) by continuity from below of \( \mu \), and \( \epsilon > 0 \) is arbitrary, it follows that \( F(x) = \mu \left[ \prod_{i=1}^{j} (-\infty, x_i] \right] \). Also, as \( x \to \infty \) the above analysis proves that \( F(x) \to 1 \) and thus \( F \) is a distribution function.

The final step is to prove that \( \mu_{n_k} \Rightarrow \mu \), which is to prove that \( F_{n_k}(x) \to F(x) \) for all continuity points of \( F \). Given such a point \( x \) and \( \epsilon > 0 \), choose rational \( r > x \) so that \( A(r) < F(x) + \epsilon \) as above. By assumed continuity at \( x \) there is \( y < x \) so that \( F(y) - \epsilon < F(y) \), and choosing rational \( q \) with \( y < q < x \), we have that \( F(y) \leq A(q) \). Now since \( A(q) \leq A(r) \) as above:
\[
F(x) - \epsilon < A(q) \leq A(r) < F(x) + \epsilon.
\]
By construction \( F_{n_k}(r_i) \to A(r_i) \) for all rational \( r_i \), and this obtains:
\[
F(x) - \epsilon \leq \lim F_{n_k}(q) \leq \lim F_{n_k}(r) < F(x) + \epsilon.
\]
But since \( F_{n_k}(q) \leq F_{n_k}(x) \leq F_{n_k}(r) \) for all \( n_k \), this implies that
\[
F(x) - \epsilon \leq \liminf_k F_{n_k}(x) \leq \limsup_k F_{n_k}(x) \leq F(x) + \epsilon.
\]
As \( \epsilon > 0 \) is arbitrary, we conclude that \( \lim_k F_{n_k}(x) = F(x) \). □

**Corollary 4.42 (Prokhorov’s theorem)** Given a tight sequence of probability measures \( \{\mu_n\}_{n=1}^\infty \) on \( \mathbb{R}^J \), if every subsequence \( \{\mu_{n_k}\}_{k=1}^\infty \) that converges weakly, converges to a given probability measure \( \mu \), then \( \mu_n \Rightarrow \mu \).

**Remark 4.43** Note that under the hypothesis that \( \{\mu_n\}_{n=1}^\infty \) is tight, any such \( \mu \) is of necessity a probability measure, so this is not really an assumption. Specifically, proposition 4.41 provides a specified convergent subsequence that converges weakly to some probability measure which we denote here \( \mu' \), and thus by the hypothesis of this corollary it is the case that \( \mu = \mu' \).

**Proof.** Assume that for the given \( \mu \) that \( \mu_n \not\Rightarrow \mu \), and thus there is a continuity point of the associated distribution function \( F \), say \( x \), so that \( F_n(x) \nrightarrow F(x) \), where \( F_n \) is the distribution function associated with \( \mu_n \). This then implies that for any \( \epsilon > 0 \) there is a subsequence \( \{n_k\}_{k=1}^\infty \) so that \( |F_{n_k}(x) - F(x)| \geq \epsilon \) for all \( k \). But \( \{\mu_{n_k}\} \) are tight, and thus by the above proposition there is a subsequence of \( \{\mu_{n_k}\}_{k=1}^\infty \) which converges weakly, and hence by hypothesis must converge to \( \mu \). But this is a contradiction, since for the \( F \)-continuity point \( x \), \( |F_{n_k}(x) - F(x)| \geq \epsilon \) for all \( k \). Hence \( \mu_n \Rightarrow \mu \). □
Chapter 5

Expectations of Random Variables 2

Chapter 3 of book 4 introduced the notion of the expectation of a random variable $X$ defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$, and more generally the expectation of $g(x)$ where $g$ is a Borel measurable real-valued function. That introduction provided a link between the intuitive notion of expectation usually given in introductory probability texts, and the more formal framework that was promised to be developed in this book using the tools of book 5. In that introduction was noted that the informal approach to expectation is only readily applicable to random variables with density functions, either Lebesgue integrable or discrete, and not to the more general cases of distribution functions characterized in section 1.1 of book 4. And even in these special cases of random variables with density functions, it was seen in section 3.1.1 of that book that these definitions were not obviously consistent, and required some additional analysis to verify that they were. In this section we apply the tools developed in book 5 to put the theory of expectations on the firm foundation outlined in section 3.1.2 of book 4.

5.1 General Definition of $E[g(X)]$

We begin by formalizing the definition given in the above noted section 3.1.2 of book 4.

**Definition 5.1** If $X : \mathcal{S} \rightarrow \mathbb{R}$ is a random variable defined on a probability
space \((\mathcal{S}, \mathcal{E}, \lambda)\), the **expectation** of \(X\), denoted \(E[X]\), is defined:

\[
E[X] = \int_{\mathcal{S}} X(s) d\lambda,
\]

where this integral is defined in chapter 2 of book 5.

More generally, if \(g : \mathbb{R} \to \mathbb{R}\) is a Borel measurable function, then \(Y = g(X)\) is a random variable on \((\mathcal{S}, \mathcal{E}, \lambda)\) and the **expectation** of \(g(X)\), denoted \(E[g(X)]\), is defined:

\[
E[g(X)] = \int_{\mathcal{S}} g(X(s)) d\lambda.
\]

If \(X : \mathcal{S} \to \mathbb{R}^j\) is a random vector, \(E[X]\) is defined componentwise:

\[
E[X] = (E[X_1], ..., E[X_j]).
\]

Similarly, if \(g : \mathbb{R}^j \to \mathbb{R}^k\) then \(E[g(X)]\) is defined by 5.1 if \(k = 1\), and by 5.3 if \(k \geq 2\).

**Remark 5.2** Consistent with the general integration theory of book 5, these integrals are defined only when the respective integrands are absolutely integrable. For example, \(E[g(X)]\) is defined as above only in the case:

\[
\int_{\mathcal{S}} |g(X(s))| d\lambda < \infty.
\]

Although these integrals are well defined by the development in chapter 2 of book 5, in order to evaluate \(E[g(X)]\) for given \(g\) and \(X\) it is usually necessary to transform this \(\lambda\)-integral to an integral on \(\mathbb{R}\) which can then be evaluated directly. This first step is typically accomplished with a Lebesgue-Stieltjes integral, and in the special case of absolutely continuous distribution functions, this Lebesgue-Stieltjes integral can then be transformed into a Lebesgue integral. We proceed in two steps. But to simplify notation we develop the case where \(X : \mathcal{S} \to \mathbb{R}\), noting that the referenced change of variables results from chapter 3 of book 5 are quite general, and thus analogous transformations are valid in the more general case where \(X : \mathcal{S} \to \mathbb{R}^j\), etc. Details are left as an exercise.

1. **Transformation to a Lebesgue-Stieltjes Integral on \(\mathbb{R}\):**

The random variable \(X : \mathcal{S} \to \mathbb{R}\) is a measurable transformation between measure spaces \((\mathcal{S}, \mathcal{E}, \lambda)\) and \((\mathbb{R}, \mathcal{B}(\mathbb{R}), m)\) by definition 3.9 of
book 5, there denoted $T$. This transformation induces a Borel measure
\( \lambda_X \) on \( \mathbb{R} \) by this definition and recalled in 4.5 above. Specifically, for
\( A \in \mathcal{B}(\mathbb{R}) \):
\[
\lambda_X(A) \equiv \lambda \left[ X^{-1}(A) \right].
\]
Applying the change of variables result in proposition 3.14 of book 5:
\[
\int_S g(X(s))d\lambda = \int_{\mathbb{R}} g(x)d\lambda_X,
\]
where the integral on the right is a Lebesgue-Stieltjes integral. If
\( A = (-\infty, x] \) it follows that \( \lambda_X(A) = F(x) \) where \( F \) is the distribution
function of \( X \), and so \( \lambda_X \) is the Borel measure on \( \mathbb{R} \) induced by \( F \).
That is, \( d\lambda_X = d\mu_F \), and this expectation can also be expressed:
\[
E[g(X)] = \int_{\mathbb{R}} g(x)d\mu_F. \tag{5.5}
\]
Thus, in the most general case, \( E[g(X)] \) can be expressed as a Lebesgue-
Stieltjes integral of \( g(x) \) on \( \mathbb{R} \).

By proposition 2.56 of book 5, it then follows that for continuous \( g \),
\( E[g(X)] \) can also be expressed as a Riemann-Stieltjes integral of
book 3:
\[
E[g(X)] = \int_{\mathbb{R}} g(x)dF. \tag{5.6}
\]

**Example 5.3** Consider the special case where \( X \) has a discrete distribu-
tion, so \( X(S) = \{ x_i \}_{i=1}^{\infty} \), and assume for simplicity that this collec-
tion has no accumulation points and is indexed in increasing order.
If this collection is unbounded positively and negatively, it can be par-
titioned into two monotone sequences and the following logic applied
to each separately. By 4.5 the distribution function is given by
\[
F(x) \equiv \lambda \left[ X^{-1}((-\infty, x]) \right] = \mu_F \left[ (-\infty, x] \right].
\]
Defining \( f(x) \), the density function associated with \( F(x) \), by:
\[
f(x) = F(x) - F(x^-),
\]
it follows that
\[
F(x) = \sum_{x_i \leq x} f(x_i).
\]
The standard expectations formula for discrete distributions is usually presented:

\[ E[g(X)] = \sum_{i=1}^{\infty} g(x_i) f(x_i), \]  

(5.7)

provided that as in 5.4:

\[ \sum_{i=1}^{\infty} |g(x_i)| f(x_i) < \infty. \]

The formula for \( E[g(X)] \) in 5.7 follows from definitions 2.11 and 2.37 of book 5 applied to 5.5. This can be summarized as:

\[ \int_{\mathbb{R}} g(x)d\mu_F \equiv \sup_{\varphi \leq g^+} \int_{\mathbb{R}} \varphi(x)d\mu_F - \sup_{\psi \leq g^-} \int_{\mathbb{R}} \psi(x)d\mu_F, \]

where \( \varphi \) and \( \psi \) are simple functions, and \( g^+ \) and \( g^- \) the positive and negative parts of \( g \). To evaluate this integral, assume now that \( g \) is nonnegative, since the following can be applied to each of \( g^+ \) and \( g^- \) separately.

If \( A \subset (x_i, x_{i+1}) \), then \( \int_{\mathbb{R}} \chi_A(x)d\mu_F = 0 \) where \( \chi_A(x) \) denotes the characteristic function of \( A \), defined as 1 for \( x \in A \) and 0 otherwise. Now let \( A_i \equiv \{x_i\}, a_i = \inf_{A_i} g(x) = g(x_i), \) and define \( \varphi_n \) in terms of \( \{a_i\}_{i=1}^{n}, \{A_i\}_{i=1}^{n} : \)

\[ \varphi_n(x) = \sum_{i=1}^{n} a_i \chi_{A_i}(x). \]

Then by 2.1 of book 5:

\[ \int_{\mathbb{R}} \varphi_n(x)d\mu_F = \sum_{i=1}^{n} g(x_i)f(x_i), \]

since by construction:

\[ \mu_F [A_i] = f(x_i). \]

It now follows that for such \( 0 \leq \varphi_n \leq g \), that:

\[ \int_{\mathbb{R}} \varphi_n(x)d\mu_F \leq \int_{\mathbb{R}} g(x)d\mu_F. \]

Letting \( n \to \infty \), Lebesgue’s dominated convergence theorem obtains that:

\[ \sup_{n} \int_{\mathbb{R}} \varphi_n(x)d\mu_F = \sum_{i=1}^{\infty} g(x_i)f(x_i). \]

As \( E[g(X)] \) is defined as a supremum over all simple functions, it then follows that:

\[ E[g(X)] \geq \sum_{i=1}^{\infty} g(x_i)f(x_i). \]
Finally, given an arbitrary simple function, \( \varphi(x) = \sum_{j=1}^{n} b_j \chi_{B_j}(x) \) with disjoint \( \{B_i\}_{i=1}^{n} \) and \( \varphi \leq g \):

\[
\int_{\mathbb{R}} \varphi(x) d\mu_F = \sum_{j=1}^{n} b_j \sum_{i=1}^{\infty} \int_{\mathbb{R}} \chi_{A_i \cap B_j}(x) d\mu_F,
\]

where \( A_i = \{x_i\} \) as above. But then by the above observation for \( B \subset (x_i, x_{i+1}) \), it follows that \( \int_{\mathbb{R}} \chi_{A_i \cap B_j}(x) d\mu_F \) equals 0 or \( f(x_j) \), and since \( b_j \leq \inf_{B_j} g(x) \) and \( \{B_i\}_{i=1}^{n} \) are disjoint:

\[
\int_{\mathbb{R}} \varphi(x) d\mu_F \leq \sum_{i=1}^{\infty} g(x_i) f(x_i).
\]

This upper bound then applies to the supremum of such integrals, and the result in 5.7 follows.

2. Transformation to a Lebesgue Integral on \( \mathbb{R} \):

In the special case where \( F(x) \) is absolutely continuous, propositions 3.58 and 3.61 of book 3 obtain that \( F'(x) \) exists almost everywhere, and if \( f(x) \) is measurable and \( f(x) = F'(x) \) a.e. then

\[
F(x) = \int_{-\infty}^{x} f(y)dy.
\]

Such \( f \) is a density function associated with \( F \) by definition 1.8. By proposition 7.18 of book 5, \( F(x) \) is absolutely continuous if and only if the induced measure \( \mu_F \) is absolutely continuous relative to Lebesgue measure \( m \), denoted \( \mu_F \ll m \). In the multivariate case, the same conclusion on the existence of a density function for \( F \) is obtained when \( \mu_F \ll m^j \) by the Radon-Nikodým theorem of proposition 7.22 of book 5. Here \( m^j \) is Lebesgue measure on \( \mathbb{R}^j \). Then by proposition 3.3 of book 5, the set function \( \nu \) defined on \( A \in B(\mathbb{R}) \) by:

\[
\nu(A) = \int_{A} f(y)dy
\]

is a Borel measure. It then follows that \( \nu(A) = \mu_F(A) \) on the semi-algebra of right semi-closed intervals \( A \equiv \{(a, b]\} \), or more generally \( A \equiv \{[a, b]\} \), and proposition 6.14 of book 1 assures that \( \nu(A) = \mu_F(A) \) for all \( A \in B(\mathbb{R}) \). Finally by proposition 3.6 of book 5, \( E[g(X)] \) can be expressed as a Lebesgue integral:

\[
E[g(X)] = \int_{\mathbb{R}} g(x) d\mu_F = (\mathcal{L}) \int_{\mathbb{R}} g(x) f(x) dx, \quad (5.8)
\]
which is the standard formula for such expectations.

In the special case where \( g(x) \) and \( f(x) = F'(x) \) are continuous, then \( E[g(X)] \) can be equivalently expressed as a **Riemann integral**, and this result follows from 5.6 by proposition 4.30 of book 3:

\[
E[g(X)] \equiv \int_{\mathbb{R}} g(x) dF = (\mathcal{R}) \int_{\mathbb{R}} g(x) f(x) dx. \tag{5.9}
\]

**Summary 5.4** With the aid of an integration theory on a probability space \((S, \mathcal{E}, \lambda)\), it is possible to define \( E[g(X)] \) in a consistent way for arbitrary random variables \( X : S \to \mathbb{R} \) and Borel measurable \( g : \mathbb{R} \to \mathbb{R} \). The same conclusion follows for multivariate random vectors and Borel measurable functions or transformations. In the special cases where the distribution function for \( X \) is discrete or absolutely continuous, this definition can be translated to a calculation involving the probability density function of \( X \) as in 5.7 for the discrete case, or 5.8 or 5.9 in the absolutely continuous case.

For more general distribution functions, the value of \( E[g(X)] \) can be expressed as a Lebesgue-Stieltjes integral using the Borel measure \( \mu_F \) induced by \( F \).

### 5.2 Weak Convergence and Moment Limits

In this section we investigate an open question which continues the investigation underlying the method of moments of section 3.2.8 of book 4. If \( X_n, X : S \to \mathbb{R} \) are random variables defined on a probability space \((S, \mathcal{E}, \lambda)\) with associated distribution functions \( F_n(x) \) and \( F(x) \) such that \( F_n \) converges weakly to \( F \), notationally \( F_n \Rightarrow F \), is it true that \( E[g(X_n)] \to E[g(X)] \) for certain measurable functions \( g \)? Equivalently, using 5.5, is it true that:

\[
\int_{\mathbb{R}} g(x) d\mu_{F_n} \to \int_{\mathbb{R}} g(x) d\mu_F.
\]

In the case that \( g \) is a continuous and bounded real value function, the portmanteau theorem of proposition 4.4 above assures that \( E[g(X_n)] \to E[g(X)] \). For unbounded \( g \), we present two results related to the convergence of moments for which \( g(x) = x^m \).

For this investigation, recall the definition of **uniform integrability** introduced in definition 3.63 of book 4 for a related investigation, and again in definition 2.50 of book 5 for a result on **integration to the limit**.
5.2 WEAK CONVERGENCE AND MOMENT LIMITS

Definition 5.5 Given a probability space \((S, \mathcal{E}, \lambda)\), a sequence of random variables \(\{X_n\}_{n=1}^{\infty}\) is said to be uniformly integrable if:

\[
\lim_{N \to \infty} \sup_{n} \int_{|X_n| \geq N} |X_n(s)| \, d\lambda = 0. \tag{5.10}
\]

Proposition 5.6 Let \(X_n, X : S \to \mathbb{R}\) be random variables defined on a probability space \((S, \mathcal{E}, \lambda)\) with associated distribution functions \(F_n(x)\) and \(F(x)\) such that \(F_n \Rightarrow F\). If \(\{X_n^m\}_{n=1}^{\infty}\) are uniformly integrable for some integer \(m > 0\), then \(E[|X|^m] < \infty\) and

\[
E[X_n^m] \to E[X^m]. \tag{5.11}
\]

Proof. The initial construction is identical to that in the proof of proposition 3.66 of book 4. By Skorokhod’s representation theorem of proposition 8.30 of book 2, define random variables \(\{Y_n\}_{n=1}^{\infty}\) and \(Y\) on the Lebesgue measure space \((\mathbb{R}, \mathcal{B}(\mathbb{R}), m_L)\) with respective distribution functions \(\{F_n\}_{n=1}^{\infty}\) and \(F\), and for which \(Y_n \to Y\) for all \(t \in (0, 1)\). Then with \(m_L\) denoting Lebesgue measure, it is an exercise below to show that:

\[
\int_{S} |X_n(s)|^m \, d\lambda = \int_{(0,1)} |Y_n(t)|^m \, dm_L, \tag{(*)}
\]

and it thus follows that \(\{Y_n^m\}_{n=1}^{\infty}\) are also uniformly integrable. Choosing \(N\) large enough so that the supremum of the associated Lebesgue integrals is less than \(\delta\), it follows that for all \(n\):

\[
\int_{(0,1)} |Y_n(t)|^m \, dm_L \leq \int_{|Y_n| < N} |Y_n(t)|^m \, dm_L + \int_{|Y_n| \geq N} |Y_n(t)|^m \, dm_L \\
\leq N^m + \delta.
\]

Since \(Y_n \to Y\) for all \(t\), Fatou’s lemma in proposition 2.46 of book 3 obtains that \(E[|Y|^m] < N^m + \delta\), and thus by the equivalence of the Lebesgue and \(\lambda\)-integral this assures that \(E[|X|^m] < N^m + \delta\).

To prove convergence of expectations, first note that:

\[
\limsup_n \left| \int_{(0,1)} |Y_n(t)|^m \, dm_L - \int_{(0,1)} |Y(t)|^m \, dm_L \right| \\
\leq \limsup_n \left| \int_{|Y_n(t)| \leq N} |Y_n(t)|^m \, dm_L - \int_{|Y(t)| \leq N} |Y(t)|^m \, dm_L \right| \\
+ \sup_n \int_{|Y_n(t)| \geq N} |Y_n(t)|^m \, dm_L + \int_{|Y(t)| \geq N} |Y(t)|^m \, dm_L.
\]
Since $Y_n \to Y$ pointwise, the first limit superior on the right is 0 by the bounded convergence theorem of proposition 2.37 of book 3. By uniform integrability of $\{Y_n^m\}_{n=1}^{\infty}$, the supremum of integrals can be made as small as desired by choosing $N$ large, and similarly for the second integral since $E[|Y|^m] < \infty$. Hence

$$\limsup_n \left| \int_{(0,1)} [Y_n(t)]^m \, dm_L - \int_{(0,1)} [Y(t)]^m \, dm_L \right| = 0,$$

and the result follows by restating this expression in terms of $\lambda$-integrals using $(\ast)$. ■

**Exercise 5.7** Derive $(\ast)$ in the above proof. Hint: Use the change of variables results of the above section to convert both integrals to the associated Lebesgue-Stieltjes integrals.

The next proposition leads to the same conclusion but is often easier to apply in practice. As one example of an application, uniformly bounded variances assure convergence of the sequence of means. For a finance application of this result see the section below, Limiting Options Price Under the Risk Neutral Measure.

**Proposition 5.8** Let $X_n, X : S \to \mathbb{R}$ be random variables defined on a probability space $(S, \mathcal{E}, \lambda)$ with associated distribution functions $F_n(x)$ and $F(x)$ such that $F_n \Rightarrow F$. If for some integer $m > 0$ and $\epsilon > 0$:

$$\sup_n E \left[ |X_n|^m + \epsilon \right] < \infty,$$

then $E[|X|^m] < \infty$ and 5.11 holds.

**Proof.** Similar to the proof of Chebyshev’s inequality of proposition 3.32 of book 4:

$$\int_S |X_n(s)|^{m+\epsilon} \, d\lambda \geq \int_{|X_n| \geq N} |X_n(s)|^{m+\epsilon} \, d\lambda \geq N^\epsilon \int_{|X_n| \geq N} |X_n(s)|^m \, d\lambda,$$

and so as $N \to \infty$:

$$\sup_n \int_{|X_n| \geq N} |X_n(s)|^m \, d\lambda \leq \frac{1}{N^\epsilon} \sup_n E \left[ |X_n|^{m+\epsilon} \right] \to 0.$$

Hence the collection of random variables $\{X_n^m\}_{n=1}^{\infty}$ is uniformly integrable, and the prior proposition applies. ■
Example 5.9 For the above conclusion of convergence of moments, it is not difficult to exemplify why some assumption is needed related to uniform integrability. For example, with $\delta > 0$ define discrete density functions $\{f_n\}_{n=1}^{\infty}$ so that $f_n(n^{1+\delta}) = 1/n$, $f_n(0) = 1 - 1/n$. Then $F_n \Rightarrow F$ where $f(0) = 1$, but $E[X_n] \rightarrow E[X]$. Indeed, $E[X_n] = n^\delta$ is unbounded while $E[X] = 0$. Also $\{X_n\}_{n=1}^{\infty}$ is not uniformly integrable since $\sup \int_{|X_n| \geq N} |X_n(s)| d\lambda \geq N^\delta$.

Note that the requirement of the second proposition, that $\sup_n E[|X_n|^{1+\epsilon}] < \infty$ for some $\epsilon > 0$, is satisfied if $\delta < 0$, in which case it can also be verified directly that $E[X_n] \rightarrow E[X] = 0$.

5.3 Conditional Expectations

Given a random variable $X$ defined on a probability space $(\mathcal{S}, \mathcal{E}, \mu)$ and a sigma (sub)algebra $\mathcal{F} \subset \mathcal{E}$, the goal of this section is to define $E[X|\mathcal{F}]$, or in words, the conditional expectation of $X$ given $\mathcal{F}$. Though this appears to be an entirely different concept than that seen in the book 2 introductory developments of conditional probabilities and conditional distributions, these prior notions will be seen to be special cases of the current more general development. We begin with a generalization of the earlier work on conditional probability measures.

5.3.1 Conditional Probability Measures

Given a probability space $(\mathcal{S}, \mathcal{E}, \mu)$ and fixed $B \in \mathcal{E}$ with $\mu(B) > 0$, the conditional probability measure $\mu(\cdot|B)$ was defined on this space in definition 1.31 of book 2 by:

$$\mu(A|B) \equiv \frac{\mu(A \cap B)}{\mu(B)}, \quad \text{for } A \in \mathcal{E}. \quad (5.12)$$

For any set $B \in \mathcal{E}$ with $\mu(B) > 0$, $\mu(\cdot|B)$ is indeed a probability measure on $(\mathcal{S}, \mathcal{E})$ and it is natural to wonder if this definition can be extended to all $B \in \mathcal{E}$. Put another way by now fixing $A$, this definition provided $\mu(A|B)$, the conditional measure of $A$ relative to every set in the collection $\{B \in \mathcal{E} | \mu(B) > 0\}$, and one might wonder if this definition can be extended to all $B \in \mathcal{E}$. It turns out that this latter extension is possible, and though little explicit use of this result will be made in this book, its development introduces machinery that will be critical below in the study of conditional expectations.

To this end, let $A \in \mathcal{E}$ be given and assume that $B \in \mathcal{F}$ where $\mathcal{F}$ is a sigma subalgebra, $\mathcal{F} \subset \mathcal{E}$. Of course, this inclusion allows for the case $\mathcal{F} = \mathcal{E}$,
but gives an important generalization that will be of use below. Now the idea of defining \( \mu(A|B) \) relative to every set \( B \in \mathcal{F} \) is a small step from defining \( \mu(A|s) \) as a function, or random variable, for \( s \in \mathcal{S} \):

\[
\mu(A|s) : \mathcal{S} \rightarrow \mathbb{R}.
\]

The following definition identifies the properties that we want this function to have. The remark following provides some intuition.

**Definition 5.10** Given a probability space \((\mathcal{S}, \mathcal{E}, \mu)\), \( A \in \mathcal{E} \), and \( \mathcal{F} \subset \mathcal{E} \) a sigma subalgebra, the **conditional probability function** \( \mu(A|\mathcal{F}) \) **given** \( \mathcal{F} \), and sometimes denoted \( P[A|\mathcal{F}] \) or \( \mu[A|\mathcal{F}] \), is defined as **any** function \( f_A(s) \) so that:

1. \( f_A(s) \) is an \( \mathcal{F} \)-measurable function on \( \mathcal{S} \), and so \( f_A^{-1}(H) \in \mathcal{F} \) for all Borel sets \( H \in \mathcal{B}(\mathbb{R}) \).
2. \( f_A(s) \) is \( \mu \)-integrable, and denoting \( f_A(s) \equiv \mu(A|\mathcal{F}) \):

\[
\int_B \mu(A|\mathcal{F}) d\mu = \mu(A \cap B), \tag{5.13}
\]

for all \( B \in \mathcal{F} \), and in particular:

\[
\int_{\mathcal{S}} \mu(A|\mathcal{F}) d\mu = \mu(A). \tag{5.14}
\]

**Remark 5.11** a. Intuitively, this definition states that the conditional probability function is a random variable on \((\mathcal{S}, \mathcal{F}, \mu)\), noting the sigma algebra \( \mathcal{F} \) here. By taking various expectations of this random variable we can recover the \( \mu \)-measures of various \( \mathcal{F} \)-sets. For example, 5.13 states that for all \( B \in \mathcal{F} \) and \( \chi_B(s) \) the characteristic function of \( B \):

\[
E[\chi_B(s)\mu(A|\mathcal{F})] = \mu(A \cap B),
\]

while for 5.14:

\[
E[\mu(A|\mathcal{F})] = \mu(A).
\]

Looked at this way, it is natural to think that this is not such a big idea since it is easy to produce the \( \mu \)-measures of such sets with integration. Indeed, why not simply define \( f_A(s) \equiv \chi_A(s) \)? It is easy to see that this function is \( \mu \)-integrable and satisfies 5.13 and 5.14. The problem is that \( \chi_A(s) \) is not in general \( \mathcal{F} \)-measurable. Specifically,

\[
\chi_A^{-1}[\mathcal{B}(\mathbb{R})] = \{\emptyset, A, \bar{A}, \mathcal{S}\}.
\]
and though \( \chi_A(s) \) is always \( \mathcal{E} \)-measurable, it is \( \mathcal{F} \)-measurable if and only if \( A \in \mathcal{F} \). For general \( A \in \mathcal{E} - \mathcal{F} \), \( \chi_A(s) \) has the right integrability properties, but not the correct measurability.

In summary, the big idea in the above definition is \( \mathcal{F} \)-measurability! Thus \( \mu(A|\mathcal{F}) \) is basically \( \chi_A(s) \), but changed just enough to be \( \mathcal{F} \)-measurable, but not so much as to change the value of its integrals over \( \mathcal{F} \)-sets. With that said, it should not be obvious that such a function exists.

b. The connection between the new definition of conditional probability as a function on \( S \), \( f_A(s) \equiv \mu(A|\mathcal{F}) \), and the previous book 2 notion of a set function \( \mu(A|B) \) definable for \( B \) with \( \mu(B) > 0 \), follows from condition 2 of the definition. Specifically, for \( B \in \mathcal{F} \) with \( \mu(B) > 0 \), 5.13 and 5.12 produce:

\[
\frac{1}{\mu(B)} \int_B \mu(A|\mathcal{F}) \, d\mu = \mu(A|B).
\]

In other words, \( \mu(A|\mathcal{F}) \) is an \( \mathcal{F} \)-measurable function such that its average value over every such \( B \in \mathcal{F} \) equals the original probability of \( A \) conditional on \( B \).

c. This definition characterizes \( \mu(A|\mathcal{F}) \) as any function \( f_A(s) \) with the identified properties. Thus, \( \mu(A|\mathcal{F}) \) will not in general be unique since if \( g(s) = f_A(s) \) \( \mu \)-a.e., 5.13 and 5.14 are automatically satisfied. But one additional limitation on such \( g \) is \( \mathcal{F} \)-measurability.

Example 5.12 Let \( \{B_i\}_{i=1}^\infty \subset \mathcal{E} \) form a disjoint partition of \( S \), so \( B_i \cap B_j = \emptyset \) for \( i \neq j \) and \( \bigcup_{i=1}^\infty B_i = S \), assume all have positive measure, and let \( \mathcal{F} = \sigma (\{B_i\}) \) the smallest sigma algebra that contains these sets. Of course \( \mathcal{F} \subset \mathcal{E} \) and a general set \( B \in \mathcal{F} \) is a finite or countable union of the \( B_i \)-sets. In this case \( \mu(A|B_i) \) is well-defined for all \( A \in \mathcal{E} \) by 5.12 as noted above. For fixed such \( A \), define a function \( f_A(s) \) on \( S \) by:

\[
f_A(s) = \mu(A|B_i) \text{ for } s \in B_i,
\]

and so \( f_A(s) \equiv \mu(A|\mathcal{F}) \) is definable as a countable version of a simple function:

\[
f_A(s) = \sum_{i=1}^\infty \mu(A|B_i) \chi_{B_i}(s).
\]

Note that \( f_A(s) \) is an \( \mathcal{F} \)-measurable function on \( S \) since for any Borel set \( H \in \mathcal{B}(\mathbb{R}) \), \( f_A^{-1}(H) \) is empty or a finite or countable union of \( B_i \)-sets.
Also, $f_A$ is integrable since $f_A(s) \leq 1$ (recall $\mu$ is a probability measure), and thus by corollary 2.29 of book 5 applied to the disjoint union $S = \bigcup_{i=1}^{\infty} B_i$:

$$
\int_S \mu(A|\mathcal{F})d\mu = \sum_{i=1}^{\infty} \mu(A|B_i) \int_{B_i} \chi_{B_i}(s)d\mu
= \sum_{i=1}^{\infty} \mu(A|B_i)\mu(B_i)
= \sum_{i=1}^{\infty} \mu(A \cap B_i)
= \mu(A).
$$

Finally, if $B \in \mathcal{F}$, say $B = \bigcup_{k=1}^{\infty} B_{jk}$, then by the same calculation,

$$
\int_B \mu(A|\mathcal{F})d\mu = \mu(A \cap B).
$$

Hence, for $\mathcal{F}$ given as a sigma algebra generated by a disjoint partition of sets of positive measure, the elementary definition of $\mu(A|\cdot)$ provides a measurable function with the requisite properties. But as noted in remark 5.11 above, this definition of $f_A(s) = \mu(A|\mathcal{F})$ is not uniquely defined. Indeed, if $g_A(s)$ is any $\mathcal{F}$-measurable function on $S$ with

$$
g_A(s) = f_A(s) \mu\text{-a.e.,}
$$

then $g_A(s)$ will be perfectly usable as another version of $\mu(A|\mathcal{F})$.

Example 5.12 notwithstanding, given $A \in \mathcal{E}$ and a general sigma subalgebra $\mathcal{F} \subset \mathcal{E}$, it is not at all obvious that an $\mathcal{F}$-measurable function exists that possesses the requisite properties of definition 5.10. Indeed, to prove existence requires the power of the **Radon-Nikodým theorem** of proposition 7.22 of book 5, which then also assures uniqueness $\mu$-a.e.

**Proposition 5.13** Given a probability space $(S, \mathcal{E}, \mu)$, $A \in \mathcal{E}$, and a sigma subalgebra $\mathcal{F} \subset \mathcal{E}$, the **conditional probability function** $\mu(A|\mathcal{F})$ exists and is unique $\mu$-a.e.

**Proof.** On the probability space $(S, \mathcal{F}, \mu)$, noting that we use the sigma algebra $\mathcal{F}$ here, define a set function $\nu_A$ on $\mathcal{F}$ by:

$$
\nu_A(B) = \mu(A \cap B), \text{ for } B \in \mathcal{F}.
$$

Since $A \cap B \in \mathcal{E}$ for all such $B$, $\nu_A(B)$ is well defined and it is an exercise to check that it is in fact a measure on $\mathcal{F}$. It is also the case that $\mu|\mathcal{F}$, the restriction of $\mu$ to $\mathcal{F}$ and defined simply by

$$
\mu|\mathcal{F}(B) = \mu(B), \text{ for } B \in \mathcal{F},
$$

is a $\sigma$-finite measure on $(S, \mathcal{F})$.
is another measure on $\mathcal{F}$. Further, $\nu_A$ is absolutely continuous with respect to $\mu|_\mathcal{F}$, denoted $\nu_A \ll \mu|_\mathcal{F}$ (definition 7.3, book 5), since $\mu|_\mathcal{F}(B) = 0$ implies that $\mu(A \cap B) = 0$ and hence $\nu(B) = 0$.

The Radon-Nikodým theorem then assures that there exists a nonnegative $\mathcal{F}$-measurable function $f_A(s)$ so that for all $B \in \mathcal{F}$:

$$\int_B f_A(s) \, d\mu = \nu_A(B).$$

The function $f_A$ is necessarily $\mu$-integrable since $S \in \mathcal{F}$ and so by $(*)$,

$$\int_S f_A(s) \, d\mu = \nu_A(S).$$

Since $\nu_A(B) = \mu(A \cap B)$ and $\nu_A(S) = \mu(A)$ we can define $\mu(A|\mathcal{F}) = f_A(s)$, proving existence.

Further, the Radon-Nikodým theorem assures that $f_A$ is unique $\mu$-a.e., meaning if $g_A$ is an $\mathcal{F}$-measurable function so that $(*)$ is true with $g_A$, then $g_A = f_A$, $\mu$-a.e.

Example 5.14 It is interesting to explicitly identify $f_A(s)$ in the two extreme cases of sigma sub-algebras, $\mathcal{F} \subset \mathcal{E}$:

1. If $\mathcal{F} = \{\emptyset, S\}$, then $\mu(A|\mathcal{F}) = \mu(A)$.

2. If $\mathcal{F} = \mathcal{E}$, then $\mu(A|\mathcal{F}) = \chi_A(s)$.

Details are left as an exercise.

Notation 5.15 The conditional probability function is sometimes denoted $\mu[A|\mathcal{F}]$, which again identifies the defining set $A \in \mathcal{E}$, as well as the sigma subalgebra $\mathcal{F} \subset \mathcal{E}$. This notation is also more consistent with the notation for conditional expectation discussed below. The disadvantage of this type notation, at least initially, is that it must be kept in mind that this is not simply a probability, but is an $\mathcal{F}$-measurable function which integrates to certain probabilities. A notation such as $\mu(A|s)$ or $f_A(s)$ remedies the latter problem in that it clearly now a function and also identifies the defining set $A \in \mathcal{E}$, but does not now identify the sigma subalgebra $\mathcal{F}$. One can imagine more explicit notation that packs-in all the information, but this is rarely if ever used.
5.3.2 Conditional Expectations

Elementary Conditional Expectations

In this section we take an informal approach to motivate some ideas. Assume that $X, Y$ are random variables on a probability space $(\mathcal{S}, \mathcal{E}, \mu)$, so $X, Y : \mathcal{S} \to \mathbb{R}$ with a joint density function $f(x, y)$ and marginal densities $f(x)$ and $f(y)$. For any point $y$ for which the marginal density function $f(y) \neq 0$, the conditional density function $f(x|y)$ can often be defined by:

$$f(x|y) = \frac{f(x, y)}{f(y)}.$$ 

In the discrete case this follows from 1.42, while a similar formula is sometimes possible when $f(x, y)$ is a continuously differentiable density function, and this was seen in example 3.42 of book 2.

In such cases, the conditional expectation of $X$ given $y$ and denoted $E[X|y]$, is defined as the expectation of $X$ calculated with the conditional density function $f(x|y)$. In other words:

$$E[X|y] \equiv \int_{\mathbb{R}} xf(x|y)dx$$ \hspace{1cm} (5.15)

for continuous densities, and:

$$E[X|y] \equiv \sum x_i f(x_i|y)$$ \hspace{1cm} (5.16)

in the discrete case. Similarly, one can define the conditional variance of $X$ given $y$ by $E \left[ (X - E[X|y])^2 \right]$, and denoted by $Var[X|y]$, which with a little algebra becomes:

$$Var[X|y] = \int_{\mathbb{R}} x^2 f(x|y)dx - (E[X|y])^2$$ \hspace{1cm} (5.17)

for continuous densities, and with a comparable formula in the discrete case.

Remark 5.16 The above approach, while perhaps reflecting familiar manipulations, suffers from the following shortcoming. The above conditional expectations have been defined directly in terms of a density function which does not always exist, rather than in terms of the $\mu$-integral of a well defined random variable defined on $(\mathcal{S}, \mathcal{E}, \mu)$. To correct this and be formal, we would need to identify a random variable $Z \equiv X|y$ on $(\mathcal{S}, \mathcal{E}, \mu)$, and then define:

$$E[X|y] \equiv \int_{\mathcal{S}} Zd\mu.$$
This random variable would then have an associated distribution function, defined in the usual way by $F(z) = \mu [Z^{-1}(-\infty, z)]$. Then by 5.6:

$$E[X|y] = \int_{\mathbb{R}} zdF,$$

definable as a Riemann-Stieltjes integral. In the special cases of proposition 4.30 of book 3, this integral reduces to a Riemann integral or a summation:

$$E[X|y] = \int_{\mathbb{R}} zf(z)dz, \quad E[X|y] = \sum z_i f(z_i).$$

These are starting to look like 5.15 and 5.16, but what is the random variable $Z$ that makes these formulas reconcile?

This is easy if $X$ and $Y$ are independent, since then formulaically $f(x|y) = f(x)$ and thus $Z \equiv X$. But in the general case it is not apparent how to formalize this calculation in terms of $(\mathcal{S}, \mathcal{E}, \mu)$, even for discrete random variables. However, this idea will follow from the generalization in the following section.

Assuming for the moment that we define $f(x|y)$ as above and avoid $y$ with $f(y) = 0$, it is an exercise to derive the following identities, named analogously for the law of total probability of proposition 1.35 of book 2. We state these results in the integral format, but note that similar results follow for discrete random variables.

1. **Law of Total Probability**

$$f(x) = \int_{\mathbb{R}} f(x|y)f(y)dy. \quad (5.18)$$

2. **Law of Total Expectation**

$$E[X] = E[E[X|Y]], \quad (5.19)$$

where

$$E[E[X|Y]] = \int_{\mathbb{R}} E[X|y] f(y)dy.$$

3. **Law of Total Variance**

$$\text{Var}[X] = E[\text{Var}[X|Y]] + \text{Var}[E[X|Y]], \quad (5.20)$$

where $E[\text{Var}[X|Y]]$ is defined analogously to $E[E[X|Y]]$, and

$$\text{Var}[E[X|Y]] = \int_{\mathbb{R}} (E[X|y])^2 f(y)dy - (E[E[X|Y]])^2.$$
Example 5.17  An insurance company has a medium quality bond portfolio of 250 bonds, each with a par value of $10 million, and each with a probability of default over the next year of $p = 0.003$. On default, the company’s loss given default (LGD) is assumed to be uniformly distributed with losses of $25 – 75\%$. It is assumed that the event of default and the losses on defaulted bonds are independent random variables, and that the events of default on different bonds are independent. Neither assumption is perfectly realistic but used for simplicity. The goal is to model the mean and standard deviation of losses over the next year.

Using the so-called individual loss model, the random variable of losses $L$ can be expressed:

$$L = \sum_{i=1}^{250} B_i L_i,$$

where $B_i L_i$ is the loss on the $i$th bond. This is modelled as the product of a standard binomial $B_i$, which indicates the event of default, and the LGD variable $L_i$. The binomial equals 0 with probability 0.997, and 1 with probability 0.003, while the LGD distribution function is given as assumed above, with units in millions:

$$F_L(l) = \begin{cases} 
0, & l \leq 2.5, \\
(2l - 5.0)/10, & 2.5 \leq l \leq 7.5, \\
1, & 7.5 \leq l.
\end{cases}$$

Conditioning on $B_i$ it follows that with par $P = 10$ million:

$$E[B_i L_i | B_i] = 0.5PB_i, \quad Var[B_i L_i | B_i] = P^2B_i^2/48.$$

Applying 5.19,

$$E[B_i L_i] = (0.5P) E[B_i] = 15000.0.$$

Similarly, with 5.20,

$$Var[B_i L_i] = \left(\frac{P^2}{48}\right) E[B_i^2] + (0.5P)^2 Var[B_i]$$

$$= \left(\frac{P^2}{48}\right) (0.003) + (0.5P)^2 (0.003) (0.997)$$

$$= 8.1025 \times 10^{10}.$$

By independence it follows that $E[L] = 3.75 \times 10^6$, and $Var[L] = 2.0256 \times 10^{13}$, and so the standard deviation of losses is $SD[L] = 4.5007 \times 10^6$. 


Alternatively using the aggregate loss model, $L$ is expressed:

$$L = \sum_{i=1}^{N} L_i,$$

where $N$ is the random variable representing the number of defaults, and $L_i$ is the LGD variable defined above. The distributional model for $N$ is exactly binomial with density

$$f_N(n) = \binom{250}{n} (0.003)^n (0.997)^{250-n},$$

or approximately Poisson with parameter $\lambda = 250(0.003) = 0.75$. This follows from the Poisson limit theorem of proposition 5.5 of book 4. Conditioning on $N$ it follows that

$$E[L|N] = 0.5PN, \quad \text{Var}[L|N] = P^2N/48.$$  

Completing the calculation with 5.19 reproduces $E[L] = 3.75 \times 10^6$, with either the binomial or Poisson model for $N$, while with 5.20 we obtain the individual loss model result for binomial $N$, and with Poisson $N$ obtain $\text{Var}[L] = 2.0313 \times 10^{13}$ and $\text{SD}[L] = 4.5070 \times 10^6$.

**General Conditional Expectations - An Introduction**

Even in the elementary formulation of 5.15 or 5.16 for random variables on $(\mathcal{S}, \mathcal{E}, \mu)$ with density functions, $E[X|Y]$ can be imagined to be given by a random variable $E[X|Y]$ defined on $\mathcal{S}$ by $E[X|Y](s) \equiv E[X|Y(s)]$. Indeed, this could be boldly defined pointwise in the most direct way by:

$$E[X|Y](s) \equiv E[X|y], \text{ if } f(y) \equiv f(Y(s)) > 0.$$

But such a definition has at least two problems:

1. As defined, it is not at all clear what measurability properties this function possesses, and so it is not easily seen to be a random variable on $\mathcal{S}$.

2. A pointwise definition of a random variable can certainly be interpreted as overkill in a measure space where arbitrary redefinitions on sets of $\mu$-measure zero have no influence on the most important properties of a random variable.
As a second attempt we might instead require only that the $\mu$-integrals of $E[X|Y](s)$ over certain $\mathcal{E}$-sets be defined in terms of the Lebesgue-Stieltjes integral of $E[X|y]$ over associated Borel sets. This then overcomes problem 2, since this approach is indifferent to sets of $\mu$-measure zero, and will ultimately be seen to solve problem 1 as well. To this end, assume $f(y) > 0$ for all $y$ for simplicity. Then using 3.11 of proposition 3.14 of book 5, we could require that for all $H \in \mathcal{B}(\mathbb{R})$:

$$
\int_{Y^{-1}(H)} E[X|Y](s)d\mu = \int_H E[X|y]d\mu_Y(y),
$$

where $\mu_Y$ is the Borel measure induced by $Y$ as in definition 3.9 of book 5, and also noted in 4.5. The integral on the right is a Lebesgue-Stieltjes integral, and by definition of $\mu_Y$, the distribution function $F(y)$ of $Y$ is also the distribution function induced by $\mu_Y$, or notationally $d\mu_Y = dF$. Since $F(y)$ has an associated density function $f(y)$ by assumption, proposition 3.6 of book 5 applies to obtain the requirement that for all $H \in \mathcal{B}(\mathbb{R})$:

$$
\int_{Y^{-1}(H)} E[X|Y](s)d\mu = \int_H E[X|y]f(y)dy.
$$

From the form of this identity we see that the $\mu$-integral of $E[X|Y](s)$ need only be defined over sets of a sigma-subalgebra of $\mathcal{E}$. Define $\mathcal{F} \subset \mathcal{E}$ by $\mathcal{F} \equiv Y^{-1}(\mathcal{B}(\mathbb{R}))$, or in other words $\mathcal{F} \equiv \sigma(Y)$, the sigma algebra generated by $Y$ (definition 3.43 of book 2). The above requirement can be restated for all $B \in \mathcal{F}$:

$$
\int_B E[X|Y](s)d\mu = \int_{Y(B)} E[X|y]f(y)dy.
$$

For the integral on the left to make sense, it will be necessary to assure that $E[X|Y](s)$ is $\mathcal{F}$-measurable, or equivalently, $\sigma(Y)$-measurable.

Finally, substituting the definition of $E[X|y]$ from 5.15 and continuing to assume that $f(y) > 0$ for all $y$, this $\mu$-integral requirement can be transformed by an application of change of variables. First by 5.15:

$$
\int_B E[X|Y](s)d\mu = \int_\mathbb{R} \int_\mathbb{R} x_{\chi_{Y}(B)}(y)f(x,y)dxdy.
$$

Now define the transformation $T: \mathcal{S} \to \mathbb{R}^2$ by $T(s) = (X(s), Y(s))$. Then as above, with $\mu_T$ the measure induced by $T$:

$$
\int_\mathbb{R} \int_\mathbb{R} x_{\chi_{Y}(B)}(y)f(x,y)dxdy = \int_\mathbb{R} \int_\mathbb{R} x_{\chi_{Y}(B)}(y)d\mu_T.
$$
5.3 CONDITIONAL EXPECTATIONS

Applying 3.11 of proposition 3.14 of book 5:

\[ \int \int \mathbb{R}^2 x \chi_{(B)}(y) d\mu_T = \int \chi_B(s)X(s) d\mu \equiv \int_B X d\mu. \]

Thus combining steps, the \(\mu\)-integrals of the yet-to-be-specified random variable \(E[X|Y]\) over \(\mathcal{F}\)-sets must equal the respective \(\mu\)-integrals of \(X\):

\[ \int_B E[X|Y](s) d\mu = \int_B X d\mu. \quad (*\)

This is a promising development, because even though the above manipulations required the strong assumptions that the random variables \(X,Y\) had a joint density function and that \(f(y) > 0\) for all \(y\), this final conclusion makes sense without these restrictive assumptions.

Summary 5.18 If \(E[X|Y] \equiv E[X|Y](s)\) can be well-defined as a random variable on \(S\), the constraints on this function must be:

1. That \(E[X|Y]\) is measurable relative to \(\sigma(Y)\), and,

2. For all \(B \in \sigma(Y)\), that:

\[ \int_B E[X|Y] d\mu = \int_B X d\mu. \]

General Definition of Conditional Expectation

In this section we derive the existence of the \(\sigma(Y)\)-measurable function \(E[X|Y]\) by applying the same approach as for the existence of the conditional probability function given \(\mathcal{F}\), \(\mu(A|\mathcal{F})\) or \(\mu[A|\mathcal{F}]\). Recall that this was defined as an \(\mathcal{F}\)-measurable function on \(S\) with the property that for all \(B \in \mathcal{F}\),

\[ \int_B \mu(A|\mathcal{F})d\mu = \mu(A \cap B). \]

Here we develop the analogous argument to define the conditional expectation of \(X\) given \(\mathcal{F}\), denoted \(E[X|\mathcal{F}]\), where \(X : S \to \mathbb{R}\) is a random variable, and \(\mathcal{F} \subset \mathcal{E}\) a sigma subalgebra. When \(\mathcal{F} = \sigma(Y)\), where \(\sigma(Y) \subset \mathcal{E}\) by measurability of \(Y\), this notion is called the conditional expectation of \(X\) given \(Y\), and denoted \(E[X|Y]\).

The properties desired for \(E[X|\mathcal{F}]\) are defined next, and reflect the above discussion.
Definition 5.19 Given a probability space \((S, \mathcal{E}, \mu)\), a \(\mu\)-integrable random variable \(X\), and \(\mathcal{F} \subseteq \mathcal{E}\) a sigma subalgebra, the conditional expectation of \(X\) given \(\mathcal{F}\), denoted \(E[X|\mathcal{F}]\), is defined as any function \(f_X(s)\) so that:

1. \(f_X(s)\) is an \(\mathcal{F}\)-measurable function on \(S\), and so \(f_X^{-1}(H) \in \mathcal{F}\) for all Borel sets \(H \in \mathcal{B}({\mathbb{R}})\).

2. \(f_X(s)\) is \(\mu\)-integrable, and denoting \(f_X(s) \equiv E[X|\mathcal{F}]\):

\[
\int_B E[X|\mathcal{F}] \, d\mu = \int_B X \, d\mu, \tag{5.21}
\]

for all \(B \in \mathcal{F}\), and in particular:

\[
\int_S E[X|\mathcal{F}] \, d\mu = E[X]. \tag{5.22}
\]

If \(Y\) is a random variable on \((S, \mathcal{E}, \mu)\), the conditional expectation of \(X\) given \(Y\), denoted \(E[X|Y]\), is defined as any function \(f_X(s)\) that satisfies the above definition with \(\mathcal{F} \equiv \sigma(Y)\) where \(\sigma(Y) = Y^{-1}[\mathcal{B}({\mathbb{R}})]\).

Remark 5.20 This definition states that the conditional expectation function \(E[X|\mathcal{F}]\) is a random variable on \((S, \mathcal{F}, \mu)\), noting the sigma subalgebra \(\mathcal{F}\) here, and with the property that by taking various expectations we derive various expectations of \(X\). So for any \(B \in \mathcal{F}\) and \(\chi_B(s)\) the characteristic function of \(B\), 5.21 requires that:

\[
E[\chi_B(s)E[X|\mathcal{F}]] = E[\chi_B(s)X],
\]

while 5.22 requires that

\[
E[E[X|\mathcal{F}]] = E[X].
\]

Looked at this way, it is natural to think that this is not such a big idea, since it is easy to produce such expectations.

Indeed, simply define \(E[X|\mathcal{F}] = X\). This function is \(\mu\)-integrable by assumption, and satisfies 5.21 and 5.22. But, \(f_X(s) = X\) is not in general \(\mathcal{F}\)-measurable, since while \(f_X^{-1}[\mathcal{B}({\mathbb{R}})] \subset \mathcal{E}\), \(X\) need not be \(\mathcal{F}\)-measurable in the sense that \(X^{-1}[H] \notin \mathcal{F}\) for some \(H \in \mathcal{B}({\mathbb{R}})\). Hence, while \(X\) has the right expectation properties, it need not have the right measurability property.

Thus the big idea in the above definition, as it was for conditional probability, is again \(\mathcal{F}\)-measurability! In essence, \(E[X|\mathcal{F}]\) is a version of \(X\) that is changed just enough to produce the required \(\mathcal{F}\)-measurability. In the case of \(\mathcal{F} \equiv \sigma(Y)\) where \(\sigma(Y) = Y^{-1}[\mathcal{B}({\mathbb{R}})]\), \(E[X|Y]\) is again basically \(X\), but is changed just enough to be \(\sigma(Y)\)-measurable.
5.3 CONDITIONAL EXPECTATIONS

Given a sigma subalgebra \( \mathcal{F} \subset \mathcal{E} \), while it is again not obvious that a measurable function with the needed properties exists, the **Radon-Nikodým theorem** of proposition 7.22 of book 5 again provides a solution, and one that is unique \( \mu \)-a.e.

**Proposition 5.21** Given an integrable random variable \( X \) on \((\mathcal{S}, \mathcal{E}, \mu)\) and sigma subalgebra \( \mathcal{F} \subset \mathcal{E} \), the conditional expectation \( E[X|\mathcal{F}] \) exists and is unique \( \mu \)-a.e.

**Proof.** As for conditional probabilities, we again prove the existence of \( E[X|\mathcal{F}] \) using the Radon-Nikodým theorem but need an extra step. On the probability space \((\mathcal{S}, \mathcal{F}, \mu)\), noting the sigma algebra \( \mathcal{F} \) here, define a set function \( \nu_X \) on \( \mathcal{F} \) by:

\[
\nu_X(B) = \int_B X d\mu, \text{ for } B \in \mathcal{F}.
\]

Since \( \mathcal{F} \subset \mathcal{E} \), \( \nu_X(B) \) is well defined. But as such integrals need not be non-negative, \( \nu_X \) is in fact a **signed** measure on \( \mathcal{F} \) as in definition 7.5 of book 5. Recalling definition 2.36 of book 5, split \( X \) into positive and negative parts,

\[
X = X^+ - X^-,
\]

and define

\[
\nu_X(B) = \nu_{X^+}(B) - \nu_{X^-}(B),
\]

where \( \nu_{X^+}(B) \) and \( \nu_{X^-}(B) \) are defined analogously. Each of \( \nu_{X^+} \) and \( \nu_{X^-} \) is a measure using the same proof as for proposition 3.3 of book 5.

The restriction of \( \mu \) to \( \mathcal{F} \), denoted \( \mu|_{\mathcal{F}} \), is another measure on \( \mathcal{F} \) and both \( \nu_{X^+} \) and \( \nu_{X^-} \) are absolutely continuous with respect to \( \mu|_{\mathcal{F}} \). For example, \( \nu_{X^+} \ll \mu|_{\mathcal{F}} \) since \( \mu|_{\mathcal{F}}(B) = 0 \) implies that \( \nu_{X^+}(B) = \int_B X^+ d\mu = 0 \).

The Radon-Nikodým theorem thus assures that there exists nonnegative \( \mathcal{F} \)-measurable functions, \( f_{X^+}(s) \) and \( f_{X^-}(s) \), so that with \( f_X(s) = f_{X^+}(s) - f_{X^-}(s) \), we have for all \( B \in \mathcal{F} \):

\[
\int_B f_X(s) d\mu = \int_B X d\mu. \quad (**)
\]

The function \( f_X \) is necessarily \( \mu \)-integrable since \( |f_X(s)| = f_{X^+}(s) + f_{X^-}(s) \), and thus by (**) applied to the component functions:

\[
\int_S |f_X(s)| \, d\mu = E[|X|] < \infty,
\]

since \( X \) is integrable.
Further, the Radon-Nikodým theorem assures that each of \( f_X^+ (s) \) and \( f_X^- (s) \) are unique \( \mu \)-a.e., and thus \( f_X \) is unique \( \mu \)-a.e. That is, if \( g_X \) is an \( \mathcal{F} \)-measurable function so that \((*)\) is true with \( g_X \), then \( g_X = f_X \), \( \mu \)-a.e. ■

**Exercise 5.22** It is interesting to explicitly identify \( f_X (s) = E [X | \mathcal{F}] \) in the two extreme cases of sigma sub-algebras, \( \mathcal{F} \subset \mathcal{E} \):

1. If \( \mathcal{F} = \{\emptyset, \mathcal{S}\} \), then \( E [X | \mathcal{F}] = E [X] \).
2. If \( \mathcal{F} = \mathcal{E} \), then \( E [X | \mathcal{F}] = X \).

**Exercise 5.23** Confirm that if \( A \in \mathcal{E} \), the random variable \( X \equiv \chi_A (s) \) satisfies:

\[
E [X | \mathcal{F}] = \mu (A | \mathcal{F}) \mu \text{-a.e.} \tag{5.23}
\]

That is, the conditional expectation of the characteristic function \( \chi_A (s) \) equals the conditional probability function, \( \mu \)-a.e. Hint: Verify that \( E [X | \mathcal{F}] \) satisfies the conditions of the \( \mu (A | \mathcal{F}) \) definition, and thus by the Radon-Nikodým theorem, these are equal \( \mu \)-a.e.

**Example 5.24** In the introductory comments of the last section it was shown that when \( X \) and \( Y \) have a joint density function and \( f(y) > 0 \) for all \( y \), the definition of \( E [X | Y] \) given in 5.15 above is consistent with the definition of \( E [X | Y] \) interpreted in this more general context. Here we provide a more explicit example that is common in discrete probability theory.

Assume that \( X, Y \) are random variables on a probability space \( (\mathcal{S}, \mathcal{E}, \mu) \), so \( X, Y : \mathcal{S} \to \mathbb{R} \), and that \( Y \) has finite or countable range which we denote by \( \{y_i\}_{i=1}^{N} \) with \( N \leq \infty \). Then \( \{Y^{-1}(y_i)\}_{i=1}^{N} \equiv \{B_i\}_{i=1}^{N} \) forms a disjoint partition of \( \mathcal{S} \) in that \( \bigcup_{i=1}^{N} B_i = \mathcal{S} \) and \( B_i \cap B_j = \emptyset \) for \( i \neq j \). In addition, the sigma algebra \( \sigma (Y) \) is the collection of all unions of such sets. In order for \( E [X | Y] \) to be \( \sigma (Y) \)-measurable, it must assume a constant value on each \( B_i \). Otherwise \( E [X | Y]^{-1} (A) \) would be a proper subset of a \( B_i \)-set for some \( A \in \mathcal{B} (\mathbb{R}) \), contradicting measurability. Hence \( E [X | Y] (s) = \alpha_i \) on \( B_i \).

To determine these constants, it follows from 5.21 that:

\[
\int_{B_i} E [X | Y] \, d\mu = \int_{B_i} X \, d\mu.
\]

As \( E [X | Y] = \alpha_i \) on \( B_i \), we derive that

\[
\alpha_i = \frac{1}{\mu (B_i)} \int_{B_i} X \, d\mu.
\]

In other words, \( E [X | Y] (s) \) equals the "average" value of \( X \) over \( B_i \), where this average is calculated relative to the \( \mu \)-measure.
Remark 5.25 The last example provides a useful illustration of a general property of the random variable $E[X|Y]$. First, the integrability condition in 5.21 implies that for every $B \in \sigma(Y)$ with $\mu(B) > 0$, the average value of $E[X|Y]$ equals the average value of $X$. This follows by dividing this constraint by $\mu(B)$. If $\mu(B) = 0$, both integrals are independent of the definition of $E[X|Y]$. This example shows that when $\sigma(Y)$ is generated by countably many disjoint sets, that $E[X|Y]$ is constant on such sets and exactly equal to the given averages of $X$.

More generally, on any set $B \in \sigma(Y)$ with $\mu(B) > 0$ and for which $Y(s) = c$ for $s \in B$, it is again the case that $E[X|Y] = \alpha$ on $B$, and this constant is again equal to the average value of $X$ on $B$.

That $E[X|Y]$ can be arbitrarily defined on sets $B \in \sigma(Y)$ with $\mu(B) = 0$ supports the observation there are many "versions" of $E[X|Y]$.

5.3.3 Properties of Conditional Expectations

In the next book we will require some facility with manipulations involving conditional expectations. This section summarizes some of the most important properties that will be needed into the following proposition. To standardize terminology, we state results in the context of $E[X|F]$ since that is the context that will appear most often later. It may well be the case that such $F = \sigma(Y)$ for some random variable $Y$, in which case all these results apply to $E[X|Y]$ without change. The reader is encouraged to make these results more concrete by recasting them in the context of sigma subalgebras such as $F$ in example 5.24 above, which are generated by countably many disjoint sets. This is especially true for the tower property in 5, for which one might choose $F_2 = F$, and $F_1$ generated by the sets $\{C_i\}_{i=1}^N$ with say $C_i = B_{2i} \cup B_{2i-1}$.

Proposition 5.26 Given a probability space $(S, \mathcal{E}, \mu)$, $\mu$-integrable random variables $X, Y$, and $\{X_n\}$, and $\mathcal{F} \subset \mathcal{E}$ a sigma subalgebra, we have the following properties, all to be interpreted as $\mu$-a.e.:

1. **Linearity:** For $a, b \in \mathbb{R}$:
   \[ E[aX + bY|\mathcal{F}] = aE[X|\mathcal{F}] + bE[Y|\mathcal{F}] . \]  
   (5.24)

2. **Monotonicity:** If $X \leq Y$ $\mu$-a.e., then:
   \[ E[X|\mathcal{F}] \leq E[Y|\mathcal{F}] . \]  
   (5.25)
3. **Triangle Inequality:**

\[ |E[X|\mathcal{F}]| \leq E[|X||\mathcal{F}] \]  

(5.26)

4. **Jensen’s Inequality:** If \( \varphi \) is convex and \( \varphi(X) \) is \( \mu \)-integrable, then

\[ \varphi(E[X|\mathcal{F}]) \leq E[\varphi(X)|\mathcal{F}] \]  

(5.27)

5. **Tower Property:** If \( \mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{E} \), then:

\[ E[E[X|\mathcal{F}_1]|\mathcal{F}_2] = E[X|\mathcal{F}_1], \quad E[E[X|\mathcal{F}_2]|\mathcal{F}_1] = E[X|\mathcal{F}_1]. \]  

(5.28)

6. **\( \mathcal{F} \)-Measurability Property:** If \( Z \) is an \( \mathcal{F} \)-measurable (not necessarily \( \mu \)-integrable) random variable, but with \( XZ \) \( \mu \)-integrable, then

\[ E[XZ|\mathcal{F}] = ZE[X|\mathcal{F}]. \]  

(5.29)

7. **Independence:** If \( X \) is independent of \( \mathcal{F} \), meaning \( \sigma(X) \) and \( \mathcal{F} \) are independent sigma algebras (definition 1.15, book 2), then:

\[ E[X|\mathcal{F}] = E[X]. \]  

(5.30)

8. **\( L_p \)-Convergence for \( 1 \leq p < \infty \):** If \( X_n \to X \) in \( L_p(\mathcal{S}, \mu) \), meaning:

\[ \|X_n - X\|_p = \left[ E(|X_n - X|^p) \right]^{1/p} \to 0, \]

then \( E[X_n|\mathcal{F}] \to E[X|\mathcal{F}] \) in \( L_p(\mathcal{S}, \mu) \):

\[ \|E[X_n|\mathcal{F}] - E[X|\mathcal{F}]\|_p \to 0. \]

(5.31)

9. **Monotone Convergence:** If \( X_n \geq 0 \) is increasing, \( X_{n+1} \geq X_n \), and \( X_n \to X \) \( \mu \)-a.e. with \( E[X] < \infty \), then:

\[ E[X_n|\mathcal{F}] \to E[X|\mathcal{F}] \mu \text{-a.e.} \]  

(5.32)

**Proof.** We take each statement in turn. For needed properties of the integral, see proposition 2.40 of book 5 unless otherwise noted.
1. First, $aE[X|\mathcal{F}] + bE[Y|\mathcal{F}]$ is $\mathcal{F}$-measurable since both conditional expectations have this property. If $B \in \mathcal{F}$, then by linearity of the integral:

$$\int_B (aE[X|\mathcal{F}] + bE[Y|\mathcal{F}]) d\mu = a\int_B E[X|\mathcal{F}] d\mu + b\int_B E[Y|\mathcal{F}] d\mu = a\int_B X d\mu + b\int_B Y d\mu = \int_B (aX + bY) d\mu.$$ 

Thus $aE[X|\mathcal{F}] + bE[Y|\mathcal{F}]$ satisfies 5.21, and hence equals $E[aX + bY|\mathcal{F}]$ $\mu$-a.e.

2. For any $B \in \mathcal{F}$, monotonicity of the integral implies that:

$$\int_B E[X|\mathcal{F}] d\mu = \int_B X d\mu \leq \int_B Y d\mu = \int_B E[Y|\mathcal{F}] d\mu.$$ 

Now given $\epsilon > 0$ define $B' = \{s|E[X|\mathcal{F}] - E[Y|\mathcal{F}] \geq \epsilon > 0\}$. Then $B' \in \mathcal{F}$ by $\mathcal{F}$-measurability of $E[X|\mathcal{F}]$ and $E[Y|\mathcal{F}]$, and so by the previous estimate:

$$\int_{B'} E[X|\mathcal{F}] d\mu \leq \int_{B'} E[Y|\mathcal{F}] d\mu.$$ 

This is only possible if $\mu[B'] = 0$, and hence $E[X|\mathcal{F}] \leq E[Y|\mathcal{F}]$ $\mu$-a.e.

3. Since both $X \leq |X|$ and $-X \leq |X|$, this result follows from 2 and 1.

4. If $ax + b \leq \varphi(x)$ for all $x$, then by 1 and 2 we have that $\mu$-a.e.,

$$aE[X|\mathcal{F}] + b \leq E[\varphi(X)|\mathcal{F}],$$ 

since $E[1|\mathcal{F}] = 1$. Restricting to rational $a,b$, it follows that

$$\sup_{(a,b)} [aE[X|\mathcal{F}] + b] \leq E[\varphi(X)|\mathcal{F}],$$ 

and since this supremum is over countably many pairs $(a,b)$, it is again true $\mu$-a.e. Finally, for any convex function,

$$\varphi(y) = \sup_{(a,b)} \{ay + b\}$$ 

where the supremum is over all $(a,b)$, or all rational $(a,b)$ with $ax + b \leq \varphi(x)$ for all $x$. So $\mu$-a.e., the result follows.
5. For the first identity, $E[X|\mathcal{F}_1]$ is $\mathcal{F}_1$-measurable and hence $\mathcal{F}_2$-measurable, since $\mathcal{F}_1 \subset \mathcal{F}_2$. It will then follow from part 6 that:

$$E[E[X|\mathcal{F}_1]|\mathcal{F}_2] = E[X|\mathcal{F}_1],$$

since $E[1|\mathcal{F}_2] = 1$.

For the second identity, both functions are $\mathcal{F}_1$-measurable by definition and thus only the integrability condition in 5.21 need be checked. Let $B \in \mathcal{F}_1$, then from 5.21:

$$\int_B E[X|\mathcal{F}_1] d\mu = \int_B X d\mu,$$

and

$$\int_B E[E[X|\mathcal{F}_2]|\mathcal{F}_1] d\mu = \int_B E[X|\mathcal{F}_2] d\mu.$$

But also $B \in \mathcal{F}_1 \subset \mathcal{F}_2$, thus:

$$\int_B E[X|\mathcal{F}_2] d\mu = \int_B X d\mu,$$

and so $E[E[X|\mathcal{F}_2]|\mathcal{F}_1] = E[X|\mathcal{F}_1]$ $\mu$-a.e.

6. Since both functions are $\mathcal{F}$-measurable, we again only need to verify the integrability condition of 5.21, that for $B \in \mathcal{F}$:

$$\int_B Z E[X|\mathcal{F}] d\mu = \int_B Z X d\mu.$$

If $Z_n$ is a simple function, $Z_n = \sum_{i=1}^n a_i \chi_{A_i}(s)$ with $\{A_i\} \subset \mathcal{F}$, then by 1 and the linearity of the integral:

$$\int_B Z_n E[X|\mathcal{F}] d\mu = \sum_{i=1}^n a_i \int_{B \cap A_i} E[X|\mathcal{F}] d\mu$$

$$= \sum_{i=1}^n a_i \int_{B \cap A_i} X d\mu$$

$$= \int_B Z_n X d\mu.$$

Next, write $Z = Z^+ - Z^-$ as a decomposition into positive and negative parts (definition 2.36, book 5), and let $\{Z^+_n\}_{n=1}^\infty$ and $\{Z^-_n\}_{n=1}^\infty$ be increasing sequences of simple functions so that $Z^+_n \leq Z^+$ and $Z^+_n \to Z^+$ pointwise, and similarly define $Z^-_n$ (proposition 1.18, book
5.3 CONDITIONAL EXPECTATIONS

5). Also split $E[X|\mathcal{F}]$ and $X$ into positive and negative parts. Using Lebesgue’s monotone convergence theorem (proposition 2.21, book 5) on each of the 4 terms in the product, it follows that:

$$
\int_B Z_n E[X|\mathcal{F}] d\mu = \int_B (Z_n^+ - Z_n^-) \left( E[X|\mathcal{F}]^+ - E[X|\mathcal{F}]^- \right) d\mu \to \int_B Z \ E[X|\mathcal{F}] d\mu,
$$

and similarly,

$$
\int_B Z_n X d\mu \to \int_B Z X d\mu.
$$

The result now follows since each of the terms of these integral sequences agree by the previous result.

7. Given $B \in \mathcal{F}$, the independence assumption assures that $X$ and $\chi_B$ are independent random variables (definition 3.47, book 2), and so:

$$
\int_B X d\mu = \int_S X \chi_B d\mu = \int_S X d\mu \int_S \chi_B d\mu = E[X] \mu[B] = \int_B E[X] d\mu.
$$

Hence, $E[X]$ satisfies 5.21 and since constant, it is apparently $\mathcal{F}$-measurable.

8. By linearity, Jensen’s inequality with $\varphi(x) = |x|^p$, and the tower property:

$$
\| E[X_n|\mathcal{F}] - E[X|\mathcal{F}] \|^p_p = E \| E[X_n - X|\mathcal{F}] \|^p_p \leq E (E \| X_n - X \|^p |\mathcal{F})) = E (|X_n - X|^p).
$$

9. By assumption, $X - X_n \geq 0$ and is decreasing, integrable, and $X - X_n \to 0 \mu$-a.e., and so for all $n$:

$$
\int_A (X - X_n) d\mu \leq \int_A (X - X_1) d\mu < \infty.
$$

By definition of conditional expectation it follows that for all $B \in \mathcal{F}$,

$$
\int_B E[X - X_n|\mathcal{F}] d\mu = \int_B (X - X_n) d\mu.
$$

Lebesgue’s dominated convergence theorem (proposition 2.43, book 5) assures that as $n \to \infty$,

$$
\int_B (X - X_n) d\mu \to 0,
$$
and so

$$\int_B E[X - X_n|\mathcal{F}] d\mu \to 0$$

for all $B \in \mathcal{F}$. By linearity this implies that $E[X_n|\mathcal{F}] \to E[X|\mathcal{F}]$ \(\mu\)-a.e..

\[ \blacksquare \]

**Remark 5.27** One of the most common applications of the tower property is in evaluating $E[X]$ for a given random variable. By exercise 5.22 of the previous section, $E[X] = E[X|\{\emptyset, S\}]$, and thus also $E[E[X|\mathcal{F}]] = E[E[X|\mathcal{F}]|\{\emptyset, S\}]$. Since $\{\emptyset, S\} \subset \mathcal{F}$ for any sigma algebra $\mathcal{F}$, it follows from 5.28 applied to $E[E[X|\mathcal{F}]|\{\emptyset, S\}]$ that:

$$E[X] = E[E[X|\mathcal{F}]].$$

This will be especially useful in later books’ developments of martingale properties of the Itô and other integrals for which a proof that $E[X|\mathcal{F}] = 0$ for some $\mathcal{F}$ will be sufficient to allow the conclusion that $E[X] = 0$. 
Chapter 6

The Characteristic Function

This chapter will focus primarily on the characteristic function because it is a more general tool for analysis than is the moment generating function. That this is the case follows because the characteristic function exists for any distribution function, a property not shared by the moment generating function. See for example proposition 3.22 and remark 3.30 of book 4. That said, we begin with a short section on the multivariate moment generating function to develop a few of its properties and to complement the analogous one variables results from chapter 3 of book 4.

6.1 The Moment Generating Function

The multivariate moment generating function was introduced in definition 3.2 above. Generalizing definition 3.9 of book 4, the moment generating function of $Y \equiv (Y_1, Y_2, ..., Y_n)$ with distribution function $F_Y(y)$ is defined by:

$$M_Y(t) \equiv \int_{\mathbb{R}^n} e^{t \cdot y} dF_Y$$

when this Riemann-Stieltjes integral exists for $|t| < t_0$ for some $t_0 > 0$. Here $t = (t_1, t_2, ..., t_n)$, $|t|^2 = \sum_{j=1}^n t_j^2$, and $y \cdot t = \sum_{j=1}^n y_j t_j$ is the usual dot product or inner product on $\mathbb{R}^n$. The function $M_Y(t)$ is also called the moment generating function of $F_Y(y)$, and sometimes denoted $M_F(t)$.

First, generalizing exercise 3.4.

Exercise 6.1 Let $Y \equiv (Y_1, Y_2, ..., Y_n)$ be defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with distribution function $F_Y(y)$ and moment generating function $M_Y(t)$ which exists for
Given a matrix $A: \mathbb{R}^n \to \mathbb{R}^m$ and fixed $\mu \in \mathbb{R}^m$, let $Z: S \to \mathbb{R}^m$ be defined by $Z = AY + \mu$. Prove that $M_Z(s)$ exists for $s \in \mathbb{R}^m$ with $|s| < s_0$ for some $s_0 > 0$, and that:

$$M_Z(s) = e^{\mu^T s} M_Y(A^T s), \quad (6.1)$$

where $A^T : \mathbb{R}^m \to \mathbb{R}^n$ is the transpose of $A$ defined by $a_{ij}^T = a_{ji}$. Hint: $|A^T s|$ is continuous $\mathbb{R}^m \to \mathbb{R}$, and thus $s_0$ can be chosen so that $|s| < s_0$ implies $|A^T s| < t_0$.

For the next corollary, recall definition 1.15.

**Corollary 6.2** Given $Y \equiv (Y_1, Y_2, \ldots, Y_n)$, assume that $M_Y(t)$ exists for $|t| < t_0$ for some $t_0 > 0$, and let $I \equiv \{i_1, \ldots, i_m\} \subset \{1, 2, \ldots, n\}$. Then the moment generating function $M_{Y_I}(t)$ of the **marginal distribution function** $F_Y(y_{i_1}, y_{i_2}, \ldots, y_{i_m})$ exists for $s \in \mathbb{R}^m$ with $|s| < s_0$ for some $s_0 > 0$, and:

$$M_{Y_I}(s) = M_Y(s'), \quad (6.2)$$

where $s_j' \equiv s_j$ for $j \in I$ and $s_j' \equiv 0$ otherwise.

**Proof.** Let the matrix $A: \mathbb{R}^n \to \mathbb{R}^m$ be defined so that the $j$th row has a 1 in column $i_j$ and 0s elsewhere. Then $AY = Y_I$, and by exercise 6.1:

$$M_{Y_I}(s) = M_Y(A^T s).$$

To see that $A^T s = s'$ defined above, note that of the $n$ rows of $A^T$, rows $i_1, \ldots, i_m$ have a 1 in columns $1, \ldots, m$, respectively, and 0s elsewhere, while the other $n - m$ rows are identically 0. □

General expectations, denoted $E[g(X)]$, are defined in definition 5.1, where here we introduce the notion of **multivariate moments** generalizing section 3.2.1 of book 4. There is no single notational convention for such moments, and of necessity any such convention will be cumbersome to write, and virtually impossible to articulate. We again set this definition in $\mathbb{R}^n$, recalling the first transformation in section 5.1, but such expectations originate as integrals on the probability space $(\mathcal{S}, \mathcal{E}, \lambda)$ on which $Y$ is defined.

**Definition 6.3** Given the random vector $Y \equiv (Y_1, Y_2, \ldots, Y_n)$ with distribution function $F_Y(y)$, for any $n$-tuple of nonnegative integers $m \equiv (m_1, \ldots, m_n)$, define the associated $(m_1, \ldots, m_n)$th **multivariate moment of** $Y$, denoted $\mu'_(m_1, \ldots, m_n)$ by:

$$\mu'_(m_1, \ldots, m_n) \equiv \int_{\mathbb{R}^n} y_1^{m_1} \cdots y_n^{m_n} dF_Y, \quad (6.3)$$
when this integral exists. For notational simplicity this integrand is sometimes written as $y^m = y_1^{m_1} ... y_n^{m_n}$.

In this section we derive a few results that will be reminiscent of book 4 results where $n = 1$. The first result generalizes that book’s proposition 3.22, that existence of $M_Y(t)$ assures the existence of all moments.

**Proposition 6.4** If $M_Y(t)$ exists for $|t| < t_0$ for some $t_0 > 0$, then $\mu'_m \equiv \mu'_{(m_1, ..., m_n)}$ exists for all $n$-tuples of nonnegative integers $m \equiv (m_1, ..., m_n)$.

**Proof.** We must prove that $\int_R^n |y^m| dF_Y < \infty$, where $|y^m| = |y_1|^{m_1} ... |y_n|^{m_n}$. Note that since $|y|^k \leq |y_1| \equiv \sum_{j=1}^n |y_j|$ for all $j$, that $|y^m| \leq |y_1|^k$ where $k \equiv \sum_{j=1}^n m_j$. Thus the proof is completed by proving that $\int_R^n |y_1|^k dF_Y < \infty$ for all nonnegative integers $k$. To this end, choose $t' \equiv (t, ..., t) \in R^n$ with $|t'| < t_0$. Now $|y_1|^k \leq e^{t|y_1|}$ if $|y_1| / \ln |y_1| > k/t$, and since $x / \ln x$ is increasing and unbounded as $x \to \infty$ choose $a$ so that $|y_1| / \ln |y_1| > k/t$ if $|y_1| > a$. Then for such $y$:

$$|y_1|^k \leq e^{t|y_1|} = \exp \left[ \sum_{j=1}^n t |y_j| \right] \leq e^{t' \cdot y} + e^{-t' \cdot y}.$$ This obtains:

$$\int_R^n |y_1|^k dF_Y = \int_{|y_1| \leq a} |y_1|^k dF_Y + \int_{|y_1| > a} |y_1|^k dF_Y$$

$$\leq ca^k + \int_{|y_1| > a} (e^{t' \cdot y} + e^{-t' \cdot y}) dF_Y$$

$$\leq ca^k + M_Y(t') + M_Y(-t').$$

Thus $\int_R^n |y_1|^k dF_Y < \infty$ for all $k$. $\blacksquare$

In order to generalize proposition 3.24 of book 4 to the current context and derive an expression for the value of these moments in terms of $M_Y(t)$, it is necessary to develop some combinatorial results. First, recalling the Taylor series for $e^x$:

$$e^{t \cdot y} = \prod_{j=1}^n e^{t_j y_j}$$

$$= \prod_{j=1}^n \left( \sum_{m=0}^\infty \frac{(t_j y_j)^m}{m!} \right)$$

$$= \sum_{(m_1, ..., m_n)} \prod_{j=1}^n \frac{t_j^{m_j} y_j^{m_j}}{m_j!},$$
where this last summation is over all \( n \)-tuples of nonnegative integers, meaning \( m_j \geq 0 \) for all \( j \). When needed, this summation can also be represented:

\[
e^t y = \sum_{m=0}^{\infty} \sum_{(m_1, \ldots, m_n)} \prod_{j=1}^{n} \frac{t_j^{m_j} y_j^{m_j}}{m_j!},
\]

where the inner summation is over all nonnegative integer \( n \)-tuples \( \{m_j\}_{j=1}^{n} \) with \( \sum_{j=1}^{n} m_j = m \).

This latter expression then obtains that for all nonnegative integer \( n \)-tuples \( \{m_j\}_{j=1}^{n} \) and \( m \equiv \sum_{j=1}^{n} m_j \):

\[
\frac{\partial^m e^t y}{\partial t_1^{m_1} \cdots \partial t_n^{m_n}} \bigg|_{t=0} = y_1^{m_1} \cdots y_n^{m_n},
\]

where \( t = 0 \) is shorthand for \( t_j = 0 \) for all \( j \). This follows because this summation is absolutely convergent, and thus differentiation is valid term by term. Given \( \{m'_1, \ldots, m'_n\} \) and \( f = \prod_{j=1}^{n} t_j^{m'_j} y_j^{m'_j} / m'_j! \), then \( \frac{\partial^m f}{\partial t_1^{m_1} \cdots \partial t_n^{m_n}} \equiv 0 \) if \( \sum_{j=1}^{n} m'_j < m \), while if \( \sum_{j=1}^{n} m'_j > m \), this derivative will have at least one \( t_j \)-factor and thus is 0 again when evaluated at \( t = 0 \). Finally when \( \sum_{j=1}^{n} m'_j = m \) it will again be the case that \( \frac{\partial^m f}{\partial t_1^{m_1} \cdots \partial t_n^{m_n}} \equiv 0 \) except when \( \{m'_1, \ldots, m'_n\} = \{m_1, \ldots, m_n\} \), and then \( \frac{\partial^m f}{\partial t_1^{m_1} \cdots \partial t_n^{m_n}} = y_1^{m_1} \cdots y_n^{m_n} \).

**Proposition 6.5** If \( M_Y(t) \) exists for \( |t| < t_0 \) for some \( t_0 > 0 \), then:

\[
M_Y(t) = \sum_{(m_1, \ldots, m_n)} \prod_{j=1}^{n} \frac{t_j^{m_j} y_j^{m_j}}{m_j!} \mu'(m_1, \ldots, m_n),
\]

where this summation is over all \( n \)-tuples of nonnegative integers \( m \equiv (m_1, \ldots, m_n) \), and \( \mu'(m_1, \ldots, m_n) \) is given as in 6.3. Thus:

\[
\mu'(m_1, \ldots, m_n) = \frac{\partial^m M_Y(t)}{\partial t_1^{m_1} \cdots \partial t_n^{m_n}} \bigg|_{t=0}.
\]

**Proof.** In the notation of the prior proof, let \( t' \equiv (t, \ldots, t) \in \mathbb{R}^n \) with \( t > 0 \) and \( |t'| < t_0 \), and recall that \( e^{|t|y} \) is integrable since:

\[
e^{|t|y} = e^{t' y} + e^{-t' y}.
\]

Now with \( m \equiv \sum_{j=1}^{n} m_j \) and \( |t_j| \leq t \) for all \( j \):

\[
\left| \sum_{m \leq N} \prod_{j=1}^{n} \frac{t_j^{m_j} y_j^{m_j}}{m_j!} \right| \leq \sum_{m \leq N} \prod_{j=1}^{n} \frac{|t_j|^{m_j} |y_j|^{m_j}}{m_j!} \leq \sum_{m \leq N} \prod_{j=1}^{n} \frac{t_j^{m_j} |y_j|^{m_j}}{m_j!}.
\]
6.2 THE CHARACTERISTIC FUNCTION

On the other hand, using the calculations above:

\[ e^{t|y_1|} = \prod_{j=1}^{n} e^{t|y_j|} = \sum_{(m_1, \ldots, m_n)} \prod_{j=1}^{n} \frac{t^{m_j} |y_j|^{m_j}}{m_j!}. \]

Combining concludes that for all \( N \):

\[ \left| \sum_{m \leq N} \prod_{j=1}^{n} \frac{t^{m_j} y_j^{m_j}}{m_j!} \right| \leq e^{t|y_1|}, \]

and the expression on the left is thus integrable. By Lebesgue’s dominated convergence corollary 2.48 of book 5:

\[ M_Y(t) = \sum_{(m_1, \ldots, m_n)} \int_{\mathbb{R}^n} \prod_{j=1}^{n} \frac{t^{m_j} y_j^{m_j}}{m_j!} dF_Y \]

\[ = \sum_{(m_1, \ldots, m_n)} \prod_{j=1}^{n} \frac{t^{m_j} y_j^{m_j}}{m_j!} \int_{\mathbb{R}^n} y_1^{m_1} \ldots y_n^{m_n} dF_Y \]

\[ = \sum_{(m_1, \ldots, m_n)} \prod_{j=1}^{n} \frac{t^{m_j}}{m_j!} \mu'(m_1, \ldots, m_n). \]

Hence since \( M_Y(t) \) has a multivariate Taylor series expansion that is absolutely convergent for \( |t| < t_0 \), 6.5 follows by direct calculation as above.

6.2 The Characteristic Function

In this section we translate some of the results of the book 5 development of Fourier analysis to the context of probability theory. Specifically, given a random variable \( X \) defined on a probability space \((S, \mathcal{E}, \lambda)\) with distribution function \( F \) and induced probability measure \( \mu_F \), the characteristic function of \( \mu_F \) is defined as the Fourier transform of this finite Borel measures as defined in book 5. When the probability space \((S, \mathcal{E}, \lambda)\), random variable \( X \), and distribution function \( F \) are unimportant for the discussion, we present results directly in terms of the characteristic function of a general probability measure \( \mu \) defined on \( \mathbb{R} \). After the development of some of the important properties of characteristic functions in this section, the next section will explore applications of this theory in probability theory, extending some of the results seen in book 4.
Definition 6.6 Given a probability measure $\mu$ defined on $\mathbb{R}$, the characteristic function of $\mu$ is defined as the complex valued function of $t \in \mathbb{R}$:

$$C_\mu(t) \equiv \int_{-\infty}^{\infty} e^{ixt} d\mu(x),$$

(6.6)

where $i \equiv \sqrt{-1}$. If a probability measure $\mu$ is defined on $\mathbb{R}^n$, the characteristic function of $\mu$ is defined as the complex valued function of $t \in \mathbb{R}^n$,

$$C_\mu(t) \equiv \int_{\mathbb{R}^n} e^{ix \cdot t} d\mu(x),$$

(6.7)

where $x \cdot t = \sum_{j=1}^{n} x_j t_j$ is the usual dot product or inner product on $\mathbb{R}^n$.

Notation 6.7 If $\mu = \mu_F$, the Borel measure induced by the distribution function $F$ of a random variable $X$ defined on $(\mathcal{S}, \mathcal{E}, \lambda)$, $C_{\mu_F}(t)$ is often denoted $C_F(t)$ and called the characteristic function of the distribution function $F$. Since $e^{ixt}$ is continuous, the Lebesgue-Stieltjes integral in 6.6 is then expressible as a Riemann-Stieltjes integral using proposition 2.56 of book 5:

$$C_F(t) = \int_{-\infty}^{\infty} e^{ixt} dF(x),$$

(6.8)

in the one-variable case, and analogously for distribution functions defined on $\mathbb{R}^n$ (using proposition 2.59 of book 5). Thus for $n = 1$, $C_F(t)$ is identical to the Fourier-Stieltjes transform of the distribution function $F$. Since $e^{ixt}$ is continuous, the Lebesgue-Stieltjes integral in 6.6 is then expressible as a Riemann-Stieltjes integral using proposition 2.56 of book 5.

If $X$ is the random variable defined on a probability space $(\mathcal{S}, \mathcal{E}, \lambda)$ with distribution function $F$, then consistent with 5.2 and 5.5 it follows that

$$C_F(t) \equiv E[e^{iXt}]$$

in the 1-dimensional case, while analogously:

$$C_F(t) \equiv E[e^{iX \cdot t}],$$

in the $n$-dimensional case. In these cases it is also common to use $C_X(t)$ to denote these characteristic functions, and call them the characteristic function of $X$.

If $F$ is absolutely continuous and has a density function $f$, then using proposition 3.6 of book 5, it is also the case that as a Lebesgue integral:

$$C_F(t) = (\mathcal{L}) \int_{-\infty}^{\infty} e^{ixt} f(x) dx,$$

(6.8)
with an analogous formula in the n-dimensional case using the same book 5 result. This equals the associated Riemann integral when \( f(x) \) is continuous by proposition 2.64 of book 3. Hence \( C_F(t) \) is also the Fourier transform of \( f(x) \) in the 1-dimensional case, denoted \( \hat{f}(t) \) in definition 6.26 of book 5, and this is also the convention in the n-dimensional case though not explicitly studied in book 5.

If \( F \) is a discrete distribution, then consistent with the above noted book 3 propositions:

\[
C_F(t) = \sum_{i=1}^{\infty} e^{ix_i t} f(x_i). \tag{6.9}
\]

It is rare that any confusion is caused by these multiple notations or terminology.

**Remark 6.8** Because \( |e^{ixt}| = 1 \) by 6.13 of book 5, \( C_F(t) = E[e^{ixt}] \) is always well defined for all \( t \) since by the triangle inequality:

\[
|C_F(t)| \leq E[|e^{ixt}|] \leq 1. \tag{6.10}
\]

The same result is true for general \( n \) since again \( |e^{ixt}| = 1 \). This is in stark contrast to the moment generating function of book 4, \( M_X(t) \) or \( M_F(t) \), generalized in the previous section. By that book’s proposition 3.24, the existence of \( M_X(t) \) on an open interval \((-t_0, t_0)\) required all moments to be finite. But existence of all moments was seen to be not sufficient for the existence of \( M_X(t) \) as the lognormal distribution exemplified in remark 3.30.

When \( M_F(t) \) exists on \((-t_0, t_0)\), then \( C_F(t) = M_F(it) \) on this interval as will be shown below 6.29. This result makes the calculation of \( C_F(t) \) easy when \( M_F(t) \) exists for all \( t \), whereas when it exists on a more limited domain, or not at all, \( C_F(t) \) must be calculated directly and requires complex variable summations or integration.

### 6.3 Examples of Characteristic Functions

In this section we list the characteristic functions of probability densities introduced in previous books. Derivations are mostly left to the reader as exercises, although as noted above, \( C_F(t) = M_F(it) \) for \( t \in (-t_0, t_0) \) if \( M_F(t) \) exists on this interval. When this interval is bounded, the functional form of \( C_F(t) \) must be verified outside of this interval. In cases where \( M_F(t) \) does not exists, one obtains \( C_F(t) \) by complex variable summations or integration.
Consistent with section 3.2.5 of book 4, we split the section into discrete and continuous distributions.

6.3.1 Discrete Distributions

1. **Discrete Rectangular Distribution on \([0, 1]\):**

   The discrete rectangular distribution is defined by the density function:
   \[
   f_R(j/n) = 1/n, \quad j = 1, 2, \ldots, n.
   \]
   Applying 6.29:
   \[
   C_R(t) = \frac{\exp[i(1 + 1/n)t] - \exp[it/n]}{n(\exp[it/n] - 1)}. \tag{6.11}
   \]
   These results can be generalized to a discrete rectangular distribution on \([a, b]\) defined in 3.45 of book 4, and then:
   \[
   C_{R_{a,b}}(t) = e^{iat}C_R([b - a]t). \tag{6.12}
   \]
   A special case of this distribution with \(n = 1\) and all probability mass at \(x_0\) is sometimes called a **delta function** and denoted \(\delta_{x_0}\).
   By definition: \(\delta_{x_0}(x_0) = 1\), and \(\delta_{x_0}(x) = 0\) for \(x \neq x_0\). Thus:
   \[
   C_\delta(t) = e^{itx_0}. \tag{6.13}
   \]

2. **Binomial Distribution:**

   The binomial distribution has parameters \(n\) and \(0 < p < 1\) and is defined by the density:
   \[
   f_B(j) = \binom{n}{j} p^j (1 - p)^{n-j}, \quad j = 0, 1, \ldots, n.
   \]
   Applying 6.29:
   \[
   C_B(t) = (1 + p(e^{it} - 1))^n. \tag{6.14}
   \]

3. **Geometric Distribution:**

   The geometric distribution has parameter \(0 < p < 1\) and is defined by the density:
   \[
   f_G(j) = p(1 - p)^j, \quad j = 0, 1, 2, \ldots
   \]
The geometric summation in 6.9 is absolutely convergent for all $p$ and thus:

$$C_G(t) = \frac{p}{1 - (1-p)e^{it}}.$$  \hspace{1cm} (6.15)

In contrast, $M_G(t)$ exists only for $t$ with $(1-p)e^t < 1$, or $t < -\ln(1-p)$.

4. **Negative Binomial Distribution**:  
The negative binomial distribution has parameters $k \in \mathbb{N}$ and $0 < p < 1$ and is defined by the density:

$$f_{NB}(j) = \binom{j+k-1}{k-1}p^k(1-p)^j, \ j = 0, 1, 2, \ldots$$

The summation in 6.9 is again absolutely convergent for all $p$ and thus:

$$C_{NB}(t) = \left( \frac{p}{1 - (1-p)e^{it}} \right)^k,$$ \hspace{1cm} (6.16)

while in contrast, $M_{NB}(t)$ exists only for $t < -\ln(1-p)$.

5. **Poisson Distribution**:  
The Poisson distribution has parameter $\lambda > 0$ and is defined by the density:

$$f_P(j) = e^{-\lambda} \frac{\lambda^j}{j!}, \ j = 0, 1, 2, \ldots$$

Applying 6.29:

$$C_P(t) = \exp[\lambda(e^{it} - 1)].$$ \hspace{1cm} (6.17)

### 6.3.2 Continuous Distributions

1. **Continuous Uniform Distribution**:  
The continuous uniform distribution is defined on $x \in [a, b]$ by density:

$$f_U(x) = \begin{cases} 1/(b-a), & x \in [a, b], \\ 0, & x \notin [a, b]. \end{cases}$$

Applying 6.29:

$$C_U(t) = \frac{e^{ibt} - e^{iat}}{t(b-a)}.$$ \hspace{1cm} (6.18)
2. Exponential Distribution and Gamma Distribution

The exponential distribution has parameter $\lambda > 0$ and is defined by the density:

$$f_E(x) = \begin{cases} 0, & x < 0, \\ \lambda e^{-\lambda x}, & x \geq 0. \end{cases}$$

The exponential is a special case of the more general gamma distribution defined with parameters $\lambda > 0$ and $\alpha > 0$ by:

$$f_G(x) = \begin{cases} 0, & x < 0, \\ \lambda^\alpha x^{\alpha-1} e^{-\lambda x}/\Gamma(\alpha), & x \geq 0, \end{cases}$$

where the gamma function $\Gamma(\alpha)$ is defined in 2.25. The characteristic functions are then defined for all $t$ by:

$$C_G(t) = (1 - it/\lambda)^{-\alpha}, \quad C_E(t) = (1 - it/\lambda)^{-1}. \quad (6.19)$$

In contrast, $M_G(t)$ and $M_E(t)$ exist only for $t < \lambda$.

3. Beta Distribution

The beta distribution has parameters $\nu > 0$, $w > 0$ and is defined on the interval $[0, 1]$ by the density function:

$$f_B(x) = \frac{x^{\nu-1}(1 - x)^{w-1}}{B(\nu, w)},$$

where the beta function $B(u, v)$ is defined in 4.26. The characteristic function is then given as in the book 4 moment generating function derivation by:

$$C_B(t) = 1 + \sum_{n=1}^{\infty} \prod_{j=0}^{n-1} \left( \frac{v + j}{v + w + j} \right) \frac{(it)^n}{n!}. \quad (6.20)$$

4. Cauchy Distribution

The Cauchy density has parameters $\gamma > 0$ and $x_0$, and is defined by:

$$f_C(x) = \frac{1}{\pi \gamma} \frac{1}{1 + \left( \frac{x - x_0}{\gamma} \right)^2},$$
while the standard Cauchy distribution is parameterized with $x_0 = 0$ and $\gamma = 1$. The Cauchy distribution has no finite moments and hence no moment generating function. However, the characteristic function can be calculated based on contour integration and an important result called **Cauchy’s integral theorem**, named for **Augustin-Louis Cauchy** (1789 – 1857), which is outside the scope of our analysis. Alternatively, we can use Fourier inversion as seen in the next section to produce:

$$C_C(t) = \exp (ix_0 t - \gamma |t|). \tag{6.21}$$

The Fourier inversion approach using 6.36 is assigned as an exercise. Note that $C_C(t)$ is not differentiable at $t = 0$ consistent with the fact that this distribution has no finite moments.

5. **Normal Distribution**

The normal distribution has parameters $\sigma > 0$ and $\mu$ and is defined by the density:

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

while the standard normal distribution is parameterized with $\mu = 0$ and $\sigma = 1$. Applying 6.29:

$$C_N(t) = \exp \left( i\mu t - \frac{1}{2} \sigma^2 t^2 \right). \tag{6.22}$$

6. **Multivariate Normal Distribution**

The density function of the multivariate normal distribution is defined in 3.3 with symmetric, positive definite $n \times n$ matrix $C$ and vector $\mu = (\mu_1, \mu_2, ..., \mu_n)$. As a function of $y = (y_1, y_2, ..., y_n)$, this density is given by:

$$f_Y(y) = (2\pi)^{-n/2} [\det C]^{-1/2} \exp \left[ -\frac{1}{2} (y - \mu)^T C^{-1} (y - \mu) \right].$$

By 3.9, $C$ is the covariance matrix of the random vector $(Y_1, Y_2, ..., Y_n)$ and defined by $C_{ij} = E[(Y_i - \mu_i) (Y_j - \mu_j)]$, and $\mu$ is the vector of first moments, $\mu_j = E[Y_j]$. Assigned as an exercise using a change of variables is to prove that generalizing 6.22:

$$C_Y(t) = \exp \left[ i\mu \cdot t - \frac{1}{2} t^T C t \right]. \tag{6.23}$$
Here $t^T Ct$ as above denotes the matrix product with $t^T$ the row vector transpose of the column vector $t$. Note that as in the 1-dimensional case, $C_Y(t) = M_Y(it)$ where $M_Y(t)$ denotes the moment generating function of $Y$ as in 3.7. We prove this general result in 6.29 of proposition 6.11.

The multivariate normal distribution can also be defined when $C$ is positive semidefinite, meaning $x^T Cx \geq 0$ but without the positive definite restriction that $x^T Cx = 0$ if and only if $x = 0$. See definition 3.6, but recall that by proposition 3.33 that such distributions do not in general have density functions.

7. Lognormal Distribution The lognormal distribution has parameters $\sigma > 0$ and $\mu$ and is defined on $(0, \infty)$ by:

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

(6.24)

Although $C_L(t)$ exists in theory, there is no known closed form version of this function.

### 6.4 Properties of Characteristic Functions on $\mathbb{R}$

Because virtually all the hard work on Fourier transforms was done in book 5, it is possible to simply quote the most important properties of this transform, adapted to the context of characteristic functions. Results with no counterpart from book 5 will be proved. Then at the end of this chapter we return to the loose end left from book 4 and stated without proof in that book’s proposition 3.57. That result states that if a distribution function $F$ has moments of all orders and the associated power series converges absolutely on an open interval about 0, then $F$ is uniquely determined by these moments.

We summarize properties in a series of propositions.

**Proposition 6.9 (Affine Transforms of $X$ and $C_F(t)$)** Given $X$ defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with distribution function $F$, let $Y = aX + b$ for $a, b \in \mathbb{R}$. Then

$$C_Y(t) = e^{ibt} C_X(at).$$

(6.25)

**Proof.** This result is similar to proposition 3.14 of book 4, which addressed an analogous property of moment generating functions. By definition,

$$C_Y(t) = E[e^{itY}] = \int e^{itY}d\lambda.$$
6.4 PROPERTIES OF CHARACTERISTIC FUNCTIONS ON \( \mathbb{R} \)

Thus

\[
C_Y(t) = \int e^{(aX+b)t} d\lambda \\
= e^{ibt} E[e^{iXat}] \\
= e^{ibt} C_X(at).
\]

Exercise 6.10 Prove this result working with integrals defined in terms of the associated probability measures \( \mu_F \) and \( \mu_G \) where \( G \) is the distribution function of \( Y \). Hint: Consider the transformation \( T : (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_F) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_G) \) defined by \( T : x \to ax + b \). Check that \( \mu_G = (\mu_F)_T \) in the notation of definition 3.9 of book 5, then see that book’s proposition 3.14.

Proposition 6.11 (Smoothness of \( C_F(t) \)) Given a distribution function \( F \) and induced Borel measure \( \mu_F \):

1. \( C_F(t) \) is uniformly continuous on \( \mathbb{R} \), and \( C_F(0) = 1 \).

2. If for positive integer \( n \) we have \( \int_{-\infty}^{\infty} |x^n| d\mu_F < \infty \), then \( C_F(t) \) is differentiable up to order \( n \),

\[
C_F^{(k)}(t) = \int_{-\infty}^{\infty} (ix)^k e^{itx} d\mu_F \quad \text{for} \quad 1 \leq k \leq n, \tag{6.26}
\]

and \( C_F^{(k)}(t) \) is uniformly continuous for \( 1 \leq k \leq n \). Thus:

\[
C_F^{(k)}(0) = i^k \mu'_k. \tag{6.27}
\]

3. If \( \int_{-\infty}^{\infty} e^{sx} d\mu_F < \infty \) for any \( s \neq 0 \), then \( C_F(t) \) is infinitely differentiable and for \( |t| \leq |s| \):

\[
C_F(t) = \sum_{j=0}^{\infty} \frac{(it)^j}{j!} \mu'_k, \tag{6.28}
\]

where \( \mu'_k = \int_{-\infty}^{\infty} x^k d\mu_F \) is the \( k \)th moment of \( X \), also denoted \( E[X^k] \).

Thus for \( |t| \leq |s| \),

\[
C_F(t) = M_F(it). \tag{6.29}
\]

Proof. This is a restatement of proposition 6.31 of book 5, except \( C_F(0) = 1 \), which is true by definition, and 6.29, which follows from 6.28 and proposition 3.24 of book 4. ■
Proposition 6.12 (Behavior of $C_F(t)$ at $\pm \infty$)  

1. (Riemann-Lebesgue Lemma) If $F$ is given by a density function $f(x)$, of necessity an integrable function, then:

$$|C_F(t)| \to 0 \text{ as } t \to \pm \infty. \quad (6.30)$$

2. If $f^{(k)}(x)$ exists and is an integrable function for $k \leq n$, then:

$$|C_F(t)| = o(|t|^{-n}) \text{ as } t \to \pm \infty. \quad (6.31)$$

**Proof.** This is a restatement of proposition 6.34 of book 5. 

Proposition 6.13 (Characteristic Functions and Independent Sums)

If $\{X_i\}_{i=1}^{n}$ are independent random variables with respective distribution functions $\{F_i\}_{i=1}^{n}$, then with $X = \sum_{i=1}^{n} X_i$:

$$C_X(t) = \prod_{i=1}^{n} C_{X_i}(t). \quad (6.32)$$

**Proof.** By induction this result only needs to be proved for $n = 2$. If these distribution functions have associated density functions $\{f_i\}_{i=1}^{2}$, then $X = \sum_{i=1}^{2} X_i$ has a density function $f(x) = f_1 * f_2(x)$ by 2.13 of corollary 2.10, and then 6.32 is a restatement of proposition 6.36 of book 5. We will generalize this in the next section, deriving this result for general joint distribution functions, and without the assumption of the existence of densities. 

Proposition 6.14 (Inverse Fourier-Stieltjes transform; Uniqueness of $C_F(t)$)

For $b > a$,

$$\lim_{T \to \infty} (2\pi)^{-1} \int_{-T}^{T} \frac{e^{-iat} - e^{-ibt}}{it} C_F(t) dm = \mu_F([a,b]) + \frac{1}{2} \mu_F([a,b]). \quad (6.33)$$

Hence if $F$ and $G$ are distribution functions and $C_F(t) = C_G(t)$ for all $t$, then $\mu_F(A) = \mu_G(A)$ for all Borel sets $A \in \mathcal{B}(\mathbb{R})$. Thus for all $x$:

$$F(x) = G(x). \quad (6.34)$$

**Proof.** This is a restatement of proposition 6.37 and corollary 6.42 of book 5. Letting $A = (-\infty, x]$ obtains the conclusion on distribution functions. 

6.5 PROPERTIES OF CHARACTERISTIC FUNCTIONS ON $\mathbb{R}^n$

Proposition 6.15 (Inversion of $C_F(t)$) If $C_F(t)$ is Lebesgue integrable, then $F$ has a continuous density function $f(x)$, and hence for all $A \in \mathcal{B}(\mathbb{R})$:

$$\mu_F[A] = \int_A f(x)dm. \quad (6.35)$$

In addition:

$$f(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} C_F(t)e^{-ixt}dt. \quad (6.36)$$

**Proof.** This is a restatement of proposition 6.50 of book 5.

Recalling that $F_n \Rightarrow F$ and $\mu_{F_n} \Rightarrow \mu_F$ are equivalent statements by definition 4.6:

Proposition 6.16 (Levy’s Continuity Theorem for $C_F(t)$) Let $\{F_n, F\}$ be a collection of distribution functions, and $\{C_{F_n}, C_F\}$ the associated characteristic functions. Then $F_n \Rightarrow F$ if and only if $C_{F_n}(t) \rightarrow C_F(t)$ for all $t$.

In addition, given $\{F_n\}$ and associated characteristic functions $\{C_{F_n}\}$, if $C_{F_n}(t) \rightarrow \phi(t)$ for all $t$ where $\phi(t)$ is continuous at $t = 0$, then there exists a distribution function $F$ so that $F_n \Rightarrow F$ and $\phi(t) = C_F(t)$.

**Proof.** This is a restatement of proposition 6.54 and corollary 6.55 of book 5.

Remark 6.17 It was an oversight in book 5 not to acknowledge that the continuity results in proposition 6.54 and corollary 6.55 combine to produce Lévy’s continuity theorem, named for Paul Lévy (1886 – 1971).

6.5 Properties of Characteristic Functions on $\mathbb{R}^n$

By definition 6.6, the characteristic function of a probability measure $\mu$ on $\mathbb{R}^n$ is a complex valued function of $t \in \mathbb{R}^n$ given in 6.7 as a Lebesgue-Stieltjes integral:

$$C_\mu(t) \equiv \int_{\mathbb{R}^n} e^{ix \cdot t}d\mu(x),$$

where $x \cdot t = \sum_{j=1}^{n} x_j t_j$ is the usual dot product or inner product on $\mathbb{R}^n$. If $F$ denotes the distribution function induced by the measure $\mu$, the characteristic function is also called the characteristic function of $F$. 

and denoted $C_F(t)$, and can be expressed as a Riemann-Stieltjes integral by proposition 2.59 of book 5:

$$C_F(t) \equiv \int_{\mathbb{R}^n} e^{ix \cdot t} dF(x).$$

When $F$ has an associated density function $f$, we have by proposition 3.6 of book 5:

$$C_F(t) \equiv (L) \int_{\mathbb{R}^n} f(x)e^{ix \cdot t} dx,$$

expressed as a Lebesgue integral. This equals the associated Riemann integral when $f(x)$ is continuous by proposition 2.64 of book 3.

With the aid of the one-dimensional theory, a number of essential results on this $n$-dimensional theory can be derived with the aid of Fubini’s theorem of proposition 5.15 of book 5. We begin with a result on transformations, analogous to exercise 6.1 for the moment generating function.

**Proposition 6.18 (Affine Transforms of $X$ and $C_F(t)$)** Given $Y \equiv (Y_1, Y_2, ..., Y_n)$ defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with distribution function $F_Y(y)$, a matrix $A : \mathbb{R}^n \to \mathbb{R}^m$ and fixed $b \in \mathbb{R}^m$, let $Z : S \to \mathbb{R}^m$ be defined by $Z = AY + b$ with distribution function $G$. Then:

$$C_G(s) = e^{ibs} C_F(A^T s),$$

where $A^T : \mathbb{R}^m \to \mathbb{R}^n$ is the transpose of $A$ defined by $a_{ij}^T \equiv a_{ji}$.

**Proof.** This is a change of variable result in Lebesgue-Stieltjes integrals, and so we need to carefully set this up. Let $\mu_F$ and $\mu_G$ denote the measures induced by $F$ and $G$ respectively, and define the transformation:

$$T : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu_F) \to (\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m), \mu_G)$$

by $T : x \to Ax + b$. This transformation induces a measure on the range space denoted $(\mu_F)_T$ in definition 3.9 of book 5, and we now show that $(\mu_F)_T = \mu_G$. If $A \in \mathcal{B}(\mathbb{R}^m)$ then by 3.9 of book 5:

$$(\mu_F)_T(A) \equiv \mu_F \left[ T^{-1}(A) \right] \equiv \lambda \left[ Y^{-1}T^{-1}(A) \right] \equiv \mu_G(A),$$

where the last step follows from $Y^{-1}T^{-1} = (TY)^{-1}$ and $TY = Z$. Thus by proposition 3.14 of book 5:

$$C_G(s) \equiv \int_{\mathbb{R}^n} e^{ix \cdot s} d\mu_G(x) = \int_{\mathbb{R}^n} e^{iTx \cdot s} d\mu_F(x).$$
6.5 PROPERTIES OF CHARACTERISTIC FUNCTIONS ON $\mathbb{R}^N$

By definition of $Tx = Ax + b$:

$$C_G(s) = e^{ib \cdot s} \int_{\mathbb{R}^n} e^{iAx \cdot x} d\mu_F(x).$$

The final step is to note that $Ax \cdot s = x \cdot A^T s$, and this obtains 6.38.

The next result addresses smoothness properties of $C_F(t)$, and generalizes proposition 6.11 of the prior section.

**Proposition 6.19 (Smoothness of $C_F(t)$)** Given $X \equiv (X_1, X_2, \ldots, X_n)$ defined on $(S, \mathcal{E}, \lambda)$ with distribution function $F_X(x)$:

1. $C_F(t)$ is uniformly continuous on $\mathbb{R}^n$ and $C_F(0) = 1$.

2. If for positive integer $k$ we have $\int_{\mathbb{R}^n} |x|^k dF(x) < \infty$ where $|x|_1 \equiv \sum_{j=1}^n |x_j|$, then $C_F(t)$ is differentiable up to order $k$. Specifically, given $(m_1, \ldots, m_n)$ with $\sum_{j=1}^n m_j = m \leq k$:

$$\frac{\partial^m C_F(t)}{\partial t_1^{m_1} \cdots \partial t_n^{m_n}} = i^m \int_{\mathbb{R}^n} \prod_{j=1}^n x_j^{m_j} e^{it_1 x_1} dF(x) \quad \text{for } 1 \leq k \leq n, \quad (6.39)$$

and $\frac{\partial^m C_F(t)}{\partial t_1^{m_1} \cdots \partial t_n^{m_n}}$ is uniformly continuous for $m \leq k$. Thus:

$$\frac{\partial^m C_F(t)}{\partial t_1^{m_1} \cdots \partial t_n^{m_n}} \bigg|_{t=0} = i^m \mu'(m_1, \ldots, m_n), \quad (6.40)$$

where $t = 0$ means $t_j = 0$ for all $j$, and $\mu'(m_1, \ldots, m_n)$ is defined in 6.3.

3. If $\int_{\mathbb{R}^n} e^{s|x|_1} dF(x) < \infty$ for any $s > 0$, then $C_F(t)$ is infinitely differentiable and for $|t| \leq s$:

$$C_F(t) = \sum_{(m_1, \ldots, m_n)} i^m \prod_{j=1}^n \frac{t_j^{m_j}}{m_j!} \mu'(m_1, \ldots, m_n), \quad (6.41)$$

where $m = \sum_{j=1}^n m_j$ and the summation is over all nonnegative integer $n$-tuples $(m_1, \ldots, m_n)$.

In addition, for $|t| \leq s$,

$$C_F(t) = M_F(it). \quad (6.42)$$

**Proof.** We address these statements in turn.
1. First, \( C_F(0) = 1 \) by definition. Since \(|e^{ix}| = 1\) for all \( x \) by 6.13 of book 5, \(|e^{i(t+h)x} - e^{itx}| = |e^{ihx} - 1|\) is independent of \( t \), and so:

\[
|C_F(t + h) - C_F(t)| \leq \int_{\mathbb{R}^n} |e^{ihx} - 1| \, dF(x) \leq 2.
\]

By continuity \( e^{ihx} - 1 \to 0 \) for all \( x \in \mathbb{R}^n \) as \( h \to 0 \), and thus by Lebesgue’s dominated convergence theorem,

\[
|C_F(t + h) - C_F(t)| \to 0
\]

as \( h \to 0 \). Further, because this convergence is dominated by an expression which is independent of \( t \), this proves uniform continuity.

2. First note that:

\[
|x|_k^k = \sum_{m=k}^{n} \prod_{j=1}^{n} |x_j|^{m_j},
\]

where the summation is over all \( n \)-tuples and \( m \equiv \sum_{j=1}^{n} m_j \). Now given \( (m_1, ..., m_n) \) with \( m \leq k \):

\[
\prod_{j=1}^{n} |x_j|^{m_j} \leq \max \left( 1, \prod_{j=1}^{n} |x_j|^{m'_j} \right),
\]

where \( (m'_1, ..., m'_n) \geq (m_1, ..., m_n) \) componentwise and \( m' \equiv \sum_{j=1}^{n} m'_j = k \). Thus \( \int_{\mathbb{R}^n} |x|_k^k \, dF(x) < \infty \) assures that for all \( m \leq k \):

\[
\int_{\mathbb{R}^n} \prod_{j=1}^{n} |x_j|^{m_j} \, dF(x) < \infty.
\]

(*)

With this estimate, the proof that first partial derivatives \( \frac{\partial C_F(t)}{\partial t} \) exist for \( 1 \leq j \leq n \), are given by 6.39 with \( m = 1 \) and are uniformly continuous, now follows the proof of 2 of proposition 6.31 of book 5. For the induction step, assume that 6.39 is true for given \( (m_1, ..., m_n) \) with associated \( m < k \), and that \( \frac{\partial^m C_F(t)}{\partial t_1^{m_1} ... \partial t_n^{m_n}} \) is uniformly continuous.

Let \( t' \) be defined with \( t'_1 = t'_1 + h \), and \( t'_j = t'_j \) otherwise. Defining \( G(t) \equiv \frac{\partial^m C_F(t)}{\partial t_1^{m_1} ... \partial t_n^{m_n}} \) for notational convenience:

\[
\frac{G(t') - G(t)}{h} = \int_{\mathbb{R}^n} \prod_{j=1}^{n} x_j^{m_j} e^{itx} e^{ihx_1} - 1 \, dF(x).
\]
6.5 PROPERTIES OF CHARACTERISTIC FUNCTIONS ON $\mathbb{R}^N$

Since $\left| \frac{e^{iht_j} - 1}{h} \right| \leq |x_1|$ and $m + 1 \leq k$:

$$\left| \frac{G(t') - G(t)}{h} \right| \leq \int_{\mathbb{R}^n} |x_1| \prod_{j=1}^n |x_j|^{m_j} dF(x) < \infty.$$ 

Thus Lebesgue’s dominated convergence theorem applies, and since $\frac{e^{iht_j} - 1}{h} \to ix_1$ pointwise, 6.39 is true for $(m_1 + 1, m_2, ..., m_n)$, and this partial derivative is uniformly continuous using the same approach.

3. Recalling the proof of proposition 6.5, if $|t_j| \leq s$ for all $j$, then for all $N$:

$$\sum_{m \leq N} \prod_{j=1}^n (it_j)^{m_j} x_j^{m_j} \leq e^{s|x|_1}.$$ 

Thus $\int_{\mathbb{R}^n} e^{s|x|_1} dF(x) < \infty$ allows the same application of Lebesgue’s dominated convergence theorem and this obtains 6.41. Comparing 6.41 and 6.4 proves 6.42.

For the next result we generalize 6.32 to be applicable to independent random vectors, but more importantly, to be applicable without the assumption that the random vectors have associated joint density functions.

**Proposition 6.20 (Characteristic Functions and Independent Sums)**

Let $\{X_i\}_{i=1}^m$ be independent random vectors with range in $\mathbb{R}^n$ and respective joint distribution functions $\{F_i\}_{i=1}^m$. Then with $X \equiv \sum_{i=1}^m X_i$:

$$C_X(t) = \prod_{i=1}^m C_{X_i}(t). \quad (6.43)$$

**Proof.** By induction it is only necessary to prove this result for $m = 2$. Letting $Z = X + Y$, we have with $\mu_Z$ the Borel measure on $\mathbb{R}^n$ induced by the distribution function of $Z$:

$$C_Z(t) \equiv \int_{\mathbb{R}^n} e^{iz \cdot t} d\mu_Z.$$ 

For notational purposes, assume that $X, Y$ are defined on $\mathcal{S}, \mathcal{E}, \lambda$. Define a new random vector by $W = (X, Y)$:

$$W : (\mathcal{S}, \mathcal{E}, \lambda) \to \left( \mathbb{R}^{2n}, \mathcal{B}(\mathbb{R}^{2n}), \mu_X \times \mu_Y \right),$$
then define by $T(x, y) = x + y$ the transformation:

$$T : (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu_X \times \mu_Y) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu_Z).$$

This transformation induces a measure on the range space denoted $(\mu_X \times \mu_Y)_T$ in definition 3.9 of book 5, and we now show that $(\mu_X \times \mu_Y)_T = \mu_Z$. For $A \in \mathcal{B}(\mathbb{R}^n)$ we have by definition:

$$(\mu_X \times \mu_Y)_T(A) \equiv \mu_X \times \mu_Y\left[T^{-1}(A)\right] \equiv \lambda\left[W^{-1}T^{-1}(A)\right] \equiv \mu_Z(A),$$

since $W^{-1}T^{-1} = (TW)^{-1}$ and $TW = Z$. Thus by proposition 3.14 of book 5:

$$C_Z(t) \equiv \int_{\mathbb{R}^n} e^{ix \cdot t} d\mu_Z = \int_{\mathbb{R}^2n} e^{i(x+y) \cdot t} d(\mu_X \times \mu_Y).$$

As $|e^{i(x+y) \cdot t}| = 1$, $e^{i(x+y) \cdot t}$ is integrable over $\mathbb{R}^n$ since $\mu_X \times \mu_Y$ is a probability measure, and so Fubini’s theorem of proposition 5.19 of book 5 applies to obtain:

$$C_Z(t) = \int_{\mathbb{R}^n} e^{i x \cdot t} d\mu_X \int_{\mathbb{R}^n} e^{i y \cdot t} d\mu_Y = C_X(t)C_Y(t).$$

**Remark 6.21** Note that by definition 2.5, $\mu_Z = \mu_X \ast \mu_Y$, but this was not needed for this proof.

**Notation 6.22** The result in 6.43 is sometimes expressed in terms of the characteristic function of the convolution of distribution functions, recalling notation 2.11. Applying 2.14 iteratively,

$$F_X(x) = F_{X_1} \ast F_{X_2} \ast \ldots \ast F_{X_m}(x),$$

and thus:

$$C_F(t) = \prod_{i=1}^m C_{F_i}(t).$$

We next turn to the **Inverse Fourier-Stieltjes transform** which generalizes proposition 6.37 of book 5. Though the results below are stated in the context of probability distributions and characteristic functions, it is a small exercise in changing notation to see that these results remain valid in the slightly more general context of the Fourier transform of finite measures.

Recall that for a given Borel set $A \subset \mathbb{R}^n$, the **boundary of $A$**, denoted $\partial(A)$, is defined as:

$$\partial(A) \equiv \{x | x \text{ is a limit point of } A \text{ and } \overline{A}\},$$
6.5 PROPERTIES OF CHARACTERISTIC FUNCTIONS ON $\mathbb{R}^N$

where $\bar{A}$ denotes the complement of $A$. When $A = \prod_{j=1}^{n}(a_j, b_j)$ it is an exercise to check that:

$$\partial(A) = \prod_{j=1}^{n}[a_j, b_j] - \prod_{j=1}^{n}(a_j, b_j).$$

**Proposition 6.23 (Inverse Fourier-Stieltjes Transform)** Let $F$ be a distribution function defined on $\mathbb{R}^n$, $\mu_F$ the induced probability measure, and $C_F(t)$ the characteristic function as defined in 6.7. Let $A = \prod_{j=1}^{n}(a_j, b_j)$ be bounded and assume that $\mu_F[A] = 0$. With $R_T \equiv \{t \mid |t_j| \leq T \text{ all } j\}$:

$$\mu_F[A] = \lim_{T \to \infty} \frac{1}{(2\pi)^n} \int_{R_T^n} \prod_{j=1}^{n} \frac{e^{-ia_j t_j} - e^{-ib_j t_j}}{it_j} C_F(t) \, dm^n,$$

where $dm^n$ denotes $n$-dimensional Lebesgue measure.

**Proof.** By 6.7,

$$\frac{1}{(2\pi)^n} \int_{R_T^n} \prod_{j=1}^{n} \frac{e^{-ia_j t_j} - e^{-ib_j t_j}}{it_j} C_F(t) \, dm^n = \frac{1}{(2\pi)^n} \int_{R_T} \int_{\mathbb{R}^n} \prod_{j=1}^{n} \exp \left[ i (x_j - a_j) t_j \right] - \exp \left[ i (x_j - b_j) t_j \right] \, d\mu_F(x) \, dm^n(t).$$

Now:

$$\exp \left[ i (x_j - a_j) t_j \right] - \exp \left[ i (x_j - b_j) t_j \right] = \int_{(x_j - a_j) \to (x_j - b_j)} e^{iyt_j} \, dy,$$

and since $|e^{ix}| = 1$ for all $x \in \mathbb{R}$ by Euler’s formula (book 5), the absolute value of this integral is bounded by $b_j - a_j$ by the triangle inequality. Consequently the integrand above is continuous in $x_i$ and $t_i$ and bounded in absolute value by $\prod_{j=1}^{n}(b_j - a_j)$ and is thus integrable. By Fubini’s theorem of proposition 5.15 of book 5, this iterated integral which we denote $I_T$, can be reversed as follows:

$$I_T = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \left[ \int_{R_T^n} \prod_{j=1}^{n} \frac{\exp \left[ i (x_j - a_j) t_j \right] - \exp \left[ i (x_j - b_j) t_j \right]}{it_j} \, dm^n(t) \right] \, d\mu_F(x)$$

$$= \int_{\mathbb{R}^n} \prod_{j=1}^{n} \left[ \frac{1}{2\pi} \int_{-T}^{T} \frac{\exp \left[ i (x_j - a_j) t_j \right] - \exp \left[ i (x_j - b_j) t_j \right]}{it_j} \, dt_j \right] \, d\mu_F(x).$$

This inner integral is now identical with the 1-dimensional expression in the proof of proposition 6.37 of book 5. Using the same steps and notation there:

$$I_T = \int_{\mathbb{R}^n} \prod_{j=1}^{n} \left[ \frac{1}{2\pi} \int_{-\infty}^{\infty} [\text{sgn}(x_j - a_j) S(|x_j - a_j| T) - \text{sgn}(x_j - b_j) S(|x_j - b_j| T)] \, dt_j \right] \, d\mu_F(x),$$
where the sign function or signum function \( \text{sgn}(\theta) \) is defined:

\[
\text{sgn}(\theta) = \left\{ \begin{array}{ll}
1, & \theta > 0, \\
0, & \theta = 0, \\
-1, & \theta < 0,
\end{array} \right.
\]

and

\[
S(t) \equiv \int_0^t \frac{\sin x}{x} \, dx.
\]

Hence by 6.27 of book 5, again following the one-dimensional proof:

\[
\lim_{T \to \infty} I_T = \int_{\mathbb{R}^n} \prod_{j=1}^{n} \lambda_{a_j,b_j}(x_j) \, d\mu_F(x),
\]

where:

\[
\lambda_{a_j,b_j}(x) = \left\{ \begin{array}{ll}
0, & x < a_j, \\
\frac{1}{2}, & x = a_j, \\
1, & a_j < x < b_j, \\
\frac{1}{2}, & x = b_j, \\
0, & x > b_j.
\end{array} \right.
\]

Letting \( \hat{A} = \prod_{j=1}^{n}(a_j, b_j) \) and \( \overline{A} = \prod_{j=1}^{n}[a_j, b_j] \), note that:

\[
\chi_{\hat{A}}(x) \leq \prod_{j=1}^{n} \lambda_{a_j,b_j}(x_j) \leq \chi_{\overline{A}}(x),
\]

and so:

\[
\int_{\mathbb{R}^n} \chi_{\hat{A}}(x) \, d\mu_F(x) \leq \int_{\mathbb{R}^n} \prod_{j=1}^{n} \lambda_{a_j,b_j}(x_j) \, d\mu_F(x) \leq \int_{\mathbb{R}^n} \chi_{\overline{A}}(x) \, d\mu_F(x).
\]

Since \( \mu_F(\partial A) = \mu_F(\overline{A} - \hat{A}) = 0 \) by assumption, these bounding integrals agree. But then since \( \chi_{\hat{A}}(x) \leq \chi_A(x) \leq \chi_{\overline{A}}(x) \):

\[
\int_{\mathbb{R}^n} \prod_{j=1}^{n} \lambda_{a_j,b_j}(x_j) \, d\mu_F(x) = \int_{\mathbb{R}^n} \chi_{A}(x) \, d\mu_F(x) \equiv \mu_F(A).
\]

In the special case of integrable \( C_F(t) \), we can say more about the inverse transform, as was the case for \( n = 1 \) stated above.
6.5 PROPERTIES OF CHARACTERISTIC FUNCTIONS ON \( \mathbb{R}^n \)

Proposition 6.24 (Inversion of \( C_F(t) \)) If \( C_F(t) \) is Lebesgue integrable, then \( F \) has a continuous density function \( f(x) \) and so for all \( A \in \mathcal{B}(\mathbb{R}^n) \):

\[
\mu_F[A] = \int_A f(x) \, dm.
\]

(6.45)

In addition:

\[
f(x) = (2\pi)^{-n} \int_{\mathbb{R}^n} C_F(t) e^{-ix \cdot t} \, dt.
\]

(6.46)

Proof. Rather than prove the existence of \( f(x) \) that satisfies 6.46 here, we defer the proof to remark 6.43 following the proof of Bochner’s theorem in proposition 6.42 below. The first part of that theorem’s proof will accomplish this in a somewhat more general setting that will be discussed in that remark.

Given the existence of the density function \( f(x) \), define the set function \( \mu[A] \) on \( \mathcal{B}(\mathbb{R}^n) \) by:

\[
\mu[A] = \int_A f(x) \, dm.
\]

Then \( \mu[A] \) is a Borel measure by proposition 3.3 of book 5. Now if \( A \equiv \prod_{i=1}^n (-\infty, x_i] \), then

\[
\mu[A] = F(x) \equiv \mu_F[A].
\]

Thus by proposition 8.9 of book 1, \( \mu[A] = \mu_F[A] \) for all right semi-closed rectangles, \( A \equiv \prod_{i=1}^n (a_i, b_i] \), and this now extends to all \( A \in \mathcal{B}(\mathbb{R}^n) \) by the uniqueness of extensions result of proposition 6.14 of book 1 (see the next proof for more detail on this).

As in the 1-dimensional case of corollary 6.42 of book 5, we now address the question of the uniqueness of the \( n \)-dimensional Fourier-Stieltjes transform, or in the current context, the uniqueness of the characteristic function. Specifically, if \( C_F(t) = C_G(t) \) for all \( t \in \mathbb{R}^n \), where \( F, G \) are given distribution functions, must it be the case that \( \mu_F(A) = \mu_G(A) \) for all Borel sets \( A \in \mathcal{B}(\mathbb{R}^n) \)?

Proposition 6.25 (Uniqueness of the Characteristic Function) If \( \mu_F \) and \( \mu_G \) are finite Borel measures on \( \mathbb{R}^n \) and their characteristic functions satisfy \( C_F(t) = C_G(t) \) for all \( t \), then \( \mu_F(A) = \mu_G(A) \) for all Borel sets, \( A \in \mathcal{B}(\mathbb{R}^n) \).

Proof. The above proposition proves that \( \mu_F(A) = \mu_G(A) \) for all bounded right semi-closed rectangles \( A = \prod_{j=1}^n (a_j, b_j] \) with \( \mu_F[\partial A] = \mu_G[\partial A] = 0 \). So the essence of the proof of this general result is to extend this identity to the semi-algebra of all right semi-closed rectangles, denoted \( \mathcal{A}' \) in book 1,
without any restriction on the measures of the boundaries. This result then extends naturally from \( \mathcal{A}' \) to the associated algebra \( \mathcal{A} \) by the Carathéodory extension theorem II of proposition 6.13 of book 1, and then to the smallest sigma algebra \( \sigma(\mathcal{A}) \) that contains \( \mathcal{A} \) by the uniqueness of extensions result of that book’s proposition 6.14. But \( \sigma(\mathcal{A}) \) must then contain the open sets, since \( \prod_{j=1}^{\infty} (a_j, b_j) = \bigcup_k \prod_{j=1}^{\infty} (a_j, b_j - r_k) \) for rational \( r_k \to 0 \). It thus follows by definition 2.13 of book 1 that \( \sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R}^n) \), completing the proof.

To prove the result for \( \mathcal{A}' \), recall part a of the proof of the multivariate Portmanteau theorem, proposition 4.11. There it was demonstrated that there is a dense and uncountable set of reals, \( D_F \subset \mathbb{R} \), with countable complement, so that for any \( k \), \( \mu_F([x \in \mathbb{R}^n | x_k = r]) = 0 \) for any \( r \in D_F \). By defining \( D = D_F \cap D_G \) with \( D_G \) analogously defined, a single dense set of reals is produced that satisfies this property for both measures. It is worth a moment of thought to verify this last statement. By this proof, \( D_F = \mathbb{R} - C_F \) with \( C_F \) countable, and similarly \( D_G = \mathbb{R} - C_G \). Using De Morgan's laws, \( D_F \cap D_G = \mathbb{R} - (C_F \cup C_G) \), which is of necessity dense and uncountable, and with countable complement.

Define the class of rectangles \( \mathcal{R} \) by \( \mathcal{R} = \{ \prod_{j=1}^{\infty} (a_j, b_j) | a_j, b_j \in D \} \), and recall that from the earlier noted proof that every vertex of such a rectangle is now a continuity point of both distribution functions \( F \) and \( G \). Since the bounding hyperplanes of such rectangles have both \( \mu_F \)-measure and \( \mu_G \)-measure 0, so too do the boundaries of all such rectangles. Thus by the previous proposition, \( \mu_F(A) = \mu_G(A) \) for all \( A \in \mathcal{R} \).

To extend this identity from \( \mathcal{R} \) to \( \mathcal{A}' \), let \( A = \prod_{j=1}^{\infty} (a_j, b_j) \) with \( b_j \in D \) and \( a_j \) finite but arbitrary, and for each \( j \) let \( \{a_j^{(m)} \}_{m=1}^{\infty} \subset D \) be a decreasing sequence with \( a_j^{(m)} \to a_j \). Then \( A^{(m)} = \prod_{j=1}^{\infty} (a_j^{(m)}, b_j) \in \mathcal{R} \) for all \( m \) so \( \mu_F(A^{(m)}) = \mu_G(A^{(m)}) \). Also, \( \{A^{(m)} \}_{m=1}^{\infty} \) is nested, \( A^{(m)} \subset A^{(m+1)} \), and since \( A = \bigcup_{m=1}^{\infty} A^{(m)} \) it follows by continuity from below (proposition 2.44, book 1, and comment below remark 2.45) that \( \mu_F(A) = \mu_G(A) \). Now define \( A = \prod_{j=1}^{\infty} (\bar{a}_j, \bar{b}_j) \) with \( \bar{b}_j \) and \( \bar{a}_j \) arbitrary, and for each \( j \) let \( \{b_j^{(m)} \}_{m=1}^{\infty} \subset D \) be a decreasing sequence with \( b_j^{(m)} \to \bar{b}_j \). Then by the previous result, if \( A^{(m)} = \prod_{j=1}^{\infty} (\bar{a}_j, b_j^{(m)}) \) then \( \mu_F(A^{(m)}) = \mu_G(A^{(m)}) \). Also, \( \{A^{(m)} \}_{m=1}^{\infty} \) is nested, \( A^{(m+1)} \subset A^{(m)} \), and since \( A = \bigcap_{m=1}^{\infty} A^{(m)} \) it follows by continuity from above (same reference) that \( \mu_F(A) = \mu_G(A) \). Hence \( \mu_F(A) = \mu_G(A) \) for all bounded \( A \in \mathcal{A}' \), and since a rectangle \( A = \prod_{j=1}^{\infty} (-\infty, b_j) \) can be decomposed into countably many disjoint bounded rectangles, the result follows by countable additivity of measures. Thus \( \mu_F(A) = \mu_G(A) \) for all \( A \in \mathcal{A}' \).
6.5 PROPERTIES OF CHARACTERISTIC FUNCTIONS ON $\mathbb{R}^N$ 157

The following definition can be compared with definition 3.6 and 6.23. Recall that a matrix $C$ is **positive semidefinite** if $x^T C x \geq 0$ for all $x$. Note that this is a weaker condition than that for **positive definite** which requires that $x^T C x = 0$ if and only if $x = 0$. This definition is more general than definition 3.6 since characteristic functions always exist. However it is consistent with the earlier definition since if such $Y$ also has a moment generating function, then 6.42 assures that 3.8 implies 6.47.

**Definition 6.26** A random vector $Y \equiv (Y_1, ..., Y_n)$ is said to have a **multivariate normal distribution** or a **multivariate Gaussian distribution** if there exists an $n \times n$ symmetric, positive semidefinite matrix $C$ and an $n$-vector so that the associated characteristic function has the form

$$C_Y(t) = \exp \left[ i \mu \cdot t - \frac{1}{2} t^T Ct \right]. \quad (6.47)$$

As was proved in proposition 3.9, it is an exercise to show that:

**Exercise 6.27** If $Y \equiv (Y_1, Y_2, ..., Y_n)$ is multivariate normally distributed by definition 6.26, and $B : \mathbb{R}^n \to \mathbb{R}^m$ a linear transformation, then $BY$ is multivariate normally distributed with $\mu_B = B\mu$ and $C_B = BCB^T$.

**Example 6.28** Since $C_Y(t)$ in 6.47 is infinitely differentiable, 6.40 can be applied to show that $C$ is the **covariance matrix** of $(Y_1, Y_2, ..., Y_n)$ defined by $C_{ij} = E[(Y_i - \mu_i)(Y_j - \mu_j)]$, and $\mu = (\mu_1, \mu_2, ..., \mu_n)$ is the vector of first moments, $\mu_j = E[Y_j]$. Given such symmetric $C$, it is always possible to find a matrix $A$ so that $C = AA^T$. When $C$ is positive semidefinite, $A$ can be found using the **eigendecomposition** of $C$:

$$C = QA\Lambda Q^T,$$

where $Q$ is an orthogonal matrix with columns equal to the eigenvectors of $C$, and $\Lambda$ is a diagonal matrix with the eigenvalues of $C$ on the diagonal, and hence $\Lambda_{jj} \geq 0$. Thus $A \equiv Q\sqrt{\Lambda}$ with obvious notation, and $A$ must of necessity be singular unless $C$ is positive definite and $\Lambda_{jj} > 0$ for all $j$.

Given such $A$, define $Y' = AX + \mu$ for $X$ an $n$-vector of independent unit normals, and observe that by proposition 3.1 and exercise 3.4, $Y'$ has the same characteristic function as this originally defined variate $Y$, and further, $C$ is the covariance matrix of $Y'$ and $\mu$ its mean vector. By the uniqueness result of proposition 6.25, it follows that $Y'$ has the same distribution as the original $Y$.

Thus, as in the special case where $Y$ has a density function, there are two corollaries of this uniqueness result.
1. A multivariate normal distribution is uniquely determined by its covariance matrix $C$ and mean vector $\mu$.

2. If $Y \equiv (Y_1, Y_2, ..., Y_n)$ is multivariate normally distributed, then $\{Y_i\}_{i=1}^n$ are independent if and only if $C_{ij} = 0$ for $i \neq j$. In other words, component normal variates of a multivariate normal distribution are independent if and only if they are uncorrelated.

Exercise 6.29 As noted in remark 3.5, the reader is invited to prove proposition 3.9 using characteristic functions rather than moment generating functions, with the last step then relying on proposition 6.25 above.

The final investigation relates to the extension of the continuity theorem to this multivariate context. While half of the proof is easy based on earlier results, the other half requires a technical result. As noted in remark 6.17, this continuity result and its corollary combine to produce Lévy’s continuity theorem, named for Paul Lévy (1886 –1971).

For the next result, recall that $F_m \Rightarrow F$ and $\mu_m \Rightarrow \mu$ denote weak convergence as defined in definition 4.6.

Proposition 6.30 (Lévy’s continuity theorem) Let $\{F_m\}_{m=1}^\infty, F$ be a collection of distribution functions on $\mathbb{R}^n$ with associated probability measures $\{\mu_m\}_{m=1}^\infty, \mu$ and characteristic functions $\{C_{F_m}\}_{m=1}^\infty, C_F$. Then $F_m \Rightarrow F$ if and only if $C_{F_m}(t) \rightarrow C_F(t)$ for all $t$.

Proof. One direction of this result is an immediate consequence of the portmanteau theorem of proposition 4.11. Specifically, by part 5 of this proposition, $\mu_m \Rightarrow \mu$ if and only if $\int g(x)d\mu_m \rightarrow \int g(x)d\mu$ for every bounded, continuous real-valued function $g$ defined on $\mathbb{R}^n$. But for any $t \in \mathbb{R}^n$:

$$e^{ix \cdot t} = (\cos(x \cdot t) + i \sin(x \cdot t)) = g_1(x) + ig_2(x),$$

with $g_1(x)$ and $g_2(x)$ continuous and bounded real valued functions, and so $C_{F_m}(t) \rightarrow C_F(t)$ for all $t$.

The other direction requires Prokhorov’s theorem of proposition 4.41 and the above characteristic function results on $\mathbb{R}$. If $C_{F_m}(t) \rightarrow C_F(t)$ for all $t$, define $e_k \in \mathbb{R}^n$ to have 1 in the $k$th component and 0s elsewhere. Then with $s$ real, and with apparent notation:

$$C_{F_m}(se_k) = \int_{\mathbb{R}^n} e^{ix_k \cdot s} dF_m(x) = \int_{\mathbb{R}^n} e^{ix_k \cdot s} \int_{\mathbb{R}^{n-1}} dF_m(x) = \int_{\mathbb{R}^n} e^{ix_k \cdot s} dF_{m(k)}(x_k),$$

where $F_{m(k)}$ is the $k$th marginal distribution function of $F_m$. Thus $C_{F_m}(se_k) = C_{F_{m(k)}}(s)$, the characteristic function of the marginal distribution function...
$F_{m(k)}$ defined on $\mathbb{R}$, and it follows by hypothesis that as $m \to \infty$, $C_{F_{m(k)}}(s) \to C_{F_{k}}(s)$ for all $s$. By Lévy’s continuity theorem on $\mathbb{R}$, $F_{m(k)} \Rightarrow F_{k}$ for all $k$.

In other words, each of the $n$ one-variate marginal distributions of the sequence $\{F_{m}\}_{m=1}^{\infty}$ converges weakly to the respective marginal distributions of $F$. By proposition 8.18 of book 2, each such sequence of marginal distributions $\{F_{m(k)}\}_{m=1}^{\infty}$ is tight. So given $\epsilon > 0$, for each $k$ there is a $T_k$ so that

$$
\mu_{m(k)}([T_k, T_k]) > 1 - \epsilon/n
$$

for all $m$. Equivalently, for each $k$ there is a $T_k$ so that for all $m$,

$$
\mu_{m}([x \in \mathbb{R}^n | T_k < x_k \leq T_k]) > 1 - \epsilon/n.
$$

Let $T = \max\{T_k\}$ and note that:

$$
\mathbb{R}^n - (-T, T)^n \subset \bigcup_{k=1}^{n} \{x | x_k \leq -T_k \text{ or } x_k > T_k\},
$$

and each set in this union has $\mu_{m}$-measure less than $\epsilon/n$. It then follows by subadditivity of measures that for all $m$, $\mu_{m}[\mathbb{R}^n - (-T, T)^n] < \epsilon$ and thus $\mu_{m}([(-T, T)^n]) > 1 - \epsilon$. That is, $\{\mu_{m}\}_{m=1}^{\infty}$ is tight, and we apply Prokhorov’s theorem of proposition 4.41 and corollary 4.42.

Given any weakly convergent subsequence $\{\mu_{m_k}\}_{k=1}^{\infty}$ and limit probability measure $\nu$, noting that such exists by Prokhorov’s theorem, the conclusion that $\mu_{m_k} \Rightarrow \nu$ assures by the first part of this proposition that $C_{\mu_{m_k}}(t) \to C_{\nu}(t)$ for all $t$. But then by hypothesis and a change of notation from distribution functions to probability measures, it is the case that $C_{\mu}(t) \to C_{\nu}(t)$ and thus also $C_{\mu_{m_k}}(t) \to C_{\mu}(t)$. Hence $C_{\nu}(t) = C_{\mu}(t)$, and by the above uniqueness theorem, $\nu = \mu$. Hence, every weakly convergent subsequence of $\{\mu_{m}\}_{m=1}^{\infty}$ converges to $\mu$, and an application of corollary 4.42 yields that $\mu_{m} \Rightarrow \mu$.

**Corollary 6.31 (Lévy’s continuity theorem)** Let $\{F_{m}\}_{m=1}^{\infty}$ be a collection of distribution functions on $\mathbb{R}^n$ with associated probability measures $\{\mu_{m}\}_{m=1}^{\infty}$ and characteristic functions $\{C_{F_{m}}\}_{m=1}^{\infty}$. If $C_{F_{m}}(t) \to \varphi(t)$ for all $t$ with $\varphi$ continuous at $t = 0$, then there exists a probability measure $\mu$ and associated distribution function $F$ so that $C_{F}(t) = \varphi(t)$ and $F_{m} \Rightarrow F$.

**Proof.** As in the proof of the above theorem, it follows that for every $k$ the characteristic function of the $k$th marginal distribution converges, $C_{F_{m(k)}}(s) \to \varphi_{k}(s)$ for all $s$, where $\varphi_{k}(s) \equiv \varphi(0, \ldots, s, 0, \ldots, 0)$ with $s$ in the $k$th component. Since $\varphi_{k}(s)$ is continuous at $s = 0$ by assumption, corollary 6.55 of book 5 obtains the existence of a probability measure $\mu_{(k)}$ on $\mathbb{R}$.
so that $\mu_{m(k)} \Rightarrow \mu(k)$ where $\mu_{m(k)}$ is the probability measure associated with $F_{m(k)}$, and $\varphi_k(s)$ is the characteristic function of $\mu(k)$. By proposition 8.18 of book 2, each such sequence of marginal distributions $\{F_{m(k)}\}_{k=1}^\infty$ is tight, and using the same steps as will above conclude that $\{\mu_{m(k)}\}_{m=1}^\infty$ is tight.

The last paragraph of the prior result now changes only a little. Given any weakly convergent subsequence $\{\mu_{m(k)}\}_{k=1}^\infty$ and limit probability measure $\nu$, noting that such exists by Prokhorov’s theorem, the conclusion that $\mu_{m(k)} \Rightarrow \nu$ assures by the first part of the above proposition that $C_{\mu_{m(k)}}(t) \to C_{\nu}(t)$ for all $t$. But then by hypothesis and a change of notation from distribution functions to probability measures, it is the case that $C_{\mu_{m(k)}}(t) \to \varphi(t)$ and thus also $C_{\mu_{m(k)}}(t) \to C_{\nu}(t) = \varphi(t)$, and $\varphi(t)$ is the characteristic function of $\nu$. Thus every weakly convergent subsequence of $\{\mu_{m(k)}\}_{m=1}^\infty$ converges to a measure with characteristic function $\varphi(t)$, and so by uniqueness this measure must always be $\nu$. An application of corollary 4.42 yields that $\mu_m \Rightarrow \nu$. ■

6.5.1 The Cramér-Wold Device

The Cramér-Wold device, also called the Cramér-Wold theorem, was introduced in proposition 4.22 in the above section, General Results on Weak Convergence of Measures. It is named for a 1936 result of Harald Cramér (1893 – 1985) and Herman Wold (1908 – 1992), and addresses the following question. If $\{X_m\}_{m=1}^\infty$, $X$, are random vectors defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with range in $\mathbb{R}^n$, what is the relationship between $X_m \Rightarrow X$ and $t \cdot X_m \Rightarrow t \cdot X$ for $t \in \mathbb{R}^n$? Here the dot product of $n$-vectors is defined as usual by $x \cdot y = \sum_{k=1}^n x_k y_k$.

The first convergence $X_m \Rightarrow X$ is a statement about weak convergence of the associated joint distributions, while the latter $t \cdot X_m \Rightarrow t \cdot X$ is a statement about the weak convergence of one-dimensional distributions defined relative to half-spaces in $\mathbb{R}^n$. This follows because given $t$, if $X'_m \equiv t \cdot X_m$ and $X' \equiv t \cdot X$, then since $\{X'_m \leq \alpha\} = \{t \cdot X_m \leq \alpha\}$, the distribution function of $X'_m$ is defined by:

$$F_{X'_m}(\alpha) = \lambda \left[ (X'_m)^{-1} (\alpha) \right] = \lambda \left[ (X_m)^{-1} (H_t(\alpha)) \right],$$

where $H_t(\alpha) \subset \mathbb{R}^n$ is the half space defined as $H_t(\alpha) \equiv \{t \cdot x \leq \alpha\}$. Hence given $t \in \mathbb{R}^n$, knowing that $X'_m \Rightarrow X'$ provides a highly summarized statement about the random vectors $X_m$ and $X$. Specifically, if $\alpha$ is a continuity point of $F_{X'}$, then $F_{X'_m}(\alpha) \to F_{X'}(\alpha)$, which to say only that $\lambda \left[ (X_m)^{-1} (H_t(\alpha)) \right] \to \lambda \left[ (X)^{-1} (H_t(\alpha)) \right].$
6.5 PROPERTIES OF CHARACTERISTIC FUNCTIONS ON $\mathbb{R}^N$

The Cramér-Wold theorem states that $X_m \Rightarrow X$ if and only if $t \cdot X_m \Rightarrow t \cdot X$ for all $t \in \mathbb{R}$. Proposition 4.22 provided the simpler half of the proof, the "only if" result. The deeper half of the proof requires the powerful tools of Fourier analysis, here interpreted in the context of characteristic functions. That results on one dimensional "marginal" distributions imply results in the multivariate context has already been seen in the various proofs of the prior section.

Proposition 6.32 (Cramér-Wold theorem) Let $\{X_m\}_{m=1}^{\infty}$ be random vectors on $(S, \mathcal{E}, \lambda)$ with range in $\mathbb{R}^n$. Then $X_m \Rightarrow X$ if and only if $t \cdot X_m \Rightarrow t \cdot X$ for all $t \in \mathbb{R}^n$.

Proof. That $X_m \Rightarrow X$ assures $t \cdot X_m \Rightarrow t \cdot X$ for all $t \in \mathbb{R}^n$ was proved in proposition 4.22 as a direct application of the mapping theorem on $\mathbb{R}^j / \mathbb{R}^k$. To prove the converse, we use Lévy’s continuity theorem in one dimensional. For given $t \in \mathbb{R}^n$, define random variables $X'_t \equiv t \cdot X_m$ and $X' \equiv t \cdot X$, then $X'_t \Rightarrow X'$ by assumption. Lévy’s continuity theorem then obtains that $C_{X'_t}(s) \to C_{X'}(s)$ for all $s$, where to simplify notation $C_{X'_t}(s)$ denotes the characteristic function of the distribution function $F_{X'_t}$ defined above, and similarly for $C_{X'}(s)$. By definition:

$$C_{X'_t}(s) \equiv \int_{-\infty}^{\infty} e^{iys} dF_{X'_t}(y),$$

with $C_{X'}(s)$ similarly defined. Thus for any $t \in \mathbb{R}^n$:

$$\int_{-\infty}^{\infty} e^{iys} dF_{X'_t}(y) \to \int_{-\infty}^{\infty} e^{iys} dF_{X'}(y),$$

for all $s$.

We now set up a change of variables. For $t \in \mathbb{R}^n$ define $T_t: \mathbb{R}^n \to \mathbb{R}$ by $T_t: x \to t \cdot x$, noting that $T_t$ is continuous and thus Borel measurable. Let $\mu_{X_m}$ denote the probability measure associated with $F_{X_m}$, and similarly define $\mu_X$, $\mu_{X'_t}$, and $\mu_{X'}$. By definition 3.9 of book 5, $\mu_{X_m}$ and the transformation $T_t$ induce a measure $\left(\mu_{X_m}\right)_{T_t}$ on the range space $\mathbb{R}$ defined on $A \in \mathcal{B}(\mathbb{R})$ by:

$$(\mu_{X_m})_{T_t}(A) \equiv \mu_{X_m}\left[T_t^{-1}(A)\right].$$

The distribution function on $\mathbb{R}$ induced by the measure $\left(\mu_{X_m}\right)_{T_t}$ is exactly $F_{X'_t}$ as can be seen by letting $A = (-\infty, a]$. Since $T_t^{-1}(A) = \{x \in \mathbb{R}^n | x \cdot t \leq a\}$.
CHAPTER 6 THE CHARACTERISTIC FUNCTION

\[ (\mu_{X_m})_{T_t}((-\infty,a]) \equiv \mu_{X_m}[T_t^{-1}(A)] \]
\[ = \lambda[X_m \cdot t \leq a] \]
\[ \equiv F_{X'_m}(a). \]

Thus by proposition 3.14 of book 5:

\[ C_{X'_m}(s) \equiv \int_{-\infty}^{\infty} e^{isy} \, d\mu_{X'_m}(y) = \int_{\mathbb{R}^n} e^{i(t \cdot x)s} \, d\mu_{X_m}(x), \]

and analogously:

\[ C_{X'_m}(s) \equiv \int_{-\infty}^{\infty} e^{isy} \, d\mu_{X'_m}(y) = \int_{\mathbb{R}^n} e^{i(t \cdot x)s} \, d\mu_{X}(x). \]

Thus \( C_{X'_m}(s) \to C_{X}(s) \) for all \( s \in \mathbb{R} \) and all \( t \in \mathbb{R}^n \) can be restated in Lebesgue-Stieltjes notation and assures that:

\[ \int_{\mathbb{R}^n} e^{i(t \cdot x)} \, d\mu_{X_m} \to \int_{\mathbb{R}^n} e^{i(t \cdot x)} \, d\mu_{X}. \]

By definition \( C_{X_m}(t) \to C_{X}(t) \) for all \( t \in \mathbb{R}^n \), and by Lévy’s continuity theorem this is equivalent to \( X_m \Rightarrow X \). \( \blacksquare \)

Remark 6.33 For an application of this result, see the section below on the \( n \)-dimensional central limit theorem.

6.6 Bochner’s Theorem

Bochner’s theorem is named for Salomon Bochner (1899 – 1982). Published in German in two papers in 1932 and 1933, this theorem provides a necessary and sufficient condition on a complex valued function \( \varphi : \mathbb{R}^n \to \mathbb{C} \) to ensure that there exists a probability measure \( \mu \), or equivalently a distribution function \( F \), so that \( \varphi(t) = C_F(t) \). His first paper developed the result for \( n = 1 \), and this was then generalized in the second paper.

Certainly any such \( \varphi \) must possess the properties common to all characteristic functions. Looking at proposition 6.11, the only general property identified there for all \( C_F(t) \) was stated in 1, and thus:
6.6 BOCHNER’S THEOREM

1. If $\varphi : \mathbb{R}^n \to \mathbb{C}$ is the characteristic function of a distribution function $F$ defined on $\mathbb{R}^n$, then $\varphi$ must be uniformly continuous on $\mathbb{R}^n$ and $\varphi(0) = 1$.

But clearly there is more. For example, since $|e^{ix \cdot t}| = 1$ it follows by the triangle inequality that:

$$|C_F(t)| \leq \int_{\mathbb{R}^n} dF(x) = 1.$$ 

Hence:

2. If $\varphi : \mathbb{R}^n \to \mathbb{C}$ is the characteristic function of a distribution function $F$ defined on $\mathbb{R}^n$, then $\varphi$ must satisfy $|\varphi(t)| \leq 1$ for all $t$.

Also, there is a certain symmetry in the value of characteristic functions on $\pm t$. But first a definition.

**Definition 6.34** The **complex conjugate** of a complex number $z \equiv a + bi$, denoted $\bar{z}$, is defined by:

$$\bar{z} = a - bi. \quad (6.48)$$

One also calls $a$ the **real part** of $z$, and $b$ the **imaginary part** of $z$, denoted:

$$a \equiv \text{Re}(z) = (z + \bar{z})/2, \quad b \equiv \text{Im}(z) = (z - \bar{z})/2i. \quad (6.49)$$

Also:

$$|z|^2 \equiv z\bar{z} = a^2 + b^2.$$ 

By Euler’s formula (6.12 of book 5) one obtains:

$$\overline{(e^{ix \cdot t})} = e^{-ix \cdot t},$$

and thus:

$$C_F(-t) = \int_{\mathbb{R}^n} e^{-ix \cdot t} dF(x) = \int_{\mathbb{R}^n} e^{ix \cdot t} dF(x) = \overline{C_F(t)}.$$ 

Hence:

3. If $\varphi : \mathbb{R}^n \to \mathbb{C}$ is the characteristic function of a distribution function $F$ defined on $\mathbb{R}^n$, then $\varphi(-t) = \overline{\varphi(t)}$ for all $t$. 

CHAPTER 6 THE CHARACTERISTIC FUNCTION

The final property of $C_F(t)$ that can be verified is perhaps a surprising one, and not because it will be difficult to derive. The surprise is that it is not a property that has been observed or indeed utilized in any of the above results, and hence it is somewhat surprising that it was identified. First a definition:

**Definition 6.35** A complex valued function $\varphi : \mathbb{R}^n \to \mathbb{C}$ is said to be *positive semidefinite* (sometimes *nonnegative definite*) if for an $m$, any $\{t_j\}_{j=1}^m \subset \mathbb{R}^n$, and any $\{z_j\}_{j=1}^m \subset \mathbb{C}$:

$$\sum_{j=1}^m \sum_{k=1}^m \varphi(t_j - t_k)z_j \bar{z}_k \geq 0.$$  \hspace{1cm} (6.50)

Implicit in 6.50 is that the double summation is real valued.

**Remark 6.36** Note that 6.50 can be expressed in matrix notation as:

$$z^T \Phi z \geq 0,$$

where $\Phi$ is an $m \times m$ matrix with $\Phi_{jk} = \varphi(t_j - t_k)$, $z \equiv (z_1, ..., z_m)$ and $\bar{z} \equiv (\bar{z}_1, ..., \bar{z}_m)$. Thus the terminology "positive semidefinite" is reminiscent of that used for a matrix $C$ in the discussion on the multivariate normal distribution. But there is an important difference between these two notions. and in fact, in some sources one will see the terminology that $\varphi : \mathbb{R}^n \to \mathbb{C}$ is said to be *positive definite*.

To clarify, recall that $C$ in the earlier discussion was a real matrix, while $\Phi$ is complex. Also, the earlier expression $x^T C x \geq 0$ was a statement about $x \in \mathbb{R}^m$, and for which it is apparent that $x^T C x$ is a real number. Above, $z^T \Phi z \geq 0$ is a statement about $z \in \mathbb{C}^m$, and first of all there is no reason this value should be real, since in general it will be complex. So an important part of the above definition is that $z^T \Phi z$ is always a real number. Proposition 6.39 identifies other properties of such $\varphi$. But first a result on characteristic functions.

**Proposition 6.37** If $F$ is a distribution function on $\mathbb{R}^n$, then the characteristic function $C_F$ is positive semidefinite.
6.6 BOCHNER’S THEOREM

Proof. With the notation above:

\[ \sum_{j=1}^{m} \sum_{k=1}^{m} C_F(t_j - t_k) z_j \bar{z}_k = \int_{\mathbb{R}^n} \sum_{j=1}^{m} \sum_{k=1}^{m} e^{ix(t_j - t_k)} z_j \bar{z}_k dF(x) \]
\[ = \int_{\mathbb{R}^n} \sum_{j=1}^{m} \sum_{k=1}^{m} e^{ix(t_j - t_k)} z_j e^{-ix(t_j)} \bar{z}_k dF(x) \]
\[ = \int_{\mathbb{R}^n} \sum_{j=1}^{m} e^{ix(t_j - t_k)} \sum_{k=1}^{m} e^{ix(t_k)} z_k dF(x) \]
\[ = \int_{\mathbb{R}^n} \left| \sum_{j=1}^{m} e^{ix(t_j - t_k)} z_j \right|^2 dF(x) \geq 0. \]

This now provides a fourth property of characteristic functions.

4. If \( \varphi : \mathbb{R}^n \to \mathbb{C} \) is the characteristic function of a distribution function \( F \) defined on \( \mathbb{R}^n \), then \( \varphi \) is positive semidefinite.

Before we state and prove Bochner’s theorem, we develop some additional properties of positive semidefinite functions.

Remark 6.38 A complex matrix \( \Phi \) that satisfies 6.51 below is said to be a Hermitian matrix, and this identity is sometimes expressed:

\[ \Phi = \Phi^T. \]

In other words, a Hermitian matrix \( \Phi \) equals its conjugate transpose. It was named for Charles Hermite (1822 – 1901) who proved that as is the case for real symmetric matrices where \( C = C^T \), such matrices have only real eigenvalues. Another property of Hermitian matrices, and one used in the proof of 4 below, is that the determinant of a Hermitian matrix is real and nonnegative: \( \det \Phi \geq 0 \).

Proposition 6.39 If \( \varphi : \mathbb{R}^n \to \mathbb{C} \) is positive semidefinite then:

1. Given \( \{t_j\}_{j=1}^{m} \subset \mathbb{R}^n \) and matrix \( \Phi \) defined above:

\[ \bar{\Phi} = \Phi^T, \quad \text{(6.51)} \]

where \( \Phi^T \) denotes the transpose of the matrix \( \Phi \). In other words, \( \bar{\Phi}_{jk} = \Phi_{kj} \), or as a statement about \( \varphi \):

\[ \bar{\varphi}(t) = \varphi(-t), \quad \text{(6.52)} \]

for all \( t = (t_1, ..., t_m) \).
2. \( \varphi(0) \) is real and \( \varphi(0) \geq 0 \).

3. \( |\varphi(t)| \leq \varphi(0) \) for all \( t \).

4. If continuous at \( t = 0 \), then \( \varphi(t) \) is uniformly continuous on \( \mathbb{R}^n \).

**Proof.** For 1, since \( z^T \Phi \bar{z} \in \mathbb{R} \) for all \( z = (z_1, \ldots, z_m) \) as noted above, it follows that \( (z^T \Phi \bar{z})^T = z^T \Phi \bar{z} \) and \( (z^T \Phi \bar{z}) = z^T \Phi \bar{z} \). Now formally:

\[
(z^T \Phi \bar{z})^T = z^T \Phi^T z,
\]

while

\[
(z^T \Phi \bar{z}) = z^T \Phi z.
\]

Thus a consequence of \( z^T \Phi \bar{z} \in \mathbb{R} \) is that \( z^T \Phi^T z = z^T \Phi z \) for all \( z \in \mathbb{C}^m \). It is an exercise to verify that this implies 6.51. Hint: Consider \( z = e_k \in \mathbb{R}^n \) to have 1 in the \( k \)th component and 0s elsewhere, as well as \( z = e_j + e_k \).

In terms of the components of \( \Phi \) and \( \Phi^T \), 6.51 states that:

\[
\varphi(t_j - t_k) = \varphi(t_k - t_j),
\]

which is 6.52, and letting \( t = 0 \) proves that \( \varphi(0) \) is real. Next recall that for any complex number \( w = a + bi \) that \( \bar{w} = a^2 + b^2 \equiv |w|^2 \). Letting \( m = 2 \), by 6.50 and 6.52 one obtains that for all \( \{z_j\}_{j=1}^2 \subset \mathbb{C} \) and all \( t \equiv t_2 - t_1 \in \mathbb{R}^n \) that:

\[
|z_1|^2 + |z_2|^2 |\varphi(0) + z_1 \bar{z}_2 \varphi(-t) + \bar{z}_1 z_2 \varphi(t)| \geq 0.
\]

(\( \ast \))

Letting \( z_1 = 0 \) and \( z_2 \neq 0 \) obtains \( \varphi(0) \geq 0 \), proving 2.

If \( \varphi(t) = 0 \) for all \( t \) then 3 is satisfied, so assume there exists \( t \in \mathbb{R}^n \) with \( \varphi(t) \neq 0 \) and then let \( z_1 = \bar{z}_2 = \sqrt{-\varphi(t)} \). If \( -\varphi(t) = a + bi \) we take the principal square root, defined as:

\[
\sqrt{-\varphi(t)} = \sqrt{\left(\sqrt{a^2 + b^2} + a\right)/2} + isgn(b)\sqrt{\left(\sqrt{a^2 + b^2} - a\right)/2}.
\]

Here we take the positive square root of real numbers and define \( sgn(b) = 1 \) if \( b \geq 0 \) and \(-1 \) otherwise. A little calculation obtains that

\[
|z_1|^2 = |z_2|^2 = |\varphi(t)|, \quad z_1 \bar{z}_2 = -\varphi(t), \quad \bar{z}_1 z_2 = -\varphi(t),
\]

and then from (\( \ast \)) and 6.52:

\[
2 |\varphi(t)| \varphi(0) - 2 |\varphi(t)|^2 \geq 0.
\]
Since $|\varphi(t)| > 0$ by assumption, 3 follows.

For 4 let $m = 3$, and $t = (t_1, t_2, t_3) \equiv (t, s, 0)$. Using 6.52:

$$
\Phi = \begin{pmatrix}
\varphi(0) & \varphi(t-s) & \varphi(t) \\
\varphi(t-s) & \varphi(0) & \varphi(s) \\
\varphi(t) & \varphi(s) & \varphi(0)
\end{pmatrix}.
$$

As noted above $\det \Phi \geq 0$, and we evaluate this determinant using the diagonal method. Recall that this means we augment the matrix:

$$
\Phi_{Aug} = \begin{pmatrix}
\varphi(0) & \varphi(t-s) & \varphi(t) & \varphi(0) & \varphi(t-s) \\
\varphi(t-s) & \varphi(0) & \varphi(s) & \varphi(t-s) & \varphi(0) \\
\varphi(t) & \varphi(s) & \varphi(0) & \varphi(t) & \varphi(s)
\end{pmatrix},
$$

and then the determinant equals the sum of the three down-diagonal products, less the sum of the three up-diagonal products.

$$
\det \Phi = \varphi(0)^3 + \varphi(t-s)\varphi(s)\varphi(t) + \varphi(t-s)\varphi(s)\varphi(t) - \varphi(0) \left[ |\varphi(t)|^2 + |\varphi(s)|^2 + |\varphi(t-s)|^2 \right].
$$

Using 6.49 and $|z|^2 + |w|^2 = |z - w|^2 + 2\Re[\bar{z}w]$ obtains:

$$
\det \Phi = \varphi(0)^3 - \varphi(0) \left[ |\varphi(t) - \varphi(s)|^2 + |\varphi(t-s)|^2 \right] - 2\Re \left[ \varphi(s)\varphi(t)(\varphi(0) - \varphi(t-s)) \right]
\leq \varphi(0)^3 - \varphi(0) \left[ |\varphi(t) - \varphi(s)|^2 + |\varphi(t-s)|^2 \right] + 2\varphi(0)^2 |\varphi(0) - \varphi(t-s)|,
$$

where in the last step we use $-\Re(\bar{z}w) \leq |z||w|$, then 3.

If $\varphi(0) = 0$ then $\varphi(t) = 0$ for all $t$ by 3 and thus 4 is assured, so assuming $\varphi(0) > 0$, and recalling that determinant of $\Phi$ is nonnegative as noted above obtains:

$$
0 \leq \varphi(0)^2 \left[ |\varphi(0) - \varphi(t-s)|^2 + |\varphi(t-s)|^2 \right] + 2\varphi(0) |\varphi(0) - \varphi(t-s)|.
$$

Thus:

$$
|\varphi(t) - \varphi(s)|^2 \leq \varphi(0)^2 - |\varphi(t-s)|^2 + 2\varphi(0) |\varphi(0) - \varphi(t-s)|
\leq 2\varphi(0) |\varphi(0) - \varphi(t-s)|.
$$

since $|\varphi(t-s)|^2 \leq \varphi(0)^2$ by 3. Hence continuity at 0 yields uniform continuity on $\mathbb{R}^n$. ■
In summary, the above proposition assures that if \( \varphi : \mathbb{R}^n \to \mathbb{C} \) is positive semidefinite, \( \varphi(0) = 1 \) and \( \varphi \) is continuous at \( t = 0 \), then this function has all of the properties of a characteristic function noted in 1 – 4 above. Bochner's theorem states that in fact this is enough to ensure that such \( \varphi \) is the characteristic function of a probability measure on \( \mathbb{R}^n \).

Before stating and proving this result, we require an additional result on positive semidefinite functions.

**Proposition 6.40** If \( \varphi : \mathbb{R}^n \to \mathbb{C} \) is positive semidefinite then:

1. If \( y \in \mathbb{R}^n \) then \( \varphi_y(t) = \varphi(t) e^{it \cdot y} \) is positive semidefinite.
2. If \( h : \mathbb{R}^n \to \mathbb{R} \) is nonnegative, continuous and Riemann integrable, then
   
   \[ \psi(t) \equiv \int_{\mathbb{R}^n} h(y) \varphi_y(t) \, dy \]

   is positive semidefinite.

**Proof.** Let \( m, \{t_j\}_{j=1}^m \subset \mathbb{R}^n \) and \( \{z_j\}_{j=1}^m \subset \mathbb{C} \) be given. For 1:

\[
\sum_{j=1}^m \sum_{k=1}^m \varphi_y(t_j - t_k) z_j z_k = \sum_{j=1}^m \sum_{k=1}^m \varphi(t_j - t_k) e^{i(t_j - t_k) \cdot y} z_j z_k \\
= \sum_{j=1}^m \sum_{k=1}^m \varphi(t_j - t_k) e^{it_j \cdot y} z_j e^{-it_k \cdot y} z_k \\
= \sum_{j=1}^m \sum_{k=1}^m \varphi(t_j - t_k) w_j w_k
\]

with \( w_j \equiv e^{it_j \cdot y} z_j \). Thus this is nonnegative by assumption on \( \varphi \).

Since \( |e^{it \cdot y}| = 1 \), and \( |\varphi_y(t)| \leq \varphi(0) \) by 3 of proposition 6.39, \( h(y) \varphi_y(t) \) is nonnegative, continuous and Riemann integrable, and thus \( \int_{\mathbb{R}^n} h(y) \varphi_y(t) \, dy \) is the limit of Riemann sums. For a given Riemann sum of mesh size \( \delta \), with \( v_i > 0 \) denoting the volume of the \( i \)th cell, define:

\[
\psi_\delta(t) \equiv \sum_i h(y_i) \varphi_y(t) v_i.
\]

See section 2.1.1 of book 3 for example. Since this summation is absolutely convergent, we can rearrange sums in the following:

\[
\sum_{j=1}^m \sum_{k=1}^m \psi_\delta(t_j - t_k) z_j z_k = \sum_{i=1}^m \sum_{k=1}^m \sum_{j=1}^m h(y_i) \varphi_y(t_j - t_k) z_j z_k \\
= \sum_{i=1}^m h(y_i) v_i \sum_{j=1}^m \varphi_y(t_j - t_k) z_j z_k.
\]

Because \( h(y_i) v_i \geq 0 \) it follows that each \( \psi_\delta(t) \) is positive semidefinite. Now \( \psi_\delta(t) \to \psi(t) \) pointwise as \( \delta \to 0 \), and thus by continuity this convergence is
uniform on compact sets. As any collection \( \{t_j - t_k\}_{j,k=1}^m \) is contained in a compact set, it follows that:

\[
\sum_{j=1}^m \sum_{k=1}^m \psi(t_j - t_k)z_j \bar{z}_k \rightarrow \sum_{j=1}^m \sum_{k=1}^m \psi(t_j - t_k)z_j \bar{z}_k \geq 0.
\]

\[\blacksquare\]

**Remark 6.41** The following proof was adapted from an online lecture of Rongfeng Sun of National University of Singapore.

**Proposition 6.42 (Bochner’s theorem)** A function \( \varphi : \mathbb{R}^n \to \mathbb{C} \) is the characteristic function of a distribution function defined on \( \mathbb{R}^n \) if and only if \( \varphi \) is positive semidefinite, continuous at \( t = 0 \), and \( \varphi(0) = 1 \).

**Proof.** Since characteristic functions necessarily have the stated properties as noted above, the essence of this result is the sufficiency of these conditions. This proof will first address the case where \( \varphi \) is also absolutely integrable on \( \mathbb{R}^n \), which is the harder step, and then we generalize.

1. **Integrable \( \varphi \):** First assume that in addition to the properties identified, that \( \varphi \) is absolutely integrable on \( \mathbb{R}^n \), and since continuous by proposition 6.39, we can assume Riemann integrable. Define \( f : \mathbb{R}^n \to \mathbb{C} \) by

\[
f(x) = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \varphi(t)e^{-ix\cdot t} dt. \quad (\text{(*)})
\]

Since \( |e^{-ix\cdot t}| = 1 \), it follows by the triangle inequality that \( f(x) \) is well defined, continuous, and absolutely bounded. Continuity follows by the triangle inequality, the continuity of \( e^{-ix\cdot t} \), and Lebesgue’s dominated convergence theorem. Details are an exercise utilizing propositions 2.61 and 2.64 of book 3.

In addition, \( f(x) \) is real valued and nonnegative. To see this, fix \( x \), define \( R_T^+ = [0,T]^n \), and let:

\[
f_T(x) \equiv \frac{1}{(2\pi)^n} \frac{1}{T^n} \int_{R_T^+ \times R_T^+} \varphi(t-s)e^{-ix\cdot (t-s)} ds dt.
\]

Since \( \varphi(t-s) \) is bounded by proposition 6.39, this Riemann integral is the limit of Riemann sums, and indeed Riemann sums with finitely many terms. Any such Riemann sum is real and nonnegative by the positive semidefinite property of \( \varphi \), and thus this integral is real and nonnegative:

\[
f_T(x) \geq 0.
\]
We next perform a change of variables using the invertible linear transformation on $\mathbb{R}^n$ defined by $T : (s, t) \to (u, v)$ where $u \equiv t - s$, $v \equiv s$. This linear transformation has matrix $A \equiv \begin{pmatrix} -I & I \\ Z & I \end{pmatrix}$, there $I$ denotes the $n \times n$ identity, and $Z$ the $n \times n$ zero matrix. The determinant is therefore equal to $(-1)^n$, so $|\det A| = 1$ for the application of proposition 3.24 of book 5. The above integral can be defined over $\mathbb{R}^n \times \mathbb{R}^n$ by multiplying the integrand by the characteristic function $\chi_{R_T^+ \times R_T^+}(s, t)$, and then, notationally ignoring the $|\det A|$ term in the second integral:

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \chi_{R_T^+ \times R_T^+}(t-s)e^{-ix(t-s)}dsdt = \int_{\mathbb{R}^n \times \mathbb{R}^n} \chi_{R_T^+ \times R_T^+}(u+v)\varphi(u)e^{-ixu}dudv.$$  

Now reversing the order of integration using Fubini’s theorem of book 5, the $dv$-integral can be explicitly evaluated. By definition $\chi_{R_T^+ \times R_T^+}(v, u + v) = 1$ if for all $j$ both $0 \leq v_i \leq T$ and $-u_i \leq v_i - T - u_i$ are true. These $v_i$-sets intersect only when $-T \leq u_i \leq T$ for all $i$, and then this intersection set is either $0 \leq v_i \leq T - u_i$ for $u_i \geq 0$ or $-u_i \leq v_i \leq T$ for $u_i < 0$, in either case this set has measure $T - |u_i|$. Thus with $R_T = [-T, T]^n$

$$f_T(x) = \frac{1}{(2\pi)^n} \frac{1}{T^n} \int_{\mathbb{R}^n} \chi_{R_T}(v) \prod_{i=1}^{n} (T - |u_i|)\varphi(u)e^{-ixu}du$$

$$= \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \prod_{i=1}^{n} \left(1 - \frac{|u_i|}{T}\right) \chi_{R_T}(v)\varphi(u)e^{-ixu}du.$$  

Denoting this integrand by $\psi_T(u)$, note that that $\int_{\mathbb{R}^n} |\psi_T(u)|du \leq \int_{\mathbb{R}^n} |\varphi(u)|du < \infty$, and that $\psi_T(u) \to \varphi(u)e^{-ixu}$ pointwise as $T \to \infty$. Thus by Lebesgue’s dominated convergence theorem of proposition 2.61 of book 3, applicable to this Riemann integral by proposition 2.64 of book 3, it follows that $f_T(x) \to f(x)$ in (*) and hence $f(x) \geq 0$ as claimed.

We now show that in this case of absolutely integrable $\varphi$, that such $\varphi$ is in fact the characteristic function of $f(x)$ defined in (*). This will then prove that if integrable, then $f(x)$ is a density function since $1 = \varphi(0) = \int f(x)dx$, and this integral can be defined as a Riemann integral since $f$ is continuous as noted above. But note that it has not yet been proven that $f$ is integrable, so we have to approximate $f$ with integrable functions. For $\lambda > 0$ define $g_\lambda(x) \equiv f(x)\exp\left(-\lambda|x|^2/2\right)$. Then since $f(x)$ is bounded, $g_\lambda(x)$ is integrable for all $\lambda > 0$, and also $g_\lambda(x) \to f(x)$ pointwise as $\lambda \to 0$. Since $g_\lambda(x)$ will not be a density, we take the Fourier transform of $g_\lambda(x)$ (definition 6.26, book 5) noting that this is identical to the characteristic function of $g_\lambda(x)$.
6.6 BOCHNER’S THEOREM

function definition in definition 6.6 above expressed with a density function (see 6.37):
\[
\hat{g}_\lambda(x) \equiv \int_{\mathbb{R}^n} g_\lambda(s) e^{ix \cdot s} ds = \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \varphi(t) \exp \left[ -\lambda |s|^2 / 2 \right] e^{-is \cdot e^{ix \cdot s} dt} ds.
\]

By the integrability assumption on \(\varphi\) this integrand is integrable over \(\mathbb{R}^n \times \mathbb{R}^n\) and thus Fubini’s theorem applies to obtain:
\[
\hat{g}_\lambda(x) = \left. \int_{\mathbb{R}^n} \varphi(t) \exp \left[ -\lambda |s|^2 / 2 \right] e^{is \cdot (x-t)} ds \right| dt.
\]

The inner integral is seen to be the characteristic function of a multivariate normal density in 3.3, with \(\mu = 0\) and \(C = \lambda I\), and evaluated at \(x - t\). Thus by 6.23 and a change of variables:
\[
\hat{g}_\lambda(x) = \left. \int_{\mathbb{R}^n} \varphi(t) \exp \left[ -|x-t|^2 / 2\lambda \right] dt \right.
= \lambda^{-n/2} \int_{\mathbb{R}^n} \varphi(x + s) \exp \left[ -|s|^2 / 2\lambda \right] ds.
\]

Again recognizing the multivariate normal density and letting \(x = 0\):
\[
\int_{\mathbb{R}^n} g_\lambda(s) ds = \hat{g}_\lambda(0) \leq \sup |\varphi(s)| = |\varphi(0)| = 1. \quad (***)
\]

Since \(g_\lambda(s) \to f(x)\) monotonically for each each \(x\), and all functions are nonnegative, Lebesgue’s monotone convergence theorem (proposition 2.50, book 3) assures that \(f\) is integrable and that \(\int_{\mathbb{R}^n} f(s) ds \leq 1\).

Connecting the results:
\[
\int_{\mathbb{R}^n} g_\lambda(s) e^{ix \cdot s} ds = \lambda^{-n/2} \int_{\mathbb{R}^n} \varphi(x + s) \exp \left[ -|s|^2 / 2\lambda \right] ds.
\]

Since \(g_\lambda(s) e^{ix \cdot s} \to f(s) e^{ix \cdot s}\) pointwise as \(\lambda \to 0\), and the integrals of \(g_\lambda(s) e^{ix \cdot s}\) are bounded by the integral of \(f(s)\), Lebesgue’s dominated convergence theorem assures that \(\int_{\mathbb{R}^n} g_\lambda(s) e^{ix \cdot s} ds \to \int_{\mathbb{R}^n} f(s) e^{ix \cdot s} ds\). For the integral on the right, note that by a change of variables:
\[
\frac{\lambda^{-n/2}}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \varphi(x + s) \exp \left[ -|s|^2 / 2\lambda \right] ds = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \varphi(x + t\sqrt{\lambda}) \exp \left[ -|t|^2 / 2 \right] dt.
\]
The integrand converges pointwise to \( \varphi(x) \exp \left[ -|t|^2 / 2 \right] \), and since \( \varphi \) is bounded, Lebesgue’s dominated convergence theorem applies to conclude that this integral converges to \( \varphi(x) \). Combining:

\[
\varphi(x) = \int_{\mathbb{R}^n} f(s) e^{ix \cdot s} ds.
\]

Letting \( x = 0 \) finally confirms that \( \int_{\mathbb{R}^n} f(s) ds = \varphi(0) = 1 \) and thus \( f \) is a density function, and \( \varphi \) is its characteristic function.

2. **General** \( \varphi \): In the general case, we approximate \( \varphi \) with integrable such functions to apply part 1. Define:

\[
\varphi_\lambda(x) \equiv \varphi(x) \exp \left[ -\lambda |x|^2 / 2 \right]
\]

\[
= \frac{1}{\lambda^{n/2} (2\pi)^{n/2}} \int_{\mathbb{R}^n} \varphi(x) \exp \left[ -|s|^2 / 2\lambda \right] e^{is \cdot x} ds,
\]

where the last step follows from 6.23, that \( \exp \left[ -\lambda |x|^2 / 2 \right] \) is the characteristic function of \( \frac{1}{\lambda^{n/2} (2\pi)^{n/2}} \exp \left[ -|s|^2 / 2\lambda \right] \). Now \( \varphi_\lambda(x) \) is continuous at \( x = 0 \) with \( \varphi(0) = 1 \), and is positive semidefinite by proposition 6.40. Thus there exists a density function \( f_\lambda(x) \) by part 1 such that \( \varphi_\lambda(x) \) is the characteristic function of \( f_\lambda(x) \). But \( \varphi_\lambda(x) \to \varphi(x) \) pointwise as \( \lambda \to 0 \), and so by corollary 6.31 to Lévy’s continuity theorem there exists a distribution function \( F \) such that \( \varphi(x) = C_F(x) \).

**Remark 6.43 (Inversion of \( C_F(t) \))** The proof of 6.46 of proposition 6.24 was deferred to this point. To this end, let \( C_F(t) \) be the characteristic function of a distribution function \( F(x) \), which is assumed integrable by the hypothesis of proposition 6.24. Since \( \varphi(t) \equiv C_F(t) \) satisfies the assumptions of Bochner’s theorem, part 1 of the above proof assures the existence of a density function \( f(x) \) so that \( C_F(t) \) is the characteristic function of \( f \). By the uniqueness theorem of proposition 6.25, it follows that \( F \) is the distribution function associated with the density \( f \).

### 6.7 A Uniqueness of Moments Result

In proposition 3.57 of book 4 was stated without proof a key result for that section’s investigation into the uniqueness of moments and the moment generating function. Ironically, the proof of this result takes place in the more general environment of characteristic functions, where we prove that
moments uniquely determine the characteristic function. The last step is then simply quoting the uniqueness theorem, which converts this uniquely defined characteristic function to a unique distribution.

**Proposition 6.44 (Uniqueness of Moments)** Let \( \mu_F \) be a Borel measure on \( \mathbb{R} \) induced by a probability distribution function \( F \) and assume that \( \mu'_n = \int_{-\infty}^{\infty} x^n \, d\mu_F \) exists for all \( n \). If the power series \( \sum_{n=0}^{\infty} \frac{\mu'_n t^n}{n!} \) converges absolutely on \(( -t_0, t_0 \)) for some \( t_0 > 0 \), then \( F \) is the only distribution function with these moments.

**Proof.** Denoting by \( \mu'_{[n]} \) the absolute moments:

\[
\mu'_{[n]} = \int_{-\infty}^{\infty} |x|^n \, d\mu_F,
\]

we first show that \( \mu'_{[n]} \frac{s^n}{n!} \to 0 \) as \( n \to \infty \) for some \( s \) with \( 0 < s < 1 \). By assumption there exists \( t \) with \( 0 < t < 1 \) so that \( \sum_{n=0}^{\infty} \frac{\mu'_n t^n}{n!} \) converges and hence \( \mu'_n t^n / n! \to 0 \) as \( n \to \infty \). Thus \( \mu'_{[n]} \frac{s^n}{n!} \to 0 \) for even \( n \) if \( s < t \). For odd indexes, the inequality \( \mu'_{[2n+1]} \leq (\mu'_{2n+1} \mu'_{2n+2})^{1/2} \) was derived in example 3.49 of book 4 as an application of the Cauchy-Schwarz inequality. Thus:

\[
\mu'_{[2n+1]} \leq \mu'_{(2n+1)} \equiv \max[\mu'_{2n}, \mu'_{2n+2}].
\]

With \( s = \lambda t \) for arbitrary \( 0 < \lambda < 1 \):

\[
\frac{\mu'_{[2n+1]} |s|^{2n+1}}{(2n+1)!} \leq c_{(2n+1)} \frac{\mu'_{(2n+1)} |t|^{2n+1}}{(2n+1)!},
\]

where \( c_{(2n+1)} = (2n+2) \lambda^{2n+1} / |t| \) if \( 2n+2 = 2n+2 \), or \( c_{(2n+1)} = |t| \lambda^{2n+1} / (2n+1) \) if \( 2n+1 = 2n \). Hence since both \( c_{(2n+1)} \to 0 \) and \( \mu'_{(2n+1)} |t|^{2n+1} / (2n+1)! \to 0 \) it follows that \( \mu'_{[2n+1]} |s|^{2n+1} / (2n+1)! \to 0 \) if \( s < t \).

Next since \( |e^{itx}| = 1 \) (6.13, book 5), it follows from proposition 6.25 of book 5 that:

\[
\left| e^{itx} \left( e^{ihx} - \sum_{j=0}^{n} \frac{(ihx)^j}{j!} \right) \right| \leq \frac{\lambda^{n+1}}{(n+1)!}.
\]

By the triangle inequality:

\[
\left| C_F(h + t) - \sum_{j=0}^{n} \frac{h^j}{j!} \int_{-\infty}^{\infty} (ix)^n e^{itx} \, d\mu_F \right| \leq \int_{-\infty}^{\infty} \left| e^{itx} \left( e^{ihx} - \sum_{j=0}^{n} \frac{(ihx)^j}{j!} \right) \right| \, d\mu_F \leq \frac{\lambda^{n+1}}{(n+1)!} \mu'_{[n]}.
\]
and then 6.26 obtains:

\[ \left| C_F(h + t) - \sum_{j=0}^{n} \frac{h^j}{j!} C_F^{(j)}(t) \right| \leq \frac{|h|^{n+1} \mu_{[n]}}{(n + 1)!}. \]

Letting \( n \to \infty \) it follows that for \(|h| \leq s\) defined above, and all \( t \), that

\[ C_F(h + t) = \sum_{j=0}^{\infty} \frac{h^j}{j!} C_F^{(j)}(t). \]

Now assume that \( G \) is another distribution function with the same moments as \( F \); then as above it will follow that for all \( t \),

\[ C_G(h + t) = \sum_{j=0}^{\infty} \frac{h^j}{j!} C_G^{(j)}(t) \]

for \(|h| \leq s\), using the same definition of \( s \) as above, since it depended only on these moments. Letting \( t = 0 \) and applying 6.26 proves that for all \( j \):

\[ C_F^{(j)}(0) = i^j \mu_j = C_G^{(j)}(0), \]

and thus \( C_F(h) = C_G(h) \) for \(|h| \leq s\). Hence \( C_F^{(j)}(h) = C_G^{(j)}(h) \) for \(|h| \leq s\) and all \( j \). Applying the same logic, now letting \( t = s - \epsilon \) and then \( t = -s + \epsilon \), it is derived that \( C_F(h) = C_G(h) \) for \(|h| < 2s - \epsilon\). Repeating with \( t = \pm 2(s - \epsilon) \), and so forth, we then conclude that for every \( m \) that \( C_F(h) = C_G(h) \) for \(|h| < (m + 1)s - m\epsilon\), and hence \( C_F(h) = C_G(h) \) for all \( h \).

By the uniqueness property in 4 above, this implies that \( \mu_F(A) = \mu_G(A) \) for all Borel sets \( A \in \mathcal{B}(\mathbb{R}) \). Taking \( A = (-\infty, x] \) produces \( F(x) = G(x) \) for all \( x \). \( \blacksquare \)
In this chapter we explore applications of characteristic functions. First, the central limit theorem of proposition 5.14 of book 4 can now be greatly generalized from the earlier context of distribution functions with moment generating functions, to those with only two moments. Because every distribution function has an associated characteristic function, and moreover is uniquely "characterized" by this characteristic function, it is the perfect tool to use to generalize the earlier result. This investigation is all the more compelling because of the similarities between moment generating function manipulations and those using characteristic functions, as well as the connection between both special functions and the moments of the given distribution.

The second area of application is to the investigation of distribution functions of sums of random variables. The method employed is similar to that applied in example 3.59 of book 4 using moment generating functions, but now using characteristic functions, and so the results will extend beyond the earlier special situations. We will also apply these methods to investigate so-called "infinitely divisible" distributions. The final section will to a far limited extent investigate the distribution function of products of random variables.
7.1 Central Limit Theorems

One version of the central limit theorem was stated and proved in proposition 5.14 of book 4. This result required the very strong assumption that the distribution function of the independent and identically distributed random variables had an associated moment generating function $M_X(t)$ for $t \in (-t_0, t_0)$. With the use of characteristic functions, we can effectively reproduce the earlier proof, but now requiring only the assumption that this distribution function has two moments. This is the "classical" statement of the central limit theorem, from which many generalizations exist. The idea of using characteristic functions to prove central limit theorems and other weak convergence results was introduced by Paul Lévy (1886 – 1971). Consequently, one often finds the name Lévy attached to statements of various versions of this theorem when the proof is based on an analysis of characteristic functions.

We present three generalizations of the earlier book 4 result, but first prove a couple of needed technical estimates related to complex numbers.

**Lemma 7.1** 1. If $z \in \mathbb{C}$, then:

$$|e^z - (1 + z)| \leq |z|^2 e^{|z|}. \quad (7.1)$$

2. Given $\{a_j\}_{j=1}^n, \{b_j\}_{j=1}^n \subset \mathbb{C}$ with $|a_j| \leq \lambda$ and $|b_j| \leq \lambda$:

$$\left| \prod_{j=1}^n a_j - \prod_{j=1}^n b_j \right| \leq \lambda^{n-1} \sum_{j=1}^n |a_j - b_j|. \quad (7.2)$$

**Proof.**

1.

2. Note that by the development in section 6.3.2 of book 5,

$$e^z = \sum_{j=0}^{\infty} \frac{z^j}{j!}, \quad (7.3)$$

and this is an absolutely convergent series for all $z$. Thus:

$$|e^z - (1 + z)| = \left| \sum_{j=2}^{\infty} \frac{z^j}{j!} \right| \leq |z|^2 \sum_{j=0}^{\infty} \frac{|z|^j}{(j+2)!} \leq |z|^2 e^{|z|}.$$
7.1 CENTRAL LIMIT THEOREMS

2. This inequality is apparently true for \( n = 1 \), so the proof is completed by induction. Assuming that this result is true for \( n - 1 \):

\[
|\prod_{j=1}^{n} a_j - \prod_{j=1}^{n} b_j| \leq |a_1 \prod_{j=2}^{n} a_j - \prod_{j=2}^{n} b_j| + |a_1 \prod_{j=2}^{n} b_j - b_1 \prod_{j=2}^{n} b_j| \\
\leq \lambda \left[ \lambda^{n-2} \sum_{j=2}^{n} |a_j - b_j| \right] + |a_1 - b_1| \lambda^{n-1} \\
= \lambda^{n-1} \sum_{j=1}^{n} |a_j - b_j|.
\]

\[\square\]

7.1.1 The Classical Central Limit Theorem

The term "classical" is typically applied to the following statement of the central limit theorem. It generalizes proposition 5.14 of book 4, requiring only the existence of two moments. Note that for this convergence result, that it is necessary to "normalize" the sum or average of independent, identically distributed variates to produce a random variable \( Y_n \) with:

\[E[Y_n] = 0, \quad Var[Y_n] = 1.\]

A calculation obtains that this normalization is achieved in either cases by the transformation:

\[Y_n = \frac{W - E[W]}{\sqrt{Var[W]}}.\]  \hspace{1cm} (7.4)

**Proposition 7.2 (Central Limit Theorem 2)** Let \( F_X \) denote the distribution function of a random variable \( X \) with mean \( \mu \) and variance \( \sigma^2 \), and

\[\Phi(x) \] the distribution of the standard normal associated with the density function in 3.2. Let \( Y_n \) denote the normalized random variable associated with the sum or average of \( n \) independent values of \( X \):

\[Y_n = \frac{X - \mu}{\sqrt{\sigma^2/n}}.\]

Then as \( n \rightarrow \infty \),

\[F_{Y_n} \Rightarrow \Phi.\]  \hspace{1cm} (7.5)

**Proof.** First using 6.25 then 6.32, it follows that with \( Y = X - \mu \) and \( Y_n = \sum_{j=1}^{n} (X_j - \mu) / \sqrt{n} \sigma \):

\[C_{Y_n}(t) = \left[ C_Y \left( t / \sqrt{n} \sigma \right) \right]^n.\]
From proposition 6.25 of book 5:

\[ |e^{iy} - \sum_{j=0}^{2} (iy)^j / j!| \leq \min \left[ |y|^3 / 6, |y|^2 \right], \]

and so with \( y = Yt/\sqrt{n}\sigma \),

\[
E \left[ e^{iYt/\sqrt{n}\sigma} \right] - E \left[ \sum_{j=0}^{2} (it)^j (Y/\sqrt{n}\sigma)^{j/j!} \right] \\
\leq E \left[ e^{iYt/\sqrt{n}\sigma} - \sum_{j=0}^{2} (it)^j (Y/\sqrt{n}\sigma)^{j/j!} \right] \\
\leq E \left[ \min \left( \left( |Yt| / \left[ \sqrt{n}\sigma \right] \right)^3 / 6, \left( |Yt| / \left[ \sqrt{n}\sigma \right] \right)^2 \right) \right] \\
= \frac{t^2}{n} E \left[ \min \left( \left| \frac{t}{\sqrt{n}} \right| \left( \frac{|Y|}{\sigma} \right)^3, \left( \frac{Y}{\sigma} \right)^2 \right) \right]. \quad (*)
\]

Fixing \( t \) and letting \( f_n(Y) = \min \left( \left| \frac{t}{\sqrt{n}} \right| \left( \frac{|Y|}{\sigma} \right)^3, \left( \frac{Y}{\sigma} \right)^2 \right) \), it follows that \( f_n(Y) \leq f(Y) \equiv \left( \frac{Y}{\sigma} \right)^2 \) for all \( n \), and that \( f \) is integrable by assumption of the existence of \( \sigma^2 \). Because \( f_n(Y) \to 0 \) pointwise as \( n \to \infty \), Lebesgue’s dominated convergence theorem (proposition 2.43 of book 5) obtains that \( E[f_n(Y)] \to 0 \) as \( n \to \infty \). Thus, the upper bound in (*) is \( o(n^{-1}) \). Since \( E \left[ e^{iYt/\sqrt{n}\sigma} \right] \equiv C_Y \left( t/\left[ \sqrt{n}\sigma \right] \right) \), \( E[Y] = 0 \) and \( E[Y^2] = \sigma^2 \), this yields:

\[ C_Y \left( t/\left[ \sqrt{n}\sigma \right] \right) = 1 - t^2/n + o(n^{-1}). \]

If it can be proved that \( (1 - t^2/2n + o(n^{-1}))^n \to \exp \left( -\frac{1}{2} t^2 \right) \) for all \( t \), then 6.22 and Lévy’s continuity theorem on \( \mathbb{R} \) will complete the proof. For this limit, it is important to remember that the error term of \( o(n^{-1}) \) is complex, and hence we must prove this as a result in \( \mathbb{C} \), and not as a result in \( \mathbb{R} \) as was proved in the derivation of proposition 5.14 of book 4. To this end, note that:

\[
\left| \exp \left( -t^2/2 \right) - (1 - t^2/2n + o(n^{-1}))^n \right| \\
\leq \left| \exp \left( -t^2/2 \right) - \exp \left[ -t^2/2 + n\left[ o(n^{-1}) \right] \right] \right| \\
+ \left| \exp \left[ -t^2/2 + n\left[ o(n^{-1}) \right] \right] - \left[ 1 - t^2/2n + o(n^{-1}) \right]^n \right|.
\]

The first expression converges to zero with \( n \) by the continuity of the exponential function since \( n\left[ o(n^{-1}) \right] \to 0 \) by definition. For the second expression, let \( w = \exp(z) \) where \( z = -t^2/2n + o(n^{-1}) \), then apply 7.2:

\[ |w^n - (1 + z)^n| \leq n\lambda^{n-1} |w - (1 + z)|, \]
where
\[
\lambda = \max \left[ |1 + z|, |\exp z| \right] \leq \max \left[ 1 + |z|, \exp |z| \right].
\]
Then applying 7.1:
\[
|w^n - (1 + z)^n| \leq n\lambda^{n-1} |z|^2 e^{|z|}.
\]
Thus as \( n \to \infty \), \( |z| = O\left( n^{-1} \right) \), \( \lambda^{n-1} = O(\exp (-t^2/2)) \), and so
\[
|w^n - (1 + z)^n| = O\left( n^{-1} \right) O(\exp (-t^2/2)).
\]
Putting the pieces together obtains:
\[
\left| \exp \left( -\frac{1}{2} t^2 + o(n^{-1}) \right) - \left( 1 - \frac{1}{2} t^2 / n + o(n^{-1}) \right)^n \right| \to 0,
\]
completing the proof. ■

7.1 CENTRAL LIMIT THEOREMS

7.1.2 Lindeberg’s Central Limit Theorem

The above result can be generalized from sums of independent, identically distributed random variables, \( S_n = \sum_{j=1}^{n} X_j \), to sums of independent, but not necessarily identically distributed random variables. Specifically, we now define \( S_n = \sum_{j=1}^{r_n} X_{n,j} \) where \( r_n \to \infty \) as \( n \to \infty \), and for each \( n \), \( \{X_{n,j}\}_{j=1}^{r_n} \) are assumed independent. This general set-up applies to two common situations:

1. For all \( n \), \( X_{n,j} = X_j \) and \( r_n = n \). In this case, \( S_n = \sum_{j=1}^{n} X_j \) and the generalization relative to proposition 7.2 above is that while independent, the collection \( \{X_j\}_{j=1}^{\infty} \) is not assumed to be identically distributed.

2. More generally, \( \{X_{n,j}\}_{j=1}^{r_n} \) are assumed independent for each \( n \) but not necessarily identically distributed. The collection \( \left\{ \{X_{n,j}\}_{j=1}^{r_n} \right\}_{n=1}^{\infty} \) is then called a triangular array of random variables because of the shape of the array that is formed by placing each set of variates \( \{X_{n,j}\}_{j=1}^{r_n} \) into the \( n \)th row of an array. Thus each row of the array contains independent random variables by assumption, while each column contains independent or dependent random variables. That columns may contain dependent random variables should not surprise. In situation 1, not only are \( \{X_{n,j}\}_{n=f(j)}^{\infty} \) dependent, where \( f(j) = \min\{n : r_n \geq j\} \), but they are identical!
CHAPTER 7 APPLICATIONS OF CHARACTERISTIC FUNCTIONS

The first generalization of this type was to the context in situation 1 above, where \( S_n \equiv \sum_{j=1}^{n} X_j \) and \( \{X_j\}_{j=1}^{\infty} \) are independent but not assumed to be identically distributed. This result was formulated in 1922 by **J. W. (Jarl Waldemar) Lindeberg** (1876 – 1932). For such a result to be proved it was necessary to impose some condition that ensures that \( S_n \) is not dominated by one or a few random variables.

**Example 7.3** Assume \( X_j \) has mean 0 and variance \( j!(j!) \). Then \( S_n \equiv \sum_{j=1}^{n} X_j \) has mean 0, and if these variables are independent, mathematical induction obtains that \( S_n \) has a variance of \( (n+1)! - 1 \). Normalizing, it follows that:

\[
S_n/\sqrt{(n+1)! - 1} = X_n/\sqrt{(n+1)! - 1} + S_{n-1}/\sqrt{(n+1)! - 1},
\]

and:

\[
\text{Var}\left[ X_n/\sqrt{(n+1)! - 1} \right] = \frac{n(n!)}{(n+1)! - 1},
\]

\[
\text{Var}\left[ S_{n-1}/\sqrt{(n+1)! - 1} \right] = \frac{n! - 1}{(n+1)! - 1}.
\]

This first variance approaches 1 as \( n \to \infty \), while the second approaches 0. An application of Chebyshev’s inequality of proposition 3.32 of book 4 shows that for every \( \epsilon > 0 \):

\[
\Pr\left[ \left| S_{n-1}/\sqrt{(n+1)! - 1} \right| \geq \epsilon \right] \leq \frac{1}{\epsilon^2} \frac{n! - 1}{(n+1)! - 1} \to 0,
\]

and so the distribution function of the second term converges weakly to 0.

Hence, while the normalized summation \( S_n/\sqrt{(n+1)! - 1} \) has mean 0 and variance 1 by construction, it has little hope of achieving a normal distribution because as \( n \to \infty \), this normalized variate approaches the distribution of \( X_n \), normalized to have a variance that approaches 1 as \( n \to \infty \).

The restriction imposed by Lindeberg to avoid this anomalous behavior has come to be known as **Lindeberg’s condition**, or the **Lindeberg condition**.

**Definition 7.4 (Lindeberg’s Condition)** If \( \{X_j\}_{j=1}^{\infty} \), respectively \( \{X_{n,j}\}_{j=1}^{\infty} \), are defined on the probability space \( (S, \mathcal{E}, \lambda) \), the **Lindeberg condition** is satisfied under the following conditions:
1. In the case of a summation of independent random variables with \( \mu_j = E[X_j] \), \( \sigma_j^2 = Var[X_j] \) and \( s_n^2 = \sum_{j=1}^{n} \sigma_j^2 \), it is assumed that for every \( t > 0 \),
\[
\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{j=1}^{n} \int_{|X_j - \mu_j| \geq ts_n} (X_j - \mu_j)^2 \, d\lambda = 0.
\] (7.6)

2. For triangular arrays with \( \mu_{n,j} = E[X_{n,j}] \), \( \sigma_{n,j}^2 = Var[X_{n,j}] \) and \( s_n^2 = \sum_{j=1}^{r_n} \sigma_{n,j}^2 \), it is assumed that for every \( t > 0 \),
\[
\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{j=1}^{r_n} \int_{|X_{n,j} - \mu_{n,j}| \geq ts_n} (X_{n,j} - \mu_{n,j})^2 \, d\lambda = 0.
\] (7.7)

Either integral is also commonly expressed as a Lebesgue-Stieltjes integral on \( \mathbb{R} \), reflecting the Borel measures induced by the respective distribution functions, \( d\mu_j \) or \( d\mu_{n,j} \).

**Remark 7.5** Note that using the notation of 1:
\[
\sigma_j^2 = \int_{|X_j - \mu_j| < ts_n} (X_j - \mu_j)^2 \, d\lambda + \int_{|X_j - \mu_j| \geq ts_n} (X_j - \mu_j)^2 \, d\lambda.
\] (**)

So the Lindeberg condition in 7.6 can be equivalently stated that for every \( t > 0 \):
\[
\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{j=1}^{n} \int_{|X_j - \mu_j| < ts_n} (X_j - \mu_j)^2 \, d\lambda = 1,
\] (7.8)

with an analogously restatement for 7.7.

Intuitively, the Lindeberg condition assures that asymptotically, all of the variance of \( S_n \) comes from the variability of the individual variates "relatively near" their respective means, where by relatively near is meant, when measured relative to any multiple of \( s_n \). More formally, using the above integral expression in (*) for \( \sigma_j^2 \), we have for any \( t > 0 \):
\[
\frac{\sigma_j^2}{s_n^2} \leq t^2 + \frac{1}{s_n^2} \int_{|X_j - \mu_j| \geq ts_n} (X_j - \mu_j)^2 \, d\lambda,
\]
and so
\[
\max_{j \leq n} \frac{\sigma_j^2}{s_n^2} \leq t^2 + \frac{1}{s_n^2} \sum_{j=1}^{r_n} \int_{|X_j - \mu_j| \geq ts_n} (X_j - \mu_j)^2 \, d\lambda.
\]
Since this integral sum converges to 0 with \( n \), and \( t \) is arbitrary, this proves that as \( n \to \infty \):
\[
\max_{j \leq n} \frac{\sigma_j^2}{s_n^2} \to 0 \text{ as } n \to \infty.
\]
The same analysis applies to triangular arrays.

Hence, the Lindeberg condition assures that as \( n \to \infty \), all variates in the summation contribute insignificantly to the total variance on a relative basis, and hence no one variate can dominate as in the above example.

The following statement and proof of the Lindeberg central limit theorem is given in the more general context of triangular arrays. Simply assuming that \( X_{n,j} = X_j \), \( \mu_{n,j} = \mu_j \), \( \sigma_{n,j}^2 = \sigma_j^2 \) and \( r_n = n \) for all \( n \) produces the version applicable to \( S_n = \sum_{j=1}^{n} X_j \), where the collection \( \{X_j\}_{j=1}^{\infty} \) is assumed independent but not necessarily identically distributed.

Note that \( Y_n \) defined below is the normalized version of \( S_n = \sum_{j=1}^{r_n} X_{n,j} \) as in 7.4, since \( E[S_n] = \sum_{j=1}^{r_n} \mu_{n,j} \) and \( \text{Var}[S_n] = s_n^2 \).

**Proposition 7.6 (Lindeberg’s Central Limit Theorem)** For each \( n \) let \( \{X_{n,j}\}_{j=1}^{r_n} \) be independent random variables where \( r_n \to \infty \) as \( n \to \infty \). Denote \( \mu_{n,j} = E[X_{n,j}], \sigma_{n,j}^2 = \text{Var}[X_{n,j}] \) and \( s_n^2 = \sum_{j=1}^{r_n} \sigma_{n,j}^2 \), and define \( Y_n = \sum_{j=1}^{r_n} (X_{n,j} - \mu_{n,j}) / s_n \). If 7.7 is satisfied for all \( t \), then as \( n \to \infty \):

\[
F_{Y_n} \Rightarrow \Phi, \tag{7.9}
\]

where \( \Phi(x) \) is the distribution of the standard normal.

**Proof.** First note that defining \( X'_{n,j} \equiv (X_{n,j} - \mu_{n,j}) / s_n \), that \( \mu'_{n,j} \equiv E[X'_{n,j}] = 0 \), \( (\sigma'_{n,j})^2 \equiv \text{Var}[X'_{n,j}] = \sigma_{n,j}^2 / s_n^2 \) and \( Y_n = \sum_{j=1}^{r_n} X'_{n,j} \). Hence proving this result for independent \( \{X'_{n,j}\}_{j=1}^{r_n} \) under the assumptions that \( \mu'_{n,j} = 0 \) and \( (s'_n)^2 = \sum_{j=1}^{r_n} (\sigma'_{n,j})^2 = 1 \) yields the general result. We drop the prime notation for convenience.

As in the proof of proposition 7.2 but with \( y = Yt \), it follows that for any \( \epsilon > 0 \), and \( C_{n,j}(t) \) denoting the characteristic function of such \( X_{n,j} \):

\[
|C_{n,j}(t) - (1 - t^2 \sigma_{n,j}^2 / 2)| \leq E \left[ \min \left( |tX_{n,j}|^3 / 6, |tX_{n,j}|^2 \right) \right]
\]

\[
\leq \int_{|X_{n,j}| < \epsilon} |tX_{n,j}|^3 / 6 d\lambda + \int_{|X_{n,j}| \geq \epsilon} |tX_{n,j}|^2 d\lambda
\]

\[
\leq \epsilon t^3 \sigma_{n,j}^2 / 6 + t^2 \sum_{j=1}^{r_n} |X_{n,j}|^2 d\lambda.
\]

Since \( \sum_{j=1}^{r_n} \sigma_{n,j}^2 = 1 \):

\[
\sum_{j=1}^{r_n} |C_{n,j}(t) - (1 - t^2 \sigma_{n,j}^2 / 2)| \leq \epsilon t^3 / 6 + t^2 \sum_{j=1}^{r_n} |X_{n,j}|^2 d\lambda,
\]
and by 7.7, since \( \epsilon \) is arbitrary:

\[
\sum_{j=1}^{r_n} |C_{n,j}(t) - (1 - t^2 \sigma_{n,j}^2/2)| \to 0 \text{ as } n \to \infty.
\]

To obtain 7.9, it must be shown that this proves that the characteristic function of \( Y_n \), 
\( C_n(t) = \prod_{j=1}^{r_n} C_{n,j}(t) \), satisfies:

\[
C_n(t) \to \exp \left[-\frac{t^2}{2}\right],
\]

for all \( t \). To this end, recall that as derived above, the Lindeberg condition assures that 
\( \max_{j \leq n} \sigma_{n,j}^2 \to 0 \) as \( n \to \infty \) since here \( s_n^2 = 1 \). So given \( t \) there is an \( N \) so that for \( n \geq N \) and all \( j \):

\[
0 \leq 1 - t^2 \sigma_{n,j}^2/2 \leq 1.
\]

Applying 7.2 and recalling that \( |C_{n,j}(t)| \leq 1 \) produces:

\[
\left| \prod_{j=1}^{r_n} C_{n,j}(t) - \prod_{j=1}^{r_n} (1 - t^2 \sigma_{n,j}^2/2) \right| \leq \sum_{j=1}^{r_n} \left| C_{n,j}(t) - (1 - t^2 \sigma_{n,j}^2/2) \right|.
\]

By the above estimate, it follows that

\[
\prod_{j=1}^{r_n} C_{n,j}(t) = \prod_{j=1}^{r_n} (1 - t^2 \sigma_{n,j}^2/2) + \epsilon_n
\]

where \( |\epsilon_n| \to 0 \) as \( n \to \infty \).

Now 7.2 and then 7.1 obtains that for \( n \geq N \):

\[
\left| \prod_{j=1}^{r_n} \exp \left(-t^2 \sigma_{n,j}^2/2\right) - \prod_{j=1}^{r_n} (1 - t^2 \sigma_{n,j}^2/2) \right|
\leq \sum_{j=1}^{r_n} \left| \exp \left(-t^2 \sigma_{n,j}^2/2\right) - (1 - t^2 \sigma_{n,j}^2/2) \right|
\leq \sum_{j=1}^{r_n} \left( t^2 \sigma_{n,j}^2/2 \right)^2 \exp \left(t^2 \sigma_{n,j}^2/2\right)
\leq t^4 \max_j \left[ \exp \left(t^2 \sigma_{n,j}^2/2\right) \right] \sum_{j=1}^{r_n} \sigma_{n,j}^4
\leq t^4 \max_j \left[ \exp \left(t^2 \sigma_{n,j}^2/2\right) \right] \max_j \sigma_{n,j}^2 \sum_{j=1}^{r_n} \sigma_{n,j}^2.
\]

As \( \sum_{j=1}^{r_n} \sigma_{n,j}^2 = 1 \) and \( \max \sigma_{n,j}^2 \to 0 \):

\[
\prod_{j=1}^{r_n} (1 - t^2 \sigma_{n,j}^2/2) = \prod_{j=1}^{r_n} \exp \left[-t^2 \sigma_{n,j}^2/2\right] + \epsilon'_n
\]

where \( |\epsilon'_n| \to 0 \) as \( n \to \infty \).

Finally, since \( \prod_{j=1}^{r_n} \exp \left[-t^2 \sigma_{n,j}^2/2\right] = \exp \left[-t^2/2\right] \) and \( C_n(t) = \prod_{j=1}^{r_n} C_{n,j}(t) \), the combined estimates yield:

\[
C_n(t) = \exp \left[-t^2/2\right] + \epsilon_n + \epsilon'_n,
\]

and the result follows.  \( \blacksquare \)
Remark 7.7 When \(X_{n,j}\) are defined on \((\mathcal{S}, \mathcal{E}, \lambda)\) and identically distributed for all \(n\) and \(j\), which is the set-up for classical central limit theorem, the Lindeberg condition is always satisfied. To see this, note that:

\[
\frac{1}{s_n^2} \sum_{j=1}^{r_n} \int_{|X_{n,j} - \mu_{n,j}| \geq ts_n} (X_{n,j} - \mu_{n,j})^2 d\lambda = \frac{1}{\sigma^2} \int \chi_{A_n}(s) (X(s) - \mu)^2 d\lambda(s),
\]

where \(\chi_{A_n}(s)\) is the characteristic function of \(A_n \subset \mathcal{S}\) defined as \(A_n = \{s \mid |X(s) - \mu| \geq \sigma \sqrt{r_n}\}\). By Chebyshev’s inequality (proposition 3.32, book 4) and the assumption that \(r_n \to \infty\), \(\lambda(A_n) \to 0\) as \(n \to \infty\) for any \(t > 0\), and thus \(\chi_{A_n}(s) (X(s) - \mu)^2 \to 0\) pointwise \(\lambda\text{-a.e.}\) as \(n \to \infty\). Now for all \(n\):

\[
\chi_{A_n}(s) (X(s) - \mu)^2 \leq (X(s) - \mu)^2,
\]

an integrable function. Thus by Lebesgue’s dominated convergence theorem (proposition 2.43, book 5), the Lindeberg condition is satisfied.

Exercise 7.8 Assume that for each \(n\), \(\{X_{n,j}\}_{j=1}^{r_n}\) are independent and also identically distributed. In other words, assume that each row of the triangular array has a fixed distribution. Show that in this case 7.7 reduces to the assumption that for all \(t > 0:\)

\[
\lim_{n \to \infty} \int_{|Z_n| > t \sqrt{s_n}} Z_n^2 d\lambda = 0,
\]

(7.10)

where \(Z_n\) denotes the normalized version of \(X_n\) as in 7.4, \(Z_n = (X_n - \mu_n)/\sigma_n\).

More generally, we have the following corollary.

Corollary 7.9 (Lindeberg’s Central Limit theorem) For each \(n\), let \(\{X_{n,j}\}_{j=1}^{r_n}\) be independent random variables with \(\mu_{n,j} = E[X_{n,j}]\) and \(\sigma_{n,j}^2 = \text{Var}[X_{n,j}]\). Assume that \(r_n \to \infty\) as \(n \to \infty\), and also that \(s_n^2 = \sum_{j=1}^{r_n} \sigma_{n,j}^2 \to \sigma^2\), and \(\mu_n \equiv \sum_{j=1}^{r_n} \mu_{n,j} \to \mu\). Define \(Z_n = \left(\sum_{j=1}^{r_n} X_{n,j} - \mu\right)/\sigma\). If 7.7 is satisfied for all \(t\), then as \(n \to \infty:\)

\[
F_{Z_n} \Rightarrow \Phi,
\]

(7.11)

where \(\Phi(x)\) is the distribution of the standard normal. In other words, \(\sum_{j=1}^{r_n} X_{n,j}\) converges in distribution to the normal distribution with mean \(\mu\) and variance \(\sigma^2\).
7.1 CENTRAL LIMIT THEOREMS

Proof. Note that with \( Y_n \) as in proposition 7.6,

\[ Z_n = c_n Y_n + d_n \]

where \( c_n \equiv s_n / \sigma \) and \( d_n \equiv (\mu_n - \mu) / \sigma \). Now applying proposition 9.16 of book 2, since \( c_n \to 1 \) and \( d_n \to 0 \), we have that \( Z_n \) and \( Y_n \) converge in distribution to the same distribution function, which is \( \Phi \) by proposition 7.6. ■

Remark 7.10 It turns out that the Lindeberg condition is nearly best possible. It was proved by William Feller (1906–1970) in 1935 that if \( s_n^2 \to \infty \) and \( \max_{j \leq n} \sigma_{n,j}^2 / s_n^2 \to 0 \) as \( n \to \infty \), then 7.9 implies the Lindeberg condition in 7.6. When the above theorem is stated in terms of Lindeberg’s sufficient and Feller’s necessary conditions, it is often referred to as the Lindeberg-Feller central limit theorem.

But if there is a shortcoming with the Lindeberg condition, it is that it can be difficult to verify, even in relatively simple cases. The following is a relatively simple example with \( \{X_j\}_{j=1}^\infty \) binomial variates.

Example 7.11 Let \( S_n = \sum_{j=1}^n X_j \) where \( \{X_j\}_{j=1}^\infty \) are independent with \( X_j \) binomially distributed with parameter \( p_j \equiv X_j^{-1}(1) \), \( 0 < p_j < 1 \), and so \( \mu_j = p_j \), \( \sigma_j^2 = p_j(1 - p_j) \) and \( s_n^2 = \sum_{j=1}^n p_j(1 - p_j) \). For any \( j \) define \( m_j = \min[p_j, 1 - p_j] \) and \( M_j = \max[p_j, 1 - p_j] \), then:

\[
\int_{|X_j - p_j| \geq ts_n} (X_j - p_j)^2 \, d\lambda = \begin{cases} 
    p_j (1 - p_j), & ts_n \leq m_j, \\
    p_j (1 - p_j)^2 \text{ or } (1 - p_j) p_j^2, & m_j < ts_n \leq M_j, \\
    0, & ts_n > M_j,
\end{cases}
\]

where the value of the second row depends on whether \( m_j \) equals \( 1 - p_j \) or \( p_j \).

Since all \( M_j \leq 1 \) by definition, it follows that if \( s_n^2 \to \infty \) then:

\[
\sum_{j=1}^n \int_{|X_j - p_j| \geq ts_n} (X_j - p_j)^2 \, d\lambda = 0
\]

if \( s_n > 1/t \), and this assures Lindeberg’s condition is satisfied. The condition on \( s_n^2 \to \infty \) is assured for example if \( 0 < a \leq p_j \leq b < 1 \), but this is not necessary as \( p_j = 1/j \) or \( p_j = 1 - 1/j \) demonstrates. In this case of divergent variance, the Feller result above can also be employed to assure that the
Lindeberg condition is satisfied, since $\sigma_j^2 \leq 1/4$ assures that $\max_{j \leq n} \sigma_j^2/s_n^2 \to 0$ as $n \to \infty$.

In the case of convergence with $s_n^2 \to s^2 > 0$ say, then $p_j \to 0$ or 1, so $m_j \to 0$ and $M_j \to 1$ and hence only the limit in the second row of the above table needs investigation. Assuming $p_j \to 0$, then for $0 < t < 1/s$:

$$\frac{1}{s_n^2} \sum_{j=1}^{n} \int_{|X_j - p_j| \geq ts_n} (X_j - p_j)^2 \, d\lambda = \frac{\sum_{j=1}^{n} p_j (1 - p_j)^2}{\sum_{j=1}^{n} p_j (1 - p_j)} \geq \min_{j \leq n} (1 - p_j).$$

Similarly, if $p_j \to 1$ this ratio exceeds $\min_{j \leq n} p_j$. So in the case of convergent variance, $s_n^2 \to s^2$, the Lindeberg condition is not satisfied.

### 7.1.3 Lyapunov’s Central Limit Theorem

An important alternative sufficient condition to ensure 7.9 was developed by Aleksandr Lyapunov (1857 – 1918), and called Lyapunov’s condition. The proof of the associated central limit theorem is relatively easy and amounts to demonstrating that Lyapunov’s condition implies Lindeberg’s condition. In theory this implies that Lyapunov’s condition is weaker than Lindeberg’s, but in applications it is often far easier to verify, and thus can be more useful in practice. As was the case for the Lindeberg central limit theorem, the following statement and proof of Lyapunov’s central limit theorem is given in the more general context of triangular arrays. Simply assuming that $X_{n,j} = X_j$, $\mu_{n,j} = \mu_j$, $\sigma_{n,j}^2 = \sigma_j^2$ and $r_n = n$ for all $n$ produces the version applicable to $S_n = \sum_{j=1}^{n} X_j$, where the collection $\{X_j\}_{j=1}^{\infty}$ is assumed independent but not necessarily identically distributed.

**Definition 7.12 (Lyapunov’s Condition)** Applied to a triangular array with notation of definition 7.4 above, Lyapunov’s Condition requires that for some $\delta > 0$, that $|X_{n,j} - \mu_{n,j}|^{2+\delta}$ is integrable for all $n, j$, and:

$$\lim_{n \to \infty} \frac{1}{s_n^{2+\delta}} \sum_{j=1}^{r_n} E \left[ |X_{n,j} - \mu_{n,j}|^{2+\delta} \right] = 0. \quad (7.12)$$

**Example 7.13** Note that when $X_{n,j}$ are defined on $(S, \mathcal{E}, \lambda)$ and identically distributed for all $n, j$, which is the set-up for classical central limit theorem, Lyapunov’s condition need not be satisfied simply because $X_n$ may only have two moments and thus $|X_n - \mu_n|^{2+\delta}$ need not be integrable for any $\delta > 0$. Thus in this special case the classical central limit theorem is a little more general than is the result from Lyapunov’s central limit theorem. However,
in this case of i.i.d. $X_n$, if $|X_n - \mu_n|^{2+\delta}$ is integrable for some $\delta > 0$, then Lyapunov’s condition is indeed satisfied by the following exercise.

**Exercise 7.14** Assume that for each $n$, $\{X_{n,j}\}_{j=1}^{r_n}$ are independent and also identically distributed. In other words, assume that each row of the triangular array has a fixed distribution. Show that in this case 7.12 reduces to the assumption that for some $\delta > 0$:

$$\lim_{n \to \infty} \frac{1}{n^{\delta/2}} \int |Z_n|^{2+\delta} d\lambda = 0,$$

(7.13)

where $Z_n$ denotes the normalized version of $X_n$ as in 7.4, $Z_n = (X_n - \mu_n)/\sigma_n$.

Prove that 7.13 obtains the result of example 7.13, and also prove that this special case of Lyapunov’s condition implies the special case of Lindeberg’s condition in 7.10.

Lyapunov’s central limit theorem is next. The proof generalizes the last demonstration of exercise 7.14, that Lyapunov’s condition implies Lindeberg’s condition.

**Proposition 7.15 (Lyapunov’s Central Limit Theorem)** For each $n$ let $\{X_{n,j}\}_{j=1}^{r_n}$ be independent random variables where $r_n \to \infty$ as $n \to \infty$. Denote $\mu_{n,j} = E[X_{n,j}]$, $\sigma_{n,j}^2 = \text{Var}[X_{n,j}]$ and $s_n^2 = \sum_{j=1}^{r_n} \sigma_{n,j}^2$, and define $Y_n = \sum_{j=1}^{r_n} (X_{n,j} - \mu_{n,j})/s_n$. If 7.12 is satisfied for some $\delta > 0$, then as $n \to \infty$:

$$F_{Y_n} \Rightarrow \Phi,$$

(7.14)

where $\Phi(x)$ is the distribution of the standard normal.

**Proof.** As noted above, we show that 7.12 implies the Lindeberg condition in 7.7. To this end, given $t > 0$:

$$\frac{1}{s_n^2} \sum_{j=1}^{r_n} \int_{|X_{n,j} - \mu_{n,j}| \geq ts_n} (X_{n,j} - \mu_{n,j})^2 d\lambda$$

$$\leq \frac{1}{s_n^2} \sum_{j=1}^{r_n} \int_{|X_{n,j} - \mu_{n,j}| \geq ts_n} \frac{|X_{n,j} - \mu_{n,j}|^{2+\delta}}{(ts_n)^\delta} d\lambda$$

$$\leq \frac{1}{t^\delta} \frac{1}{s_n^{2+\delta}} \sum_{j=1}^{r_n} E\left[|X_{n,j} - \mu_{n,j}|^{2+\delta}\right].$$

This last expression converges to 0 as $n \to \infty$ by 7.12. \qed
Example 7.16 Recall example 7.11 of the sum of independent binomial random variables in the case where \( s_n^2 \to \infty \). When 
\[
0 < a \leq p_j \leq b < 1,
\]
it is an exercise to verify that 7.12 is satisfied with \( \delta = 1 \) and hence the central limit theorem applies. But divergent variance does not imply that \( p_j \) is bounded away from 0 and 1 as the example \( p_j = 1/j \) confirms.

In the general case of \( s_n^2 \to \infty \), note that \( f_{X_j}^{(1)} \) are uniformly bounded with \( |X_j - p_j| \leq 1 \), and hence since \( s_n^2 \to \infty \),
\[
\frac{1}{s_n^3} \sum_{j=1}^n E \left[ |X_j - p_j|^3 \right] \leq \frac{s_j^2}{s_n^3} \to 0.
\]
Thus 7.12 is always satisfied in this case with \( \delta = 1 \).

Exercise 7.17 Show that the above example generalizes. If \( \{X_j\}_{j=1}^\infty \) are uniformly bounded and \( s_n^2 \to \infty \), then the central limit theorem applies.

7.1.4 A Central Limit Theorem on \( \mathbb{R}^n \)

Recall definition 6.26 that a random vector \( Y \equiv (Y_1, Y_2, \ldots, Y_n) \) is said to have multivariate normal distribution, or a multivariate Gaussian distribution, if there exists a symmetric, positive semidefinite \( n \times n \) matrix \( C \) and an \( n \)-vector \( \mu \) so that the characteristic function of \( Y \) has the form 6.23:
\[
C_Y(t) = \exp \left[ i\mu \cdot t - \frac{1}{2} t^T Ct \right].
\]
It is also common to say that the probability measure induced by this distribution is multivariate normal. As noted above, \( C \) is the covariance matrix of the component variates and \( \mu \) is the mean vector. The purpose of this section is to derive a multivariate version of the central limit theorem, and we do so using the Cramér-Wold device or the Cramér-Wold theorem of proposition 6.32. This very short proof provides a good example of the power and utility of this earlier result.

To this end, let \( \{X_m\}_{m=1}^\infty \) be a collection of independent random vectors with range in \( \mathbb{R}^n \) and with a common distribution function. In short, \( \{X_m\}_{m=1}^\infty \) are i.i.d., independent and identically distributed. Let \( X_m \equiv (X_{m1}, X_{m2}, \ldots, X_{mn}) \), and assume that \( \sigma_{jj} \equiv E \left[ X_{mj}^2 \right] < \infty \) for all \( j \). Note that this notational change for variance, \( \sigma_{jj} \equiv C_{jj} \) instead of \( \sigma_j^2 \), is common
and convenient when we also use the covariances, denoted \( \sigma_{ij} \equiv C_{ij} \). Then \( \nu_j \equiv E[X_{mj}] < \infty \) by Lyapunov’s inequality in corollary 3.51, and \( \sigma_{ij} \equiv E[(X_{mi} - \nu_i)(X_{mj} - \nu_j)] < \infty \) by the Cauchy-Schwarz inequality as applied in corollary 3.50, both of book 4.

**Proposition 7.18 (Central Limit Theorem on \( \mathbb{R}^n \))** With the notation above, let \( C \) denote the \( n \times n \) covariance matrix of component variates, and \( \nu = (\nu_1, \nu_2, \ldots, \nu_n) \). Define \( S_m \equiv \sum_{j=1}^{m} X_j \), and

\[
Y_m \equiv (S_m - m\nu) / \sqrt{m}.
\]

Then as \( m \to \infty \),

\[
F_{Y_m} \Rightarrow \Phi_C,
\]

(7.15)

where \( \Phi_C \) is the distribution function of the multivariate normal as in 6.23 with covariance matrix \( C \) and \( \mu = 0 \):

\[
C_{\Phi}(t) = \exp \left( -\frac{1}{2} t^T C t \right).
\]

**Remark 7.19** As is the case for the standard versions of the central limit theorem above, this theorem also provides a weak convergence result for the average random vector, \( S_m \equiv \sum_{j=1}^{m} X_j / \sqrt{m} \), since

\[
(S_m - m\nu) / \sqrt{m} = (S_m - \nu) \sqrt{m}.
\]

**Proof.** Let \( Z \equiv (Z_1, Z_2, \ldots, Z_n) \) denote a multivariate normal random vector with covariance matrix \( C \) and \( \mu = 0 \). By the Cramér-Wold theorem, the desired conclusion that \( Y_m \Rightarrow Z \) will follow if it can be shown that for all \( t \in \mathbb{R}^n \) that \( t \cdot Y_m \Rightarrow t \cdot Z \). To this end, define \( X'_j = t \cdot X_j \). Then \( E[X'_j] = t \cdot \nu \equiv \nu' \), and by independence:

\[
\text{Var}[X'_j] = \sum_{j=1}^{m} \sum_{i=1}^{m} t_i t_j \sigma_{ij} = t^T C t.
\]

Also, with \( \sigma_t = \sqrt{t^T C t} \),

\[
(t \cdot Y_m) / \sigma_t = \left( \sum_{j=1}^{m} X'_j - m\nu' \right) / (\sigma_t \sqrt{m}).
\]

By the classical central limit theorem of proposition 7.2, \( (t \cdot Y_m) / \sigma_t \Rightarrow Z' \), where \( Z' \) is a standard normal variate. This obtains that \( t \cdot Z \) has mean 0 and variance \( \sigma_t^2 \), either directly or by Slutsky’s theorem of proposition 5.29 of book 2. But the same calculation shows that \( t \cdot Z \) has mean 0 and variance \( \sigma_t^2 \), and is a normal variate by proposition 3.9. Thus \( t \cdot Y_m \Rightarrow t \cdot Z \) for all \( t \in \mathbb{R}^n \) and the Cramér-Wold theorem completes the proof. \( \blacksquare \)
190 CHAPTER 7 APPLICATIONS OF CHARACTERISTIC FUNCTIONS

7.2 Distribution Families Related Under Addition

It is common in applications to seek the distribution function of a sum of independent random variables given the individual distributions. In many applications, these random variables will also be identically distributed. This investigation was initiated in section 1.4 and extended in example 3.59 of book 4. The direct though cumbersome way to determine this distribution function is by evaluating the convolution of the distribution functions as in 2.10, or more typically the convolution of the associated densities as in 2.13. But this approach can quickly become tedious, even for a very simple example like the uniform distribution on [0, 1].

A far more elegant approach is to use moment generating functions or characteristic functions. For sums of independent random variables \( \{X_i\}_{i=1}^m \), say \( X = \sum_{i=1}^m X_i \), we have from section 3.2.3 of book 4 that if \( M_{X_i}(t) \) exists for \( t \in (-t_0, t_0) \), then \( M_X(t) \) exists on this same interval and:

\[
M_X(t) = \prod_{i=1}^m M_{X_i}(t).
\]

Similarly, since \( C_{X_i}(t) \) always exists for all \( t \), we have from 6.32 that

\[
C_X(t) = \prod_{i=1}^m C_{X_i}(t).
\]

Consequently, if we know \( \{M_{X_i}(t)\}_{i=1}^m \), respectively \( \{C_{X_i}(t)\}_{i=1}^m \), then we know \( M_X(t) \), respectively \( C_X(t) \). Consequently, the distribution function for \( X \) can potentially be found by inspection if this moment generating function or characteristic function is recognizable.

**Remark 7.20** Note that recognition of \( M_X(t) \), respectively \( C_X(t) \), is adequate to confirm the distribution function because of the uniqueness results proved in corollary 3.58 of book 4 (requiring proposition 6.44 above), respectively proposition 6.25. Specifically, if \( M_X(t) = M_Y(t) \) on \( (-t_0, t_0) \), or \( C_X(t) = C_Y(t) \) for all \( t \), then \( X \) and \( Y \) have the same distribution function.

Of course this approach does not always work. If \( \{X_i\}_{i=1}^m \) are independent Student \( T \) random variables, \( M_{X_i}(t) \) does not exist and the available formula for \( C_{X_i}(t) \) is unwieldy, being defined in terms of a power series, and thus unwieldy enough to make \( \prod_{i=1}^m C_{X_i}(t) \) very difficult to identify. But in this case, the convolution approach is also unwieldy, and there appears to be no accessible approach to working with sums of independent Student T random variables.
7.2 DISTRIBUTION FAMILIES RELATED UNDER ADDITION

T variables. Even when \( M_X(t) \) and/or \( C_X(t) \) are readily calculated, the resultant \( M_X(t) \) or \( C_X(t) \) may be unrecognizable. In the special case where \( C_X(t) \) is integrable, we can in theory recover the density function for \( X \) by the inversion formula in 6.36, but this may or may not be successful.

In the general case of independent and identically distributed random variables with identifiable distributions, we will say that the distribution family of \( \{X_i\}_{i=1}^m \) is related under addition to the distribution family of \( X \) if \( X = \sum_{i=1}^m X_i \). This is not a precise definition until "distribution family" is defined, and for identifiable distributions we will always mean, the collection of distribution functions formed by all valid choices of defining parameters. In some cases these distribution families will coincide, and we then say that the given distribution family is closed under addition. In both cases the terminology is sometimes extended to cases where the \( \{X_i\}_{i=1}^m \) are independent and from the same distribution family but with different parameters. Interestingly, some distribution families are related under multiplication.

In this section, we identify some useful relationships. Examples of closed and related distribution families follow, as well families for which these methods fail. Consistent with the above section on examples of characteristic functions, we split the section into discrete and continuous distributions.

7.2.1 Discrete Distributions

1. Discrete Rectangular Distribution on \([0, 1] \):

   If \( R_j \) has a discrete rectangular distribution on \([0, 1] \) with parameter \( n \) then from 6.11:

   \[
   C_{R_j}(t) = \frac{\exp[i(1 + 1/n)t] - \exp[it/n]}{n(\exp[it/n] - 1)},
   \]

   and hence \( R = \sum_{j=1}^m R_j \) has characteristic function:

   \[
   C_R(t) = \left( \frac{\exp[i(1 + 1/n)t] - \exp[it/n]}{n(\exp[it/n] - 1)} \right)^m.
   \]

   This characteristic function does not lend itself to ready identification unless \( n = 1 \), in which case this distribution reduces to an example the delta function of 6.13 with \( x_0 = 1 \).

   In general, the delta distribution family is closed under addition. If \( R_j \) has parameter \( x_j \) then \( R \) has parameter \( x_0 = \sum_{j=1}^m x_j \). But certainly, one does not need characteristic functions to elucidate this fact.
2. Binomial Distribution:

If $B_j$ has a binomial distribution with parameters $p$ and $n \in \mathbb{N}$, then from 6.14:

$$C_{B_j}(t) = \left(1 + p(e^{it} - 1)\right)^n,$$

and $B = \sum_{j=1}^{m} B_j$ has characteristic function:

$$C_B(t) = \left(1 + p(e^{it} - 1)\right)^{nm}.$$

Thus $B$ is binomial with parameters $p$ and $nm$. So the binomial distribution family is closed under addition when all variates have the same parameters $p$ and $n$. When the $B_j$ have parameters $p$ and $n_j$, then $B$ is again binomial with parameters $p$ and $\sum_{j=1}^{m} n_j$, and so this family is also closed in this more general sense. When $B_j$ have parameters $p_j$ and $n_j$, then $B$ is no longer binomial and its distribution is not readily identifiable.

3. Geometric Distribution:

If $G_j$ has a geometric distribution with parameter $p$, where $0 < p < 1$, then from 6.15:

$$C_{G_i}(t) = \frac{p}{1 - (1 - p)e^{it}},$$

and $G = \sum_{j=1}^{m} G_j$ has characteristic function:

$$C_{G}(t) = \left(\frac{p}{1 - (1 - p)e^{it}}\right)^m.$$

This is recognizable from 6.16 as the characteristic function of the negative binomial with parameters $p, m$. Thus the geometric distribution family is related under addition to the negative binomial distribution family. This is a special case of the next result.

4. Negative Binomial Distribution:

As noted in 3 above, if $NB_j$ has a negative binomial distribution with parameters $p$ and positive integer $k_j$:

$$C_{NB_j}(t) = \left(\frac{p}{1 - (1 - p)e^{it}}\right)^{k_j},$$

and $NB = \sum_{j=1}^{m} NB_j$ is apparently negative binomial with parameters $p$ and $\sum_{j=1}^{m} k_j$. So the negative binomial family is closed
7.2 DISTRIBUTION FAMILIES RELATED UNDER ADDITION

under addition, as long as \( p \) is fixed. Result 3 is a special case of this because the geometric distribution is negative binomial with \( k = 1 \). When the \( p \) parameter is not fixed, the distribution of \( NB \) is not readily identifiable.

5. Poisson Distribution:

If \( P_j \) is Poisson with parameter \( \lambda_j \), then from 6.17:

\[
C_{P_j}(t) = \exp[\lambda_j(e^{it} - 1)],
\]

and \( P = \sum_{j=1}^{m} P_j \) has characteristic function \( C_P(t) = \exp[\lambda(e^{it} - 1)] \) with \( \lambda = \sum_{j=1}^{m} \lambda_j \). Hence, the Poisson family is closed under addition.

7.2.2 Continuous Distributions

1. Continuous Uniform Distribution:

If \( U_j \) is uniform on \([a, b]\), then from 6.18:

\[
C_{U_j}(t) = \frac{e^{ibt} - e^{iat}}{t(b-a)},
\]

and like the discrete version, \( U = \sum_{j=1}^{m} U_j \) has an unrecognizable characteristic function.

2. Exponential Distribution

If \( E_j \) is exponential with parameter \( \lambda \), then from 6.19:

\[
C_{E_j}(t) = (1 - it/\lambda)^{-1}.
\]

Thus \( E = \sum_{j=1}^{m} E_j \) has characteristic function \( C_E(t) = (1 - it/\lambda)^{-m} \), which is gamma with parameters \( \lambda, m \). So for fixed \( \lambda \), the exponential distribution family is related under addition to the gamma distribution family. As for the geometric in relation to the negative binomial, this result is a special case of the next.

3. Gamma Distribution

If \( \Gamma_j \) is gamma with parameters \( \lambda, \alpha_j \), then from 6.19:

\[
C_{\Gamma_j}(t) = (1 - it/\lambda)^{-\alpha_j},
\]
and $\Gamma = \sum_{j=1}^{m} \Gamma_j$ has the characteristic function of a gamma with parameters $\lambda$ and $\alpha = \sum_{j=1}^{m} \alpha_j$. Hence, the gamma distribution family is closed under addition, as long as $\lambda$ is fixed. As noted in remark 1.8 of book 4, when $\lambda = 1/2$ and $\alpha = n/2$, a gamma variate is called a chi-squared random variable with $n$ degrees of freedom, denoted $\chi^2_n$. Thus the chi-squared distribution family is closed under addition, since $\lambda$ is fixed by definition.

4. Beta Distribution

If $\beta$ has a beta distribution with parameters $v > 0$ and $w > 0$, it has the unwieldy characteristic 6.20:

$$C_\beta(t) = 1 + \sum_{n=1}^{\infty} \prod_{j=0}^{n-1} \frac{v + j}{v + w + j} \left(\frac{it}{n!}\right)^n,$$

and it is apparent that products of such functions will not be readily identifiable.

5. Cauchy Distribution

If $C_j$ is Cauchy with parameters $\gamma_j$ and $x_j$, then from 6.21:

$$C_{C_j}(t) = \exp \left( ix_j t - \gamma_j |t| \right),$$

and so $C = \sum_{j=1}^{m} C_j$ has characteristic function $C_C(t) = \exp \left( ix_0 t - \gamma |t| \right)$ with $x_0 = \sum_{j=1}^{m} x_j$ and $\gamma = \sum_{j=1}^{m} \gamma_j$. Hence, the Cauchy family is closed under addition. From this observation, we have the somewhat surprising result:

**Corollary 7.21** If $\{C_j\}_{j=1}^{m}$ are independent Cauchy with parameters $\gamma$ and $x_0$, then $\bar{C} = \sum_{j=1}^{m} C_j/m$ is Cauchy with the same parameters.

**Proof.** Apply the above calculation and 6.25.  

Hence, not only does the Cauchy distribution have no moments, but averaging independent Cauchy variates produces the same distribution, and so there is no hope that a central limit theorem could ever apply.

6. Normal Distribution
If \( N_j \) is normal with parameters \( \mu_j, \sigma_j^2 \), then by 6.22:

\[
C_{N_j}(t) = \exp\left( i\mu_j t - \frac{1}{2} \sigma_j^2 t^2 \right),
\]

and hence \( N = \sum_{j=1}^{m} N_j \) has moment generating function \( M_N(t) = \exp(\mu t + \frac{1}{2} \sigma^2 t^2) \) with \( \mu = \sum_{j=1}^{m} \mu_j \) and \( \sigma^2 = \sum_{j=1}^{m} \sigma_j^2 \). Thus the normal family is closed under addition.

### 7.3 Infinitely Divisible Distributions

It will not have been lost on the reader that the process of the previous section, of adding i.i.d. variates from one distribution family to obtain a variate from the same or a different distribution family, can sometimes be reversed. Namely, given \( X \) and \( n \) it is sometimes possible to find independent and identically distributed \( \{X_i\}_{i=1}^{m} \) so that \( X = \sum_{i=1}^{m} X_i \).

**Definition 7.22** A random variable (or vector) \( X \) is said to be **infinitely divisible**, and sometimes the associated distribution (or joint distribution) function \( F \) is said to be **infinitely divisible**, if for any \( n \in \mathbb{N} \) there exists independent, identically distributed random variables (or vectors) \( \{X_j^{(n)}\}_{j=1}^{n} \), so that:

\[
X = d \sum_{j=1}^{n} X_j^{(n)}. \tag{7.16}
\]

Here "\( d \)" means "equal in distribution."

The criterion in 7.16 can be equivalently expressed in other ways. Given \( X \) and \( F \), proposition 2.8 assures that \( F \) is infinitely divisible if for every \( n \in \mathbb{N} \) there exists a distribution function \( F_\otimes^{(n)} \) so that recalling notation 2.11:

\[
F(x) = F_\otimes^{(n)} * F_\otimes^{(n)} * \ldots * F_\otimes^{(n)}(x), \quad n\text{-times},
\]

or equivalently by proposition 6.13:

\[
C_F(t) = \left[ C_{F_\otimes^{(n)}}(t) \right]^n. \tag{7.17}
\]

Proposition 1.7 then assures the existence of random variables \( X^{(n)} \), each defined on some probability space, so that \( F_\otimes^{(n)} = F_\otimes^{(n)} \). Given a sample \( \{X_j^{(n)}\}_{j=1}^{n} \) of such variates (recall chapter 4 of book 2) for fixed \( n \) and defining:

\[
X' = \sum_{j=1}^{n} X_j^{(n)},
\]
proposition 2.8 identifies \( F \) as the distribution function of \( X' \), and thus \( X' = d X \), and this is 7.16.

Also equivalently, \( F \) is infinitely divisible if \( \left[ C_F(t) \right]^{1/n} \) is a characteristic function for any \( n \). Then by definition there exists associated distribution functions \( F_{(n)} \) with \( C_{F_{(n)}}(t) = \left[ C_F(t) \right]^{1/n} \), and by proposition 1.7 there exists associated random variables \( X_{(n)} \) so that \( F_{X_{(n)}} = F_{(n)} \). Thus \( C_F(t) \) is the characteristic function of \( X' \) defined above by proposition 6.20, and thus by uniqueness, \( F \) is the distribution function of \( X' \).

An important subset of the infinitely divisible distributions is the class of stable distributions. We will define them here and apply them in the discussion of discrete time asset models, but will otherwise not develop any of their special properties. See Sato (1999) for more details.

**Definition 7.23** A random variable (or vector with range in \( \mathbb{R}^m \)) \( X \) is said to be stable, and sometimes the associated distribution (or joint distribution) function \( F \) is said to be stable, if for any \( n \in \mathbb{N} \) there exists real \( c_n > 0 \) and \( d_n \in \mathbb{R}^m \) so that:

\[
\sum_{j=1}^n X_j = d \cdot c_n X + d_n, \tag{7.18}
\]

where \( \{X_j\}_{j=1}^n \) are independent, identically distributed random variables (or vectors) with distribution \( F \). The variate \( X \) is strictly stable if 7.18 is satisfied with \( d_n = 0 \).

Equivalently by proposition 6.18, given \( F \), \( C_F(t) \) and \( n \), then \( X \) is stable if there exists \( c_n > 0 \) and \( d_n \in \mathbb{R}^m \) so that:

\[
\left[ C_F(t) \right]^{1/n} = e^{t \cdot d_n} C_F(c_n t). 
\]

**Exercise 7.24** By considering representations of sums in 7.18 to \( n, m, \) and \( mn \), show that:

\[
c_{mn} = c_n c_m, \quad d_{mn} = c_m d_n + nd_m. \tag{7.19}
\]

**Note:** For a stable variate, it is actually the case that \( c_n = n^{1/\alpha} \) for \( 0 < \alpha \leq 2 \). See section VI.1 of Feller (1971) for the proof; see exercise 7.27 below for examples.

**Remark 7.25** All stable distributions are infinitely divisible by choosing \( X_{(n)}^{(n)} = \left( \frac{X_j - d_n/n}{c_n} \right) \). The converse is not true, and exercise 7.27 provides at least one example.
While the general question related to identifying all infinitely divisible distributions is a difficult one, the prior section provides many examples.

**Example 7.26** 1. **Degenerate Distribution**: Defined to assign measure 1 to a single point $x_0$, the associated characteristic function is $C_D(t) = e^{ix_0t}$ in one-dimension, $C_D(t) = e^{ix_0t}$ in multivariate notation. This simple distribution is infinitely divisible since $[C_D(t)]^{1/n}$ is the characteristic function of the degenerate distribution with parameter $x_0/n$. Note that when $x_0 = 0$, $C_D(t) = 1$ and for all $n$, $X_1^{(n)} = X$ in the above definition. By uniqueness, this is the only distribution function for which $C_D(t) = 1$ for all $t$.

2. **Poisson Distribution**: By 6.17 the characteristic function of the Poisson with parameter $\lambda > 0$ is:

$$C_P(t) = \exp[\lambda(e^{it} - 1)],$$

and thus the Poisson is infinitely divisible because $[C_P(t)]^{1/n}$ is the characteristic function of the Poisson distribution with parameter $\lambda/n$.

3. **Gamma Distribution**: By 6.19 the characteristic function of the gamma with parameters $\lambda$ and $\alpha$ is:

$$C_T(t) = (1 - it/\lambda)^{-\alpha},$$

and thus the Gamma is infinitely divisible because $[C_T(t)]^{1/n}$ is the characteristic function of the gamma distribution with parameters $\lambda$ and $\alpha/n$.

As a corollary to this, the **Exponential Distribution** with parameter $\lambda$ is also infinitely divisible, but with $[C_E(t)]^{1/n}$ the characteristic function of the gamma distribution with parameters $\lambda$ and $1/n$.

4. **Cauchy Distribution**: By 6.21 the characteristic function of the Cauchy with parameters $\gamma$ and $x$ is:

$$C_C(t) = \exp(ixt - \gamma|t|),$$

and thus the Cauchy is infinitely divisible because $[C_C(t)]^{1/n}$ is the characteristic function of the Cauchy distribution with parameters $\gamma/n$ and $x/n$. 

5. **Normal Distribution:** By 6.22 the characteristic function of the normal distribution with parameters $\sigma > 0$ and $\mu$ and is given by:

$$C_N(t) = \exp \left( i\mu t - \frac{1}{2} \sigma^2 t^2 \right),$$

and thus the normal is infinitely divisible because $[C_N(t)]^{1/n}$ is the characteristic function of the normal distribution with parameters $\mu/n$ and $\sigma/\sqrt{n}$.

6. **Multivariate Normal Distribution:** By 6.23, the characteristic function of the multivariate normal random vector $Y = (Y_1, Y_2, \ldots, Y_n)$ is given by:

$$C_{MN}(t) = \exp \left[ i \mu \cdot t - \frac{1}{2} t^T C t \right],$$

where $C$ is the covariance matrix of the random vector $(Y_1, Y_2, \ldots, Y_n)$ and defined by $C_{ij} = E[(Y_i - \mu_i)(Y_j - \mu_j)]$, and $\mu$ is the vector of first moments, $\mu_j = E[Y_j]$. Thus the multivariate normal is infinitely divisible because $[C_{NB}(t)]^{1/n} = \exp \left[ i (\mu/n) \cdot t - \frac{1}{2} t^T (C/n) t \right]$ is the characteristic function of a multivariate normal, noting that $C/n$ is positive semidefinite.

7. **Negative Binomial Distribution:** By 6.16 the characteristic function of the negative binomial with parameters $p$ and positive integer $k$ is:

$$C_{NB}(t) = \left( \frac{p}{1 - (1 - p)e^{it}} \right)^k,$$

and thus $[C_{NB}(t)]^{1/n}$ looks a lot like a negative binomial, but with parameters $p$ and $k/n$, where the latter value is in general a non-integer. The same result occurs with the **Geometric Distribution** where $C_G(t) = C_{NB}(t)$ with $k = 1$. As it turns out, both are infinitely divisible distributions.

Recall that $k$ is an integer in the classical definition of the negative binomial, and the density function:

$$f_{NB}(j) = \binom{j + k - 1}{k - 1} p^k (1 - p)^j, \quad j = 0, 1, 2, \ldots$$

can be interpreted as the probability of $j$ tails before the $k$th head in a series of coin flips where the probability of a head is $p$. But mathematically, $f_{NB}(j)$ is a perfectly valid density function for any real $k > 0$. 

This follows from Newton's generalized binomial theorem, named for Isaac Newton (1642 – 1727):

\[(1 - x)^{-k} = \sum_{j=0}^{\infty} \binom{j + k - 1}{k - 1} x^j, \quad |x| < 1,\]

where

\[\binom{j + k - 1}{k - 1} \equiv \binom{j + k - 1}{j} \equiv \frac{(j + k - 1)(j + k - 2) \cdots k}{j!}.
\]

Letting \(x = 1 - p\) obtains that \(f_{NB}(j)\) is a probability density for all \(k > 0\), and \(C_{NB}(t)\) given above is the associated characteristic function.

Thus \([C_{NB}(t)]^{1/n}\) is the characteristic function of a generalized negative binomial with parameters \(p\) and \(k/n\), while the same is true for \([C_G(t)]^{1/n}\) but with parameters \(p\) and \(1/n\).

8. Compound Poisson Distribution: Note that the Poisson random variable \(X_P\) with parameter \(\lambda\) can be formally defined as a "random" sum:

\[X_P = \sum_{j=1}^{N} X_j, \quad (7.20)\]

where \(N\) has the given Poisson distribution, and \(X_j \equiv D_j\) are independent degenerate random variables with \(x_0 = 1\) and thus \(C_D(t) = e^{it}\).

This is intuitive, but should be formally checked by calculating the characteristic function of this random variable. For this we use the law of total expectation in 5.19. Thus:

\[E[e^{iX_Pt} | N] = [C_D(t)]^N = e^{itN},\]

since conditional on \(N\), \(X_P\) is an i.i.d. summation with characteristic function equal to the product of the component variate characteristic functions by proposition 6.13. Thus by 5.19:

\[C_{X_P}(t) = E[e^{itN}] = e^{-\lambda} \sum_{n=0}^{\infty} \frac{e^{itn}\lambda^n}{n!} = \exp[\lambda(e^{it} - 1)],\]

so by 6.17 and proposition 6.14 uniqueness, \(X_P\) is Poisson.

A compound Poisson random variable or random vector \(X_{CP}\) is defined exactly as in 7.20 but where the \(\{X_j\}\) is any independent,
identically distributed collection of random variables or vectors. Repeating the above derivation using general vector notation obtains:

\[ E \left[ e^{it \cdot X_{CP}^j} \right] = \left[ C_X(t) \right]^N, \]

and

\[ C_{CP}(t) = \exp[\lambda(C_X(t) - 1)]. \]  

(7.21)

From this it is clear that \( X_{CP} \) is infinitely divisible, and \( [C_{CP}(t)]^{1/n} \) is the characteristic function of a compound Poisson with parameters \( \lambda/n \) and the same component variates/vectors \( \{X_j\} \).

For reasons that will be clear below, \( C_{CP}(t) \) is sometimes written as a Riemann-Stieltjes integral:

\[ C_{CP}(t) = \exp \left[ \lambda \int_{\mathbb{R}^n} (e^{it \cdot x} - 1) \, dF_X \right], \]  

(7.22)

where \( F_X \) denotes the joint distribution function of \( X \), or in the corresponding Lebesgue-Stieltjes notation with \( d\mu_X \).

Exercise 7.27 Identify which of the above distributions are stable, and determine the parameters \( c_n > 0 \) and \( d_n \) in 7.18.

With the exception of example 7, demonstrating that the above random variables were infinitely divisible was generally very easy. Looking at other examples then proves how hard this identification can be.

Example 7.28 The binomial distribution with parameters \( p \) and \( m \) has characteristic function given in 6.14 by \( C_B(t) = (1 + p(e^{it} - 1))^m \), and thus \( [C_B(t)]^{1/n} = (1 + p(e^{it} - 1))^{m/n} \). While not the characteristic function of a standard binomial, one wonders if there is a definitional extension of the binomial as there was for the negative binomial, which then has this characteristic function. As it turns out, the answer is "no" and the binomial is not infinitely divisible because the random variable \( B \) is bounded (see result 5 of proposition 7.29).

As another example, the continuous uniform distribution on \( [a - r, a + r] \) has characteristic function given in 6.14 by:

\[ C_U(t) = e^{iat} \frac{e^{irt} - e^{-irt}}{2rt} = e^{iat} \frac{\sin rt}{rt}. \]

Again it is not obvious if \( [C_U(t)]^{1/n} \) is a characteristic function. But in this case we can conclude that it is not. Result 1 of proposition 7.29 below states
that the characteristic function of an infinitely divisible distribution can have no real zeros. That is, \( C(t) \neq 0 \) for all \( t \in \mathbb{R} \). This is a necessary but not sufficient condition. Certainly \( C_U(t) \) has real zeros, and so a conclusion in the negative is possible, that the continuous uniform distribution is not infinitely divisible. This random variable is also bounded so 5 also applies to obtain this conclusion.

Note that the binomial \( C_B(t) \) has no real zeros, but is still not infinitely divisible by 5 below. Thus, the property in 1 that \( C(t) \neq 0 \) for all \( t \in \mathbb{R} \), is necessary for infinite divisibility but is not sufficient.

**Proposition 7.29** Some of the properties of infinitely divisible (I.D.) distribution functions on \( \mathbb{R}^m \) follow.

1. If \( F \) is I.D., then \( C_F(t) \neq 0 \) for all \( t \in \mathbb{R}^m \).

2. If \( F_1 \) and \( F_2 \) are I.D, then \( F \equiv F_1 + F_2 \) is I.D., recalling notation 2.11.
   
   (a) Equivalently by proposition 2.8, if \( X_1 \) and \( X_2 \) are I.D random vectors, then \( X \equiv X_1 + X_2 \) is I.D.
   
   (b) Equivalently by proposition 6.20, if \( C_1(t) \) and \( C_2(t) \) are I.D characteristic functions, then \( C(t) \equiv C_1(t)C_2(t) \) is I.D.

   These results are then true for all finite combinations (convolutions, sums, products).

3. If \( \{F_j\} \) are I.D. and \( F_j \Rightarrow F \), then \( F \) is I.D.
   
   (a) Equivalently by definition 4.6, if \( \{X_j\} \) are I.D. random vectors and \( X_j \Rightarrow X \), then \( X \) is I.D.
   
   (b) Equivalently by proposition 6.30, if \( \{C_j(t)\} \) are I.D. characteristic functions and there exists a characteristic function \( C(t) \) so that \( C_j(t) \rightarrow C(t) \) for all \( t \), then \( C(t) \) is I.D.

4. \( C(t) \) is I.D. if and only if \( [C(t)]^\alpha \) is I.D. for all real \( \alpha > 0 \).

5. If \( X \) is I.D. and bounded, then \( X \) is degenerate and \( C_X(t) = C_D(t) \equiv e^{ix_0t} \) for some \( x_0 \).

**Proof.** We provide a proof of each in turn, leaving the details of the equivalent statements to the reader.

For 1, if \( F \) is I.D. then by definition \( [C_F(t)]^{1/n} \) is a characteristic function for all \( n \). Consider \( g(t) = \lim_{n \to \infty} [C_F(t)]^{1/n} \). Then clearly \( g(t) = 0 \) if
Chapter 7 Applications of Characteristic Functions

Let $C_F(t) = 0$, and it is the case that $g(t) = 1$ otherwise. To see this, note that if $C_F(t) \neq 0$, then $\ln C_F(t) = a + bi$ is well defined by specifying this as the principal value, meaning $b \in (-\pi, \pi]$. By Euler’s formula (6.12, book 5):

$$[C_F(t)]^{1/n} = e^{a/n}(\cos(b/n) + i \sin(b/n)) \to 1.$$ 

Now $C_F(0) = 1$ and by continuity there is a real rectangle $R = \prod_{j=1}^{m} (-\epsilon, \epsilon)_j$ so that $C_F(t) \neq 0$ on $R$, and thus $g(t) = 1$ on $R$ and $g(t)$ is continuous at $t = 0$. By corollary 6.31 to Lévy’s continuity theorem, $g(t)$ is a characteristic function and thus continuous. It now follows that $g(t) = 1$ on $\mathbb{R}^m$, since the only other value possible is 0, and thus $C_F(t) \neq 0$ on $\mathbb{R}^m$.

For 2, that $F_1$ and $F_2$ are I.D implies that $[C_{F_1}(t)]^{1/n}$ and $[C_{F_2}(t)]^{1/n}$ are characteristic functions for all $n$. Specifically, $[C_{F_1}(t)]^{1/n} = C_{F_1^{(n)}}(t)$, where $F_1^{(n)}$ is the distribution function of $X_1^{(n)}$, and similarly $[C_{F_2}(t)]^{1/n} = C_{F_2^{(n)}}(t)$. But $C_{F_1^{(n)}}(t)C_{F_2^{(n)}}(t)$ is the characteristic of $F_1^{(n)} \ast F_2^{(n)}$ by propositions 2.8 and 6.20, and similarly $C_{F_1}(t)C_{F_2}(t)$ is the characteristic of $F = F_1 \ast F_2$. Thus $F$ is I.D since $[C_F(t)]^{1/n} = C_{F^{(n)}}(t)$. Extension to finite convolutions is now true by the same argument, since proposition 6.20 is applicable to all finite convolutions.

Now $F_j \Rightarrow F$ assures that $C_{F_j}(t) \rightarrow C_F(t)$ for all $t$ by Lévy’s continuity theorem. Thus for given $n$, $[C_{F_j}(t)]^{1/n} \rightarrow [C_F(t)]^{1/n}$ for all $t$ and since $[C_{F_j}(t)]^{1/n}$ is a characteristic function by definition and $[C_F(t)]^{1/n}$ is continuous at $t = 0$, Lévy’s continuity theorem assures that $[C_F(t)]^{1/n}$ is a characteristic function. This is true for all $n$ and hence $F$ is I.D. and 3 is proved.

Only the "only if" statement in 4 needs proof. If $C(t)$ is infinitely divisible and $\alpha = p/q$ is rational, then $[C^\alpha(t)]^{1/n} = [C^p(t)]^{1/nq}$. Now $C^p(t)$ is infinitely divisible by 2, and thus $[C^p(t)]^{1/nq}$ is a characteristic function for all $n$ by definition. So $C^\alpha(t)$ is infinitely divisible for all positive rationals $\alpha$. For general $\alpha$ and $p_j/q_j \to \alpha$, $C^p_j/q_j(t) \to C^\alpha(t)$ for all $t$ and $C^\alpha(t)$ is infinitely divisible by 3.

For 5, that $X$ is bounded means that $\Pr[|X| > M] = 0$ for some $M < \infty$, where as always $\Pr[|X| > M] = \lambda [X^{-1}((-\infty, M) \cup (M, \infty))]$ with $\lambda$ the measure on the probability space on which $X$ is defined. This implies that for $X^{(n)}$ in 7.16 that $\Pr[|X^{(n)}| > M/n] = 0$. Otherwise if $\Pr[|X^{(n)}| > M/n] = p > 0$, then since $X = \sum_{j=1}^{n} X_j^{(n)}$, an independent sum:

$$\Pr[|X| > M] \geq \Pr \left( \bigcap_{j=1}^{n} \left[ |X_j^{(n)}| > M/n \right] \right) = p^n > 0,$$
a contradiction. Now

\[ \text{Var}[X_j^{(n)}] = E\left[\left(X_j^{(n)}\right)^2\right] - \left(E\left(X_j^{(n)}\right)\right)^2 \leq E\left[\left(X_j^{(n)}\right)^2\right] \leq \left(\frac{M}{n}\right)^2, \]

and thus by independence:

\[ \text{Var}[X] = \sum_{j=1}^{n} \text{Var}[X_j^{(n)}] \leq n \left(\frac{M}{n}\right)^2. \]

Letting \( n \to \infty \) obtains that \( \text{Var}[X] = 0 \) and \( X \) is degenerate by Chebyshev’s inequality of proposition 3.32 of book 4.

A simple corollary of the proof of part 1 is that the distribution functions of the random variables \( \{X^{(n)}\} \) in definition 7.22 converge weakly to the degenerate distribution with \( x_0 = 0 \).

**Corollary 7.30** If \( X \) is an infinitely divisible random vector on \( \mathbb{R}^m \), then with \( X^{(n)} \) as in definition 7.22, \( X^{(n)} \Rightarrow X_D \), the degenerate random variable of 1 of example 7.26 with \( x_0 = 0 \) in \( \mathbb{R}^m \).

**Proof.** By definition, \( [C_F(t)]^{1/n} \equiv C_{X^{(n)}}(t) \) the characteristic function of \( X^{(n)} \), and by the proof of 1 of proposition 7.29, \( C_{X^{(n)}}(t) \) converges to a characteristic function \( g(t) \). But \( g(t) = 1 \) on \( \mathbb{R}^m \), which is the characteristic function of \( X_D \). Thus by uniqueness, \( g(t) = C_D(t) \) and \( C_{X^{(n)}}(t) \to C_D(t) \) for all \( t \). Lévy’s continuity theorem now obtains that \( X^{(n)} \Rightarrow X_D \).

The next result provides an early characterization of infinitely divisible random variables by Bruno de Finetti (1906 – 1985), who is credited with introducing the study of infinite divisibility in 1929. De Finetti’s theorem in essence states that infinitely divisible distributions are characterized as the weak limit of compound Poisson distributions, recalling 8 of example 7.26. Stated another way using Lévy’s continuity theorem, infinitely divisible characteristic functions are characterized as the limits of compound Poisson characteristic functions.

For this proof we need a simple but important result.

**Lemma 7.31** If \( C(t) \) is a characteristic function of a distribution function \( F \) on \( \mathbb{R}^m \) and \( \lambda > 0 \), then \( g(t) \equiv \exp\left[\lambda (C(t) - 1)\right] \) is an infinitely divisible characteristic function.

**Proof.** Let \( F \) be the distribution function for which \( C_F(t) = C(t) \), and \( X \) a random vector with distribution \( F \) as given in proposition 1.7. Also,
let \( \{X_j\}_{j=1}^\infty \) be independent and identically distributed random vectors with distribution \( F \) as given by proposition 4.4 of book 2.

Then by 8 of example 7.26, \( g(t) \) is the infinitely divisible characteristic function of a compound Poisson random vector defined by \( \lambda \) and \( F \) as in 7.21, and this distribution is uniquely defined by proposition 6.25. ■

We state this result in the context of convergence of characteristic functions. But note that by Lévy's continuity theorem of proposition 6.30, that this is equivalent to the statement of weak convergence of the associated distribution functions.

**Proposition 7.32 (de Finetti's theorem)** A characteristic function \( C(t) \) on \( \mathbb{R}^m \) is infinitely divisible if and only if there exists a sequence of positive real numbers \( \lambda_n > 0 \), and characteristic functions \( C_n(t) \), so that:

\[
C(t) = \lim_{n \to \infty} \exp \left[ \lambda_n (C_n(t) - 1) \right].
\] (7.23)

In other words, a characteristic function \( C(t) \) on \( \mathbb{R}^m \) is infinitely divisible if and only if \( C(t) \) is the limit of characteristic functions of compound Poisson random vectors.

**Proof.** If the characteristic function \( C(t) \) is given as the limit in 7.23, then since each \( \exp \left[ \lambda_n (C_n(t) - 1) \right] \) is infinitely divisible by lemma 7.31, item 3 of proposition 7.29 assures that \( C(t) \) is infinitely divisible.

Conversely assume that \( C(t) \) is infinitely divisible. Then \( C^{1/n}(t) \) is a characteristic function for all integers \( n > 0 \) by 4 of proposition 7.29, and thus \( g_n(t) \equiv \exp \left[ n (C^{1/n}(t) - 1) \right] \) is infinitely divisible by lemma 7.31. Left to prove is that for all \( t \):

\[
\lim_{n \to \infty} \exp \left[ n \left( C^{1/n}(t) - 1 \right) \right] = C(t).
\] ((*)

Once proved, 7.23 then holds with \( \lambda_n = n \) and \( C_n(t) = C^{1/n}(t) \).

To prove (*), since \( C(t) \neq 0 \) for all \( t \in \mathbb{R}^m \) by 1 of proposition 7.29, fix \( t \) and write \( C(t) = e^{a+bi} \), where to be well defined we choose \( b \in (-\pi, \pi] \).

Then using a Taylor series (remark 6.24, book 5):

\[
n \left( C^{1/n}(t) - 1 \right) = \sum_{k=1}^{\infty} \frac{(a + bi)^k}{k!n^{k-1}}
\]

\[
= a + bi + \sum_{k=2}^{\infty} \frac{(a + bi)^k}{k!n^{k-1}}
\]
The summation converges to 0 as \( n \to \infty \) because:

\[
\sum_{k=2}^{\infty} \frac{(a + bi)^k}{k! n^{k-1}} \leq \frac{1}{n} e^a,
\]

and thus by continuity (*) is proved.

Finally, the proof of lemma 7.31 obtains that each \( \exp \{ \lambda_n (C_n(t) - 1) \} \) in 7.23 is the characteristic function of a compound Poisson random vector, and thus the equivalence of the first and second statement of the proposition.

\[\blacksquare\]

Remark 7.33 1. The above proof demonstrates that any infinitely divisible distribution can be expressed as the weak limit of compound Poisson distributions with \( \lambda_n = n \) and \( C_n(t) = C^{1/n}(t) \), and thus \( C_n(t) \) is the characteristic function of the random vectors in the random summation in 7.20. It is worth a moment to better understand this construction.

Let \( X \) be the infinitely divisible (I. D.) random vector with joint distribution function \( F \) and characteristic function \( C_F(t) \). By definition of I.D, for every \( n \) there exists i.i.d. random vectors \( \{X^{(n)}_j\}_{j=1}^n \) so that as in 7.16:

\[
X = \sum_{j=1}^n X^{(n)}_j.
\]

The characteristic function for \( X^{(n)}_j \) is then \( C^{1/n}(t) \).

The de Finetti theorem states that we can construct a sequence of compound Poisson random vectors, \( \{X_n\}_{n=1}^\infty \) as follows. Define:

\[
X_n = \sum_{j=1}^{N_n} X^{(n)}_j,
\]

where \( N_n \) is Poisson with parameter \( \lambda_n = n \), and where \( X^{(n)}_j \) here has the same distribution as \( X^{(n)}_j \) in definition 7.22. It is then no longer the case that \( X_n = X \), since the compound Poisson summation replicates \( X \) only when \( N_n = n \). But now we have: \( X_n \Rightarrow X \).

2. The de Finetti theorem does not assert uniqueness in this representation of infinitely divisible \( C(t) \) with \( \lambda = n \) and \( C_n(t) = C^{1/n}(t) \). Looking at the second half of the proof, it is also the case that \( C^\alpha(t) \) is a characteristic function for all real \( \alpha > 0 \) by 4 of proposition 7.29.
CHAPTER 7 APPLICATIONS OF CHARACTERISTIC FUNCTIONS

and thus \( g_{\alpha,\beta}(t) \equiv \exp[\beta (C^\alpha(t) - 1)] \) is infinitely divisible by lemma 7.31 for \( \beta > 0 \). Now if \( C(t) = e^{a+bi} \) as in the above proof:

\[
\beta (C^\alpha(t) - 1) = \sum_{k=1}^{\infty} \frac{(a+bi)^k}{k!} \beta \alpha^k
\]

\[
= (a + bi) \beta \alpha + \sum_{k=2}^{\infty} \frac{(a+bi)^k}{k!} \beta \alpha^k.
\]

The conclusion that \( \exp[\beta_n (C^{\alpha_n}(t) - 1)] \to C(t) \) then only requires that \( \beta_n \alpha_n \to 1 \) and \( \alpha_n \to 0 \). Thus we have many such parameterizations, though all have the structure of that above.

3. A question: Why does the proof of this representation in 2 not apply to arbitrary characteristic functions \( C(t) \)? First given that \( \beta_n \alpha_n \to 1 \) and \( \alpha_n \to 0 \), the proof that \( \exp[\beta_n (z^{\alpha_n} - 1)] \to z \) for \( z \in \mathbb{C} \) requires only that \( z \neq 0 \). So when \( z = C(t) \), the characteristic function of an I.D. distribution, this is assured by 1 of proposition 7.29. But other characteristic functions which are not infinitely divisible can also have this property. For example the binomial distribution’s characteristic function has no real zeros, yet is not infinitely divisible.

The resolution is that in addition to \( C(t) \neq 0 \), this proof also requires that \( C^{\alpha_n}(t) \) be a characteristic function for some sequence \( \alpha_n \to 0 \). The implication of de Finetti’s theorem is that any such characteristic function must be infinitely divisible.

Corollary 7.34 (de Finetti’s theorem) A characteristic function \( C(t) \) is infinitely divisible if and only if \( C^{\alpha_n}(t) \) is a characteristic function for some sequence \( \alpha_n \to 0 \).

**Proof.** Infinitely divisible characteristic functions have this property by definition with \( \alpha_n = 1/n \). Conversely, if such a sequence exists, then the proof of 1 of proposition 7.29 applies, replacing \( 1/n \) there with \( \alpha_n \), concluding that \( C(t) \neq 0 \) for all \( t \). The construction in the second half of de Finetti’s proof then obtains that:

\[
\lim_{n \to \infty} \exp[\alpha_n^{-1} (C^{\alpha_n}(t) - 1)] = C(t),
\]

and thus \( C(t) \) is infinitely divisible by de Finetti’s theorem. ■

Remark 7.35 Each of the distributions in example 7.26 can be rewritten as the weak limit of compound Poisson variates with the same underlying distribution functions for the component variates derived there, and \( \lambda_n = n \).
It is worth a few moments to think through the details of such constructions, especially when $X$ is compound Poisson. What does the compound Poisson approximation look like, and why does it work?

It is natural to wonder if there is a way to characterize all of the characteristic functions that are possible as limits of the characteristic functions of compound Poisson variates as in 7.23. Put another way and without reference to de Finetti’s theorem, given an infinitely divisible distribution function $F$ on $\mathbb{R}^m$, is there a natural way to characterize the mathematical form of $C(t)$, the associated characteristic function?

As it turns out there have been several developments from papers in the late 1920s and 1930s by Bruno de Finetti (1906 – 1985), Andrey Kolmogorov (1903 – 1987), Paul Lévy (1886 – 1971) and Alexandre Khintchine (a.k.a. Aleksandr Khinchin; 1894 – 1959). A common version of these results is known as the Lévy-Khintchine Representation theorem which we state next, discussing only parts of the proof. Full details on this result can be found in Sato (1999).

**Proposition 7.36 (Lévy-Khintchine Representation theorem)** 1) Let $F$ be an infinitely divisible distribution function on $\mathbb{R}^m$ and $C(t)$ the associated characteristic function. Then with $t \in \mathbb{R}^m$:

$$
C(t) = \exp \left[ i \gamma \cdot t - \frac{1}{2} t^T A t + \int_{\mathbb{R}^m} \left( e^{i x \cdot t} - 1 - i (x \cdot t) \chi_D(x) \right) d\nu \right],
$$

(7.24)

where $\gamma \in \mathbb{R}^m$, $A$ is a symmetric, positive semi-definite $m \times m$ matrix, $D = \{ x \in \mathbb{R}^m | |x| \leq 1 \}$, and $\nu$ is a Lévy measure. A Lévy measure is a not-necessarily finite measure with:

$$
\nu(\{0\}) = 0, \text{ and } \int_{\mathbb{R}^m} \min(|x|^2, 1) d\nu < \infty.
$$

(7.25)

2) The representation in 7.24 by $\gamma$, $A$ and $\nu$ with the given properties is unique.

3) Given $\gamma$, $A$ and $\nu$ with the given properties, there exists an infinitely divisible distribution function $F$ on $\mathbb{R}^m$ with associated characteristic function $C(t)$ given as in 7.24.

**Remark 7.37** 1. **Well-definedness:** Given a Lévy measure $\nu$, the integral in 7.24 is well defined since:

$$
\int_{\mathbb{R}^m} \left( e^{ix \cdot t} - 1 - i (x \cdot t) \chi_D(x) \right) d\nu = \int_D \left( e^{ix \cdot t} - 1 - ix \cdot t \right) d\nu + \int_{\mathbb{R}^m - D} (e^{ix \cdot t} - 1) d\nu.
$$
The first integrand is bounded on \( D \) by 7.3 and the Cauchy-Schwarz inequality:
\[
|e^{ix\cdot t} - 1 - ix\cdot t| \leq \frac{1}{2} |x|^2 |t|^2 + O \left( |x|^3 |t|^3 \right),
\]
while for the second integrand, \(|e^{ix\cdot t} - 1| \leq 2\) on \( \mathbb{R}^m - D \). Thus each integral is finite by 7.25.

2. **Special Case:** If the Lévy measure \( \nu \) is in fact finite on \( D \), meaning \( \int_D |x| \, d\nu < 1 \); then the Lévy-Khintchine representation simplifies to:
\[
C(t) = \exp \left[ i\bar{\gamma} \cdot t - \frac{1}{2} t^T A t + \int_{\mathbb{R}^m} \left( e^{ix\cdot t} - 1 \right) \, d\nu \right],
\]
where \( \bar{\gamma} \) is defined componentwise:
\[
\bar{\gamma}_j = \gamma_j - \int_D x_j d\nu.
\]
Then by example 7.26 and proposition 6.20, the underlying distribution function \( F \) is the convolution of a multivariate normal distribution with parameters \( \bar{\gamma} \) and \( A \) as in 6.23, and a compound Poisson distribution with \( \lambda \mu_X = \nu \) as in 7.22, where \( \mu_X \) is the measure associated with \( F_X \).

For this representation and result it is enough that \( \int_D |x| \, d\nu < \infty \) to make \( \bar{\gamma} \) well-defined.

3. **Comment on Proof of 1) and de Finetti’s theorem:** The proof of part 1 is the most intricate, and begins with de Finetti’s theorem. Given infinitely divisible \( F \), de Finetti’s theorem assures the existence of compound Poisson distributions \( \{F_n\} \) with \( F_n \Rightarrow F \). By 7.22 of example 7.26, each such \( F_n \) has an associated characteristic function \( C_n(t) \) that has the structure of 7.24 with \( A_n = 0, \gamma_n = 0, \) and \( \nu_n = \lambda_n \mu_{X_n} \). Here \( \mu_{X_n} \) is the measure associated with \( F_{X_n} \), the distribution function of the \( X_n \)-variates in 7.20. Thus the essence to the proof of 1) is to prove that if each \( C_n \) has the structure of 7.24, then \( F_n \Rightarrow F \) if and only if \( F \) is infinitely divisible and \( C(t) \) has the structure of 7.24. Further, the parameters in the structure for \( C(t) \) are inherited from the associated structures of the \( C_n(t) \).

4. **Proof of 3):** The proof to part 3 is accessible with the tools already developed. Given \( \gamma, A \) and \( \nu \) with the given properties, define:
\[
C_n(t) = \exp \left[ i\gamma \cdot t - \frac{1}{2} t^T A t + \int_{|x| > 1/n} \left( e^{ix\cdot t} - 1 - i(\cdot x) \chi_D(x) \right) \, d\nu \right].
\]
7.4 DISTRIBUTION FAMILIES RELATED UNDER MULTIPLICATION

By 7.25 \( v \) is a finite measure on \( \{ |x| > 1/n \} \), and thus as in part 2, \( C_n(t) \) is the characteristic function of an infinitely divisible distribution function. Now \( C_n(t) \to C(t) \) in 7.24 for all \( t \) since \( \nu(\{ 0 \}) = 0 \), so by Lévy’s continuity theorem of corollary 6.31, such \( C(t) \) is a characteristic function if it is continuous at \( t = 0 \). For this investigation only the integral in the expression for \( C(t) \) need be considered. As noted in 1 above, this integrand is bounded by a \( \nu \)-integrable function, and since the integrand is continuous, continuity of the integral follows from Lebesgue’s dominated convergence theorem (proposition 2.43, book 5). Since \( C(t) \) is a characteristic function, it now follows from proposition 7.29 that as the limit of infinitely divisible characteristic functions, \( C(t) \) is infinitely divisible.

7.4 Distribution Families Related Under Multiplication

Interestingly, some distribution families are related under multiplication. The most common is the lognormal, but this property generalizes easily to the log-\( X \) distribution defined below.

1. Lognormal Distribution:

   The lognormal distribution is defined in 6.24 and has moments of all orders (see 3.68 of book 4). But this distribution was shown to not have a moment generating function in remark 3.30 of book 4, and as noted above, no closed form version of its characteristic function is currently known. But while sums of independent lognormal variates are difficult to navigate, this distribution family has the interesting property that it is closed under multiplication. Specifically, if \( L_i \) is lognormal with parameters \( \mu_i \) and \( \sigma_i^2 \), then by definition \( L_i = \exp(N_i) \) where \( N_i \) is normal with parameters \( \mu_i, \sigma_i^2 \). By proposition 3.56 of book 2, \( \{ L_i \}_{i=1}^m \) are independent lognormal if and only if \( \{ N_i \}_{i=1}^m \) are independent normal.

   Hence

   \[
   L = \prod_{i=1}^m L_i,
   \]

   has a lognormal distribution with parameters \( \mu = \sum_{i=1}^m \mu_i \) and \( \sigma^2 = \sum_{i=1}^m \sigma_i^2 \) since \( L = \exp[\sum_{i=1}^m N_i] \).

2. Log-\( X \) Distribution
If $X$ is a random variable defined on $(\mathcal{S}, \mathcal{E}, \lambda)$ with distribution function $F_X(x)$, then we can define the log-$X$ random variable $Y$ on $(0, \infty)$ by $Y \equiv \exp(X)$. Then for $y > 0$,

$$F_Y(y) \equiv \lambda \left[ Y^{-1}(-\infty, y] \right] = \lambda \left[ X^{-1}(-\infty, \ln y] \right],$$

and thus:

$$F_Y(y) = F_X(\ln y). \quad (7.27)$$

If $X$ also has a measurable density function $f_X(x)$, then $Y$ has a density function given by an application of proposition 3.34 of book 5:

$$f_Y(y) = f_X(\ln y)/y. \quad (7.28)$$

While such $f_Y$ need not be finite at $y = 0$, for any $\epsilon > 0$:

$$\int_{\epsilon}^{\infty} f_Y(y)dy = \int_{\ln \epsilon}^{\infty} f_X(x)dx.$$ 

Thus as $\epsilon \to 0$ it follows that $f_Y$ is integrable on $[0, \infty)$ and is well-defined as an improper integral. Analogous to the lognormal example, if the distribution family of $X$ is closed under addition, then the distribution function family of $Y$ is closed under multiplication. For example, the log-gamma distribution is closed under multiplication if all gamma variates are defined with the same parameter $\lambda$.

If $X$ has a discrete density function $f_X(x)$ defined on $\{x_i\}_{i=1}^{N}$ where $N$ can be finite or infinite, then $Y = \exp(X)$ had a discrete density function defined in the obvious way on $\{y_i\}_{i=1}^{N} \equiv \{\exp x_i\}_{i=1}^{N}$ by

$$f_Y(y_i) = f_X(\ln y_i). \quad (7.29)$$

More generally, if the distribution family of $X$ is related under addition to the distribution family of $X'$, then the distribution family of log-$X$ is related under multiplication to the distribution family of log-$X'$. For example, the exponential family is related under addition to the gamma family, all with common parameter $\lambda$, and so the log-exponential family is related under multiplication to the log-gamma family.
Chapter 8

Discrete Time Asset Models in Finance

In this chapter we investigate asset price models in discrete time which are envisioned to be two "dimensional" in the sense of having both a spatial and a temporal specification. The tools of the prior chapters will then be applied to identify models that have mathematically tractable characteristics. The following chapter will then study the pricing of financial derivatives on such assets, again in discrete time. Beyond these financial applications, the next few books will develop the necessary mathematics and probability theory to allow a continuous time extension of these results.

To set the stage for this discussion, recall the book 1 result as summarized in proposition 9.20:

**Proposition 9.20 (book 1):** Given probability spaces \( \{(\mathbb{R}, \sigma_i(\mathbb{R}), \mu_i)\}_{i=1}^{\infty} \), define the **infinite product probability space** \( \mathbb{R}^\mathbb{N} = \prod_{i=1}^{\infty} \mathbb{R}_i \) by:

\[
\mathbb{R}^\mathbb{N} = \{(x_1, x_2, ...) | x_i \in \mathbb{R}\}.
\]

Let \( A^+ \) denote the algebra of **general finite dimensional measurable rectangles** in \( \mathbb{R}^\mathbb{N} \), also called **general cylinder sets**. Such sets are defined for any positive integer \( n \) and \( n \)-tuple of positive integers \( J = (j(1), j(2), ..., j(n)) \) by:

\[
A = \{x \in \mathbb{R}^\mathbb{N}| (x_{j(1)}, x_{j(2)}, ..., x_{j(n)}) \in H\}, \quad (8.1)
\]
where $H \in \sigma \left( \prod_{i=1}^{n} \mathbb{R}_{j(i)} \right)$, the finite dimensional product space sigma algebra associated with $\prod_{i=1}^{n} \mathbb{R}_{j(i)}$. The cylinder set $A$ is said to be defined by $H$ and $J$.

Let $\mu_0$ be the the product set measure defined on $A \in A^+$ by:

$$\mu_0(A) = \mu_J(H),$$

(8.2)

where $A$ is defined by $H$ and $J$, and let $\mu_J$ denote the finite dimensional product space measure associated with $\prod_{i=1}^{n} X_{j(i)}$. Then $\mu_0$ can be extended to a complete probability measure $\mu_{\Pi}$ on a sigma algebra $\sigma(\mathbb{R}^N)$, where $A^+ \subset \sigma(\mathbb{R}^N)$ and $\mu_{\Pi}(A) = \mu_0(A)$ for all $A \in A^+$.

Thus $(\mathbb{R}^N, \sigma(\mathbb{R}^N), \mu_{\Pi})$ is a complete probability space, called the infinite product probability space, and $\mu_{\Pi}$ is called the product probability measure. Further, $\mu_{\Pi}$ is uniquely defined on the smallest sigma algebra that contains $A^+$.

In this chapter we use the terminology of "asset prices" to simplify the language, though the ideas discussed apply equally well to the modelling of other financial variables.

### 8.1 Spatial Models of Asset Prices

By a spatial model of asset prices is meant a collection of distributional assumptions of asset prices or other financial variables at various points in time, with or without a specification of an associated temporal model of how prices evolve over time to produce such spatial distributions. One possible application for finance of proposition 9.20 of book 1 can be derived by declaring that the index $i$ denotes time $i\Delta t$ for some arbitrarily defined time-step $\Delta t$. Thus $(\mathbb{R}, \sigma(\mathbb{R}), \mu_i)$ denotes the probability space of prices at time $i\Delta t$ of a given asset or class of assets: stocks, bonds, currencies, commodities, etc., or the probability space of a financial variable such as interest rates, etc. The conclusion of the above result then becomes: Given any collection of probability measures for these future variables, there is a complete probability space associated with the space of all prices over time. This space is defined by:

$$\mathbb{R}^N = \{(x_{\Delta t}, x_{2\Delta t}, ...)|x_i \in \mathbb{R}\},$$

a sigma algebra $\sigma(\mathbb{R}^N)$ which contains all the finite dimensional measurable rectangles, and a measure $\mu_{\Pi}$ such that the measure of such rectangles then agrees with the finite dimensional probabilities defined relative to these index sets.
Example 8.1 Define \( \lambda_i \) as the probability measure induced by the normal distribution with mean 0 and variance \( i \Delta t \). In other words, \( \lambda_i \) is induced by the distribution function

\[
F_i(x) = (2\pi i \Delta t)^{-1/2} \int_{-\infty}^{x} \exp \left( -y^2 / (2i \Delta t) \right) dy.
\]

Then \((\mathbb{R}^N, \sigma(\mathbb{R}^N), \lambda_N)\) is the probability space of all countable sequences of real numbers. Further, given \( n, J = (j(1), j(2), ..., j(n)) \), \( H = \prod_{i=1}^{n} (a_{j(i)}, b_{j(i)}) \), and \( A \equiv \{ x \in \mathbb{R}^N | (x_{j(1)}, x_{j(2)}, ..., x_{j(n)}) \in H \} \):

\[
\lambda_N[A] = \prod_{i=1}^{n} (F_{j(i)}(b_{j(i)}) - F_{j(i)}(a_{j(i)}))
\]

\[
= \prod_{i=1}^{n} \left( (2\pi i \Delta t)^{-1/2} \int_{a_{j(i)}}^{b_{j(i)}} \exp \left( -y^2 / (2j(i) \Delta t) \right) dy \right).
\]

With \( a_{j(i)} = -\infty \) for all \( i \), it follows that

\[
\lambda_N \left[ \prod_{i=1}^{n} (-\infty, b_{j(i)}) \right] = \prod_{i=1}^{n} F_{j(i)}(b_{j(i)}).
\]

Hence, if \( X_{j(i)} \) is a random variable with distribution function \( F_{j(i)}(x) \), the joint distribution function of the random vector \( (X_{j(1)}, X_{j(2)}, ..., X_{j(n)}) \) induced by \( \lambda_N \) is given by:

\[
F(x_{j(1)}, x_{j(2)}, ..., x_{j(n)}) = \prod_{i=1}^{n} F_{j(i)}(x_{j(i)}).
\]

Thus by proposition 3.53 of book 2, this construction implies that for any \( n \), \( \{X_{j(i)}\}_{i=1}^{n} \) are independent random variables.

One may well conclude that this independence property is in violation of common sense for most applications in finance. The point of this example is simply to illustrate an application of proposition 9.20, and to note a potentially unintended consequence of blindly applying such models.

The conclusion of the above example is true in the general case. Assume we are given probability measures \( \{\mu_i\}_{i=1}^{\infty} \) and associated distribution functions \( \{F_i\}_{i=1}^{\infty} \) defined by \( F_i(x) = \mu_i [(-\infty, x)] \). By the construction of Skorokhod’s Representation theorem of proposition 8.30 of book 2, there exists random variables \( \{X_i\}_{i=1}^{\infty} \) with probability distribution functions \( \{F_i\}_{i=1}^{\infty}, \) all defined on a common probability space which can be taken to be \((0, 1), (0, 1), m_L)\) with \( m_L \) Lebesgue measure. We then have that:

\[
(X_1, X_2, ...) : (0, 1) \rightarrow \mathbb{R}^N,
\]
and that under the probability measure \( \mu_N \) defined on \( \mathbb{R}^N \), this is a random vector of independent random variables as defined in definition 3.47 of book 2. Given any finite index subcollection \( J = (j(1), j(2), ..., j(n)) \), the joint distribution of \( (X_{j(1)}, X_{j(2)}, ..., X_{j(n)}) \) is defined by \( \mu_N \):

\[
F(x_{j(1)}, x_{j(2)}, ..., x_{j(n)}) = \mu_N \left[ \prod_{i=1}^{n} (-\infty, x_{j(i)}) \right],
\]

and we have that

\[
F(x_{j(1)}, x_{j(2)}, ..., x_{j(n)}) = \prod_{i=1}^{n} F_{j(i)}(x_{j(i)}).
\]

Applying proposition 3.53 of book 2, this assures independence of \( (X_{j(1)}, X_{j(2)}, ..., X_{j(n)}) \).

**Summary 8.2** The construction of proposition 9.20 of book 1 provides a **spatial model of prices** or values of given financial variable over time, given any collection of distributions at the various time points. But this construction always obtains the conclusion that these prices or values are modelled as independent random variables. Thus if we desire a spatial model that reflects the assumption that the asset prices at these times are not independent, then the above construction does not provide a probability space with an appropriate structure. This is a problem for financial applications, of course, because it is difficult to imagine a financial asset for which the value tomorrow, say, is independent of the value today.

However, if the random variable \( X_i \) denotes a return during the \( i \)th period, this assumption of independence is more reasonable in many contexts, and the above proposition can be used to construct \((\mathbb{R}^N, \sigma(\mathbb{R}^N), \mu_N)\) as a return space. We turn to these models next, and call them "temporal models of asset prices."

### 8.2 Temporal Models of Asset Prices

By a **temporal model of asset prices** is meant a collection of distributional assumptions of how asset prices change between various points in time. Such a model may be defined with or without specifying the **spatial models** of prices at various future times implied by such temporal distributions. We investigate two structures for these models. While what follows also applies to models of random vectors, we restrict our attention to models of random variables for notational simplicity.
8.2 TEMPORAL MODELS OF ASSET PRICES

8.2.1 Additive Temporal Models

Let $X_i$ denote the asset price at time $i\Delta t$ for some given time-step $\Delta t$, with $X_0$ given and $\Delta t$ typically defined in reference to a fixed interval $[0, T]$, commonly $\Delta t = T/n$. The additive temporal model specifies that:

$$X_n = X_{n-1} + Y_n,$$

where $\{Y_i\}_{i=1}^\infty$ are independent random variables with a given distribution function. Logically, this distribution function will depend on $t$.

In theory this model could be defined without independence, or reflect different distribution functions for the various time periods. But in addition to adding complexity, this generalization will complicate assumptions for interval subdivisions contemplated for "scalability" in a later section.

To simplify notation we will often notationally suppress $X_0$ by assuming that $X_0 = 0$. Then $X_n$ can be understood as the $n$-period change in the $X$-variate, and the asset price at time $n\Delta t$ is $X_0 + X_n$.

Example 8.3 Let $X_0 = 0$ and assume that each $Y_i$ has normal density function given in 3.1 with mean $\mu = 0$ and variance $\sigma^2\Delta t$. If independent, the joint density function of $(Y_1, Y_2, ..., Y_n)$ is given in 1.36 by $f_Y(y_1, y_2, ..., y_n) = \prod_{i=1}^n f_i(y_i)$, and so:

$$f_Y(y_1, y_2, ..., y_n) = \left(2\pi\sigma^2\Delta t\right)^{-n/2} \prod_{i=1}^n \exp \left(-y_i^2 / (2\sigma^2\Delta t)\right).$$

The joint density function of $(X_1, X_2, ..., X_n)$ can then be inferred from this by the change of variables formula in proposition 3.24 of book 5. To do so, define the transformation $T_n : (y_1, y_2, ..., y_n) \rightarrow (x_1, x_2, ..., x_n)$ by $x_i = \sum_{j=1}^i y_j$. As a sum of independent normals, $X_i$ is normal with mean 0 and variance $i\Delta t$ as verified by a moment generating function argument and then applying the uniqueness result from corollary 3.58 of book 4. Also $T_n$ is a linear transformation with associated nonsingular lower triangular matrix:

$$A_n = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}.$$
That is, \( T_n(y_1, y_2, \ldots, y_n) = A_n y \), where consistent with convention the \( n \)-vector \( y \) is identified as a column matrix.

It then follows from 3.17 of book 4 that with \( x \) and \( y \) denoting the respective \( n \)-vectors, the density function of \( X = (X_1, X_2, \ldots, X_n) \) is given:

\[
f_X(x) = f_Y(A_n^{-1}x) \left| \det (A_n^{-1}) \right|
\]

Now \( A_n^{-1} : (x_1, x_2, \ldots, x_n) \rightarrow (y_1, y_2, \ldots, y_n) \) is defined by \( y_1 = x_1 \), and \( y_k = x_k - x_{k-1} \) for \( k > 1 \). The matrix of \( A_n^{-1} \) is then:

\[
A_n^{-1} = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 & 0 \\
-1 & 1 & 0 & \cdots & 0 & 0 \\
0 & -1 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -1 & 1
\end{pmatrix},
\]

and thus as a lower triangular matrix with 1s along the diagonal, \( A_n^{-1} \) has determinant equal to 1. Hence,

\[
f_X(x_1, x_2, \ldots, x_n) = f_1(x_1) \prod_{i=2}^{n} f_i(x_i - x_{i-1})
= (2\pi \sigma^2 \Delta t)^{-n/2} \exp \left(-\frac{x_1^2}{2(2\sigma^2 \Delta t)}\right) \prod_{i=2}^{n} \exp \left(-\frac{(x_i - x_{i-1})^2}{2(2\sigma^2 \Delta t)}\right).
\]

As anticipated, we can see from this expression that the \( \{X_i\} \) are not independent random variables.

Comparing with 3.3 of proposition 3.1, \( (X_1, X_2, \ldots, X_n) \) so defined has a multivariate normal distribution with covariance matrix \( C = A_n A_n^T \sigma^2 \Delta t \), where this notation implies that all components of \( A_n A_n^T \) are to be multiplied by \( \sigma^2 \Delta t \). This extra factor of \( \sigma^2 \Delta t \) adjusts for \( Y_i \) having variance \( \sigma^2 \Delta t \) in contrast to a variance of 1 in proposition 3.1. Hence \( \det C = (\sigma^2 \Delta t)^n \), and with \( C^{-1} = (A_n^T)^{-1} A_n^{-1} \sigma^2 \Delta t \) this density function can be expressed:

\[
f_X(x_1, x_2, \ldots, x_n) = (2\pi \sigma^2 \Delta t)^{-n/2} \exp \left[-\frac{1}{2} x^T C^{-1} x\right].
\]

As noted above, this density function is not a product of densities in the \( X_i \) separately, and hence the spatial values are not independent.
Remark 8.4 More generally, if independent \((Y_1, Y_2, \ldots, Y_n)\) have a distribution function \(F_Y(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} F_i(y_i)\) and associated Borel measure \(\mu_Y = \prod_{i=1}^{n} \mu_i\), then \(\mu_X\) is well-defined by proposition 3.10 of book 5. Defining the transformation \(T_n\) as above, it follows that \(\mu_X = (\mu_Y)_{T_n}\), the measure induced by \(T_n\). In other words, for \(A \in B(\mathbb{R}^{n})\), it follows from 4.5 that:

\[
\mu_X(A) \equiv \mu_Y(T_n^{-1}A).
\]

Letting \(A = \prod_{i=1}^{n} (-\infty, x_i] \):

\[
F_X(x_1, x_2, ..., x_n) \equiv \mu_Y(T_n^{-1} [\prod_{i=1}^{n} (-\infty, x_i)]) .
\]

Summary 8.5 Based on the construction of proposition 9.20 of book 2, given distribution functions \(F_i\) and associated Borel measures \(\mu_i\) for independent \(Y_i\), the \(Y\)-product space \((\mathbb{R}^{n}, \sigma(\mathbb{R}^{n}), \mu_{\mathbb{R}^{n}})\) provides a rigorous structure for probability statements on asset prices. Statements in \(X\) and \(Y\) can in general be related by 4.5, and when density functions exists, the change of variables formula from 3.17 of book 4 applies. Asset prices in the spatial model are in general not independent.

8.2.2 Multiplicative Temporal Models

In many price model applications, a temporal model can be expressed as a multiplicative temporal model whereby given \(X_0\):

\[
X_n = X_{n-1} \exp Z_n,
\]

where \(\{Z_i\}_{i=1}^{\infty}\) are independent random variables with a given distribution function. As noted above, the period distribution function will in general be fixed and dependent on the model time-step \(\Delta t\) defined in reference to an interval \([0, T]\), commonly \(\Delta t = T/n\). In this formulation it is convenient to denote the nonnegative multiplicative factor as \(\exp Z_n\) to reflect the continuously compounded period return random variable

\[
Z_n \equiv \ln [X_n/X_{n-1}].
\]

To simplify notation, we often suppress \(X_0\) by assuming \(X_0 = 1\) and hence \(X_n\) can be understood as the \(n\)-period multiplicative change in the \(X\)-variate. Then the value of asset price at time \(n\Delta t\) is \(X_0X_n\).

Example 8.6 With \(X_0 = 100\) define \(\{Z_i\}_{i=1}^{\infty}\) with binomial returns given by:

\[
Z_n = \mu \Delta t + \sigma \sqrt{\Delta t} b_n,
\]
where \( \{b_n\}_{n=1}^{\infty} \) are independent and binomially distributed to equal \( \pm 1 \) with probability \( 1/2 \). Hence the first two return moments are:

\[
E[Z_n] = \mu \Delta t, \quad \text{Var}[Z_n] = \sigma^2 \Delta t.
\]

Alternatively, we can define normally distributed \( Z_n \) with returns given by:

\[
Z_n = \mu \Delta t + \sigma \sqrt{\Delta t} N_n
\]

where \( \{N_n\}_{n=1}^{\infty} \) are independent with a standard normal distribution. The first two moments of this return model are then the same as in the above binomial model.

With binomial \( Z_n \) defined above it follows that:

\[
X_n = X_0 \exp \left( \mu n \Delta t + \sigma \sqrt{\Delta t} \sum_{j=1}^{n} b_j \right), \tag{8.3}
\]

and hence each \( X_n \) is log-binomial. Thus for \( m < n \):

\[
X_n = X_m \exp \left( \mu (n-m) \Delta t + \sigma \sqrt{\Delta t} \sum_{j=m+1}^{n} b_j \right),
\]

and \( X_n \) and \( X_m \) are apparently not independent.

To formalize this, assume \( X_m \) is defined with \( X_0 \) and given binomial variates \( \sum_{j=1}^{m} b'_j \). Then given \( X_m \),

\[
X_n = X_0 \exp \left( \mu n \Delta t + \sigma \sqrt{\Delta t} \left[ \sum_{j=1}^{m} b'_j + \sum_{j=m+1}^{n} b_j \right] \right).
\]

The joint density function \( f(X_m, X_n) \) equals 0 for any \( X_n \) with \( \sum_{j=1}^{m} b_j > \sum_{j=1}^{n} b_j + (n-m) \) or \( \sum_{j=1}^{n} b_j < \sum_{j=1}^{m} b'_j - (n-m) \), and since \( -m \leq \sum_{j=1}^{m} b'_j \leq m \) is fixed, there will always be some realization of \( X_n \) with this property. Hence we cannot have \( f(X_m, X_n) = f(X_m)f(X_n) \), and so \( X_n \) and \( X_m \) are not independent. Alternatively in terms of conditional density functions, \( f(X_n|X_m) \) is log-binomial with:

\[
E[\ln(X_n|X_m)] = \ln X_0 + \mu n \Delta t + \sigma \sqrt{\Delta t} \sum_{j=1}^{m} b'_j,
\]

\[
\text{Var}[\ln(X_n|X_m)] = \sigma^2 \Delta t (n-m),
\]

while \( f(X_n) \) is log-binomial with respective moments \( \ln X_0 + \mu n \Delta t \) and \( \sigma^2 \Delta t n \). Hence \( f(X_n|X_m) \neq f(X_n) \).

With normal \( Z_n \) above, each \( X_n \) is lognormal:

\[
X_n = X_0 \exp \left( \mu n \Delta t + \sigma \sqrt{\Delta t} \sum_{j=1}^{n} N_j \right), \tag{8.4}
\]
and it follows that for \( m < n \):

\[
X_n = X_m \exp \left( \mu(n-m)\Delta t + \sigma\sqrt{\Delta t} \sum_{j=m+1}^{n} N_j \right).
\]

As above it can then be seen that \( f(X_n|X_m) \neq f(X_n) \).

**Remark 8.7** In either model the nonlinear transformation between returns and prices, \( g_n : (z_1, z_2, ..., z_n) \rightarrow (x_1, x_2, ..., x_n) \), is given by

\[
x_i = x_0 \exp \left[ \sum_{j=1}^{i} z_j \right].
\]

Thus for \( A \in \mathcal{B} (\mathbb{R}^n) \), it follows from 4.5 that:

\[
\mu_X(A) = \mu_Z(g_n^{-1}A).
\]

Letting \( A = \prod_{i=1}^{n} (\infty, x_i) \):

\[
F_X(x_1, x_2, ..., x_n) = \mu_Z \left( g_n^{-1} \left[ \prod_{i=1}^{n} (\infty, x_i) \right] \right).
\]

**Summary 8.8** Based on the construction of proposition 9.20 of book 2, given distribution functions \( F_i \) and associated Borel measures \( \mu_i \) for each \( Z_i \), the return product space \( (\mathbb{R}^N, \sigma(\mathbb{R}^N), \mu) \) provides a rigorous and applicable structure for probability statements on asset prices. The associated asset prices in the spatial model are again not independent.

### 8.3 Induced Models

We have already introduced this notion generally above, identifying the induced spatial model for given temporal models, so here we document more detailed results. The focus will be on temporal and spatial random variables with density functions, since then more explicit representations of the induced models are possible. The more general results between spatial distribution functions and temporal probability measures are seen above in remarks 8.4 and 8.7.

#### 8.3.1 Spatial and Additive Temporal Models

Assume we are given spatial asset prices \( X = (X_1, X_2, ..., X_n) \) associated with independent, identically distributed additive temporal variates \( Y = (Y_1, Y_2, ..., Y_n) \):

\[
X_j = \sum_{i=1}^{j} Y_i.
\]  

(8.5)

As always, the temporal variates have associated with them an implied period time-step of \( \Delta t \), and thus so too do the spatial variates. We use the
convention noted above that \( X_0 \) is suppressed and hence \( X_n \) equals the \( n \)-period change in this variate, and the actual value of this asset at time \( n \Delta t \) is \( X_0 + X_n \).

Let \( A_n \) be the matrix defined as in example 8.3 so that \( A_n Y = X \). We can then apply proposition 3.24 of book 5 to obtain the following.

1. Given \( f_Y \) : The general relationship is \( f_X(x) = f_Y(A_n^{-1} x) \mid \det(A_n^{-1}) \), but since \( \det(A_n^{-1}) = 1 \):

\[
f_X(x) = f_Y(A_n^{-1} x).
\]

Hence if \( \{Y_i\}_{i=1}^n \) are independent:

\[
f_X(x_1, x_2, ..., x_n) = f_Y(x_1) \prod_{j=2}^{n} f_Y(x_j - x_{j-1}),
\]

and thus in general \( \{X_i\}_{i=1}^n \) will not be independent. An example of this result is seen in example 8.3 with normal \( \{Y_i\}_{i=1}^n \).

Joint densities of subsets of variables, \( \{X_{j(i)}\}_{i=1}^m \subset \{X_i\}_{i=1}^n \) can then be calculated as the marginal density functions of \( f_X(x_1, x_2, ..., x_n) \) (see section 1.2).

2. Given \( f_X \) : The general relationship is \( f_Y(y) = f_X(A_n y) \mid \det(A_n) \), but since \( \det(A_n) = 1 \):

\[
f_Y(y) = f_X(A_n y).
\]

Thus:

\[
f_Y(y_1, y_2, ..., y_n) = f_X \left( y_1, \sum_{i=1}^{2} y_i, ..., \sum_{i=1}^{n} y_i \right).
\]

Joint densities of subsets of variables, \( \{Y_{j(i)}\}_{i=1}^m \subset \{Y_i\}_{i=1}^n \) can then be calculated as marginal density functions of \( f_Y(y_1, y_2, ..., y_n) \).

**Example 8.9** Let \( f(x) \) denote the joint density function of \( X = (X_1, X_2, ..., X_n) \), assumed to be multivariate normally distributed as in 3.3 with mean vector \( \mu \) and covariance matrix \( C \):

\[
f_X(x) = (2\pi)^{-n/2} \det C]^{-1/2} \exp \left[ -\frac{1}{2} (x - \mu)^T C^{-1} (x - \mu) \right].
\]
8.3 INDUCED MODELS

Since $Y = A_n^{-1}X$, proposition 3.9 obtains that $Y$ is multivariate normally distributed with $\bar{\mu} = A_n^{-1}\mu$ and $\bar{C} = A_n^{-1}C(A_n^{-1})^T$:

$$f_Y(y) = (2\pi)^{-n/2} \left|\det \bar{C}\right|^{-1/2} \exp\left[-\frac{1}{2}(y - \bar{\mu})^T\bar{C}^{-1}(y - \bar{\mu})\right].$$

Since $\det(A_n^{-1}) = 1$ it follows that $\det \bar{C} = \det C$. Also, $\bar{C}^{-1} = A_n^TC^{-1}A_n$ and thus:

$$(y - \bar{\mu})^T\bar{C}^{-1}(y - \bar{\mu}) = [A_n(y - \bar{\mu})]^T C^{-1} [A_n(y - \bar{\mu})]$$

$$= [A_ny - \mu]^T C^{-1} [A_ny - \mu].$$

With $A_n$ given above, it is verified that $f_Y(y)$ from this expression agrees with 8.7.

8.3.2 Spatial and Multiplicative Temporal Models

Assume we are given spatial asset prices $X = (X_1, X_2, ..., X_n)$ associated with independent, identically distributed multiplicative temporal variates $Z = (Z_1, Z_2, ..., Z_n)$:

$$X_j = \prod_{i=1}^{j} \exp Z_i. \quad (8.8)$$

As above, these temporal variates have associated with them an implied period time-step of $\Delta t$. This model can also be defined by:

$$X_j = \exp W_j, \quad W_j = \sum_{i=1}^{j} Z_i. \quad (8.9)$$

It is again notationally convenient to suppress $X_0$ and hence $X_n$ equals the $n$-period multiplicative change in the $X$-variate, while the value of asset at time $n\Delta t$ is $X_0X_n$. Since all $X_j > 0$ it is apparent that all asset prices must have the same sign, and in particular the sign of $X_0$.

As above define the non-linear, differentiable transformation $g_n : (z_1, z_2, ..., z_n) \rightarrow (x_1, x_2, ..., x_n)$ by:

$$g_n(z_1, z_2, ..., z_n) = \left(\exp z_1, \exp \left(\sum_{i=1}^{2} z_i\right), ..., \exp \left(\sum_{i=1}^{n} z_i\right)\right).$$

Applying proposition 3.34 of book 5 to derive the relationship between the density functions of $x$ and $z$, we must determine the associated Jacobian matrix and determinant. Since $\partial x_j/\partial z_k = x_j$ for $k \leq j$ and $\partial x_j/\partial z_k = 0$
otherwise, the Jacobian matrix of \( g_n \) is lower triangular, and since \( x_j > 0 \) by assumption:

\[
\left| \det \left( \frac{\partial g_n(z)}{\partial z} \right) \right| = \prod_{j=1}^{n} x_j
\]

\[
= \prod_{j=1}^{n} \exp [(n - j + 1) z_j].
\]

Because \( g_n \) is invertible, the Jacobian matrices satisfy

\[
\left( \frac{\partial g_n^{-1}(x)}{\partial x} \right) = \left( \frac{\partial g_n(z)}{\partial z} \right)|_{g_n^{-1}(x)}^{-1},
\]

and thus:

\[
\left| \det \left( \frac{\partial g_n^{-1}(x)}{\partial x} \right) \right| = \left( \prod_{j=1}^{n} x_j \right)^{-1}.
\]

This determinant can also be calculated explicitly by noting that \( g_n^{-1} : (x_1, x_2, ..., x_n) \rightarrow (z_1, z_2, ..., z_n) \) is defined by \( z_j = \ln [x_j/x_{j-1}] \) with \( x_0 \equiv 1 \).

1. Given \( f_Z \):

\[
f_X(x) = f_Z \left( g_n^{-1}(x) \right) \left| \det \left( \frac{\partial g_n^{-1}(x)}{\partial x} \right) \right| = f_Z \left( \ln x_1, \ln [x_2/x_1], ..., \ln [x_n/x_{n-1}] \right) \left( \prod_{j=1}^{n} x_j \right)^{-1}.
\]

If \( \{Z_i\}_{i=1}^{n} \) are independent:

\[
f_X(x_1, x_2, ..., x_n) = \prod_{j=1}^{n} f_{Z_j} \left( \ln [x_j/x_{j-1}] \right) / x_j. \quad (8.10)
\]

2. Given \( f_X \):

\[
f_Z(z) = f_X(g_n(z)) \left| \det \left( \frac{\partial g_n(z)}{\partial z} \right) \right|,
\]

and substitution then provides:

\[
f_Z(z) = f_X \left( \exp z_1, \prod_{j=1}^{2} \exp z_j, ..., \prod_{j=1}^{n} \exp z_j \right) \prod_{j=1}^{n} \left[ \exp (n - j + 1) z_j \right]. \quad (8.11)
\]

**Example 8.10** Recall example 8.6.
8.3 INDUCED MODELS

1. Given normal returns, \( Z_j = \mu \Delta t + \sigma \sqrt{\Delta t} N_j \), with \( N_j \) independent standard normals with mean 0 and variance 1, we have from 8.8:

\[
x_j = \prod_{i=1}^{j} \exp \left( \mu \Delta t + \sigma \sqrt{\Delta t} N_i \right) = \exp \left[ \mu j \Delta t + \sigma \sqrt{\Delta t} \sum_{i=1}^{j} N_i \right].
\]

For such \( n \)-tuples \((x_1, x_2, ..., x_n)\), 8.10 obtains:

\[
f_X(x_1, x_2, ..., x_n) = \prod_{j=1}^{n} \frac{\exp \left[ -\left( \ln x_j - \ln x_{j-1} \right)^2/2 \right]}{x_j \sqrt{2\pi}}
= (2\pi)^{-n/2} \left( \prod_{j=1}^{n} x_j \right)^{-1} \exp \left[ -\sum_{j=1}^{n} \left( \ln x_j - \ln x_{j-1} \right)^2/2 \right].
\]

For the marginal density of \( X_n \), it is easier to return to the definition than to attempt to evaluate the associated marginal density from \( f_X(x_1, x_2, ..., x_n) \). Since \( X_n = \exp \left( \sum_{i=1}^{n} Z_i \right) \) and \( \sum_{i=1}^{n} Z_i \) is normally distributed with parameters \( n \mu \Delta t \) and \( \sigma^2 n \Delta t \), we have by definition that \( X_n \) is lognormally distributed with parameters \( n \mu \Delta t = \mu T \) and \( \sigma^2 n \Delta t = \sigma^2 T \), since typically \( \Delta t = T/n \).

A similar analysis holds for the conditional density of \( X_n \) given \( X_m \), \( m < n \), which is seen to be lognormally distributed with parameters \( \ln X_m + \mu (n - m) \Delta t \) and \( \sigma^2 (n - m) \Delta t \).

2. Given binomial returns, \( Z_j = \mu \Delta t + \sigma \sqrt{\Delta t} b_j \), with \( b_j \) independent and binomially distributed to equal 1 or -1 with probabilities \( p \) and \( 1 - p \) respectively, note that 8.10 does not apply to discrete density functions. However, based on first principles we see that \( \ln \left( x_j/x_{j-1} \right) \) must equal \( \mu \Delta t \pm \sigma \sqrt{\Delta t} \) for all \( j \), and hence \( f_X(x_1, x_2, ..., x_n) = 0 \) unless there is an \( n \)-tuple \((b_1, b_2, ..., b_n)\) of binomial variates so that for each \( j \leq n \):

\[
x_j = \prod_{i=1}^{j} \exp \left( \mu \Delta t + \sigma \sqrt{\Delta t} b_i \right) = \exp \left[ \mu j \Delta t + \sigma \sqrt{\Delta t} \sum_{i=1}^{j} b_i \right].
\]

If \((x_1, x_2, ..., x_n)\) is such an \( n \)-tuple, it follows that:

\[
f_X(x_1, x_2, ..., x_n) = \prod_{j=1}^{n} f_Z_j \left( \ln \left( x_j/x_{j-1} \right) \right),
\]

and thus \( f_X(x_1, x_2, ..., x_n) = p^k(1-p)^{n-k} \) where \( k \) denotes the number of indexes in \((b_1, b_2, ..., b_n)\) for which \( b_i = 1 \).
For any \( k \) there will be \( \binom{n}{k} \) such \( n \)-tuples \((x_1, x_2, \ldots, x_n)\) with \( f_X(x_1, x_2, \ldots, x_n) = p^k(1-p)^{n-k} \), all of which result in the same value of \( x_n \). For example, the marginal density of \( X_n \) is log-binomial with parameters \( n, p \):

\[
f_n(X_n) = f_n\left(\exp\left(\mu n\Delta t + \sigma \sqrt{\Delta t} \sum_{i=1}^{n} b_i\right)\right) = \binom{n}{k} p^k(1-p)^{n-k},
\]

where \( k \) denotes the number of indexes for which \( b_i = 1 \).

Comparing to the lognormal model for \( X_n \) above, note that in this log-binomial case, \( E[\ln X_n] = \mu n\Delta t + n(2p-1)\sigma \sqrt{n\Delta t} \) and \( Var[\ln X_n] = 4p(1-p)n\sigma^2 \Delta t \) and so these moments match the lognormal case when \( p = 1/2 \).

### 8.3.3 Simulating Asset Price Paths

In either temporal model it is easy to simulate asset price paths \((X_1, X_2, \ldots, X_n)\) in the case of independent \( Y \) increments for the additive model, or independent \( Z \) increments for the multiplicative model. Recalling chapter 4 of book 4, in either case it is often sufficient to only generate independent \( n \)-tuples of continuous, uniformly distributed variates defined on \([0, 1]\) and applying proposition 4.9 of book 2.

**Note:** In the book 4 examples these uniformly distributed variates were denoted \( \{Y_j\}_{j=1}^{n} \). To avoid notational confusion, we denote this uniformly distributed variate collection by \( \{D_j\}_{j=1}^{n} \).

In the most common case where the \( Y_i \) variates of the additive temporal model are independent and identically distributed with distribution function \( F_Y \), it follows from proposition 4.9 of book 2 that

\[
\{Y_j\}_{j=1}^{n} \equiv \{F_Y^\ast(D_j)\}_{j=1}^{n},
\]

defines a collection of independent variates with the given distribution function. For this statement, recall that \( F_Y^\ast \) denotes the left continuous inverse of \( F_Y \) as defined in definition 3.12 of book 2. For independent and identically distributed \( Z \) variates:

\[
\{Z_j\}_{j=1}^{n} \equiv \{F_Z^\ast(D_j)\}_{j=1}^{n},
\]

similarly defines an independent sample of returns in the multiplicative model.

In either case, each such sample provides a sample "path" \((X_1, X_2, \ldots, X_n)\) of asset variates given \( X_0 \) in the respective models.
8.4 Harmonious Asset Models

Though not standard terminology, below we define a model of asset prices or other financial variables to be **harmonious** when it possesses useful properties which we now discuss. The first property is necessarily subjective, reflecting the notion of "identifiability" of the spatial model, but is nonetheless useful in practice. The second notion relates to sums of the temporal variates $Y_i$ and $Z_i$ above, and we now refine this notation to explicitly link the temporal variates to the time-step $\Delta t \equiv T/n$, where $[0, T]$ denotes the time interval over which the model is applied.

Given time-step $\Delta t \equiv T/n$, we denote the independent, identically distributed temporal variates of either model by $\{V_i^{(n)}\}_{i=1}^n$, and define for $j = 1, \ldots, n$:

$$U_j^{(n)} = \sum_{i=1}^j V_i^{(n)}.$$  

Thus in the original notation of additive model $V_i^{(n)} = Y_i$ and $U_j^{(n)} = X_j$, while in the multiplicative model $V_i^{(n)} = Z_i$, and $U_j^{(n)} = W_j$. This notation is then consistent with the notation for $\Delta t$-refinements, since what will be denoted $Y_i^{(nm)}$ etc. will be associated with $\Delta t' \equiv T/nm$.

**Definition 8.11** A model of asset prices or other financial variables over the time period $[0, T]$ is said to be **harmonious** if:

1. The distributional assumptions for the (additive or multiplicative) temporal variables associated with a given time-step $\Delta t \equiv T/n$, plus independence, produce identifiable spatial distributions of asset prices at times $\{j\Delta t\}$ for $j = 1, \ldots, n$.

2. The distributional assumptions made for the (additive or multiplicative) temporal variables are scalable in the time step $\Delta t$.

A temporal model is said to be **perfectly** or **exactly** scalable if given independent $\{V_i^{(n)}\}_{i=1}^n$ associated with $\Delta t \equiv T/n$, and any $m$, there are independent, identically distributed $\{V_k^{(mn)}\}_{k=1}^{mn}$ so that for $1 \leq j \leq n$:

$$\sum_{i=1}^{mj} V_i^{(mn)} = d \sum_{i=1}^j V_i^{(n)}, \quad (8.12)$$

where $= d$ means equal in distribution. Equivalently, defining $U_{mj}^{(mn)} \equiv \sum_{i=1}^{mj} V_i^{(mn)}$, 8.12 can be stated:

$$U_{mj}^{(mn)} = d U_j^{(n)}$$
A temporal model is said to be **approximately scalable to order** $k_0$ if for $1 \leq j \leq n$:

$$E\left[\left(\sum_{i=1}^{mj} V_{ij}^{(mn)}\right)^k\right] = E\left[\left(\sum_{i=1}^{j} V_{i}^{(n)}\right)^k\right], \quad (8.13)$$

for $k \leq k_0$. Equivalently, with $U_{mj}^{(mn)}$ defined above:

$$E\left[U_{mj}^{(mn)}\right]^k = E\left[U_{j}^{(n)}\right]^k$$

for $k \leq k_0$.

**Remark 8.12** Note that while 8.13 is stated in terms of the moments $E\left[X^k\right]$, we can by proposition 3.20 of book 4 equivalently state this requirement in terms of the central moments, $E\left[(X - \mu)^k\right]$, when convenient.

Of course "identifiability" of the spatial distribution is subjective and reflects the scope of the portfolio of distributions in the user's tool kit. That said, the first condition allows the modeler to evaluate the reasonableness of "output" spatial distributions given the "input" temporal distributional assumptions and the application in hand. As a general rule, the temporal distributions are inputs to the model because empirical data in finance is almost always of the temporal type, and hence provides the greatest insights into this dimension. The spatial distribution can then be seen as simply an output of the temporal assumptions and not an independent distributional assumption. If the model provides unrealistic spatial results, this may then provide a cautionary note on reasonableness of the temporal assumptions.

Identification of the spatial distribution is also essential for the rigorous verification of the scalability criterion. In the case where the spatial distribution can only be inferred empirically, the scalability criterion can then only be approximately evaluated. The value of a scalable model as required in the second condition is that it allows an evaluation of the properties and implications of the model as $\Delta t \to 0$. This then provides a linkage between the more tractability discrete model, and the continuous time "models" of finance, the subject of later books.

### 8.4.1 Identifiability of the Spatial Distribution

In either temporal model, identifiability of the spatial distributions is ultimately a question of the identifiability of the distribution of
8.4 HARMONIOUS ASSET MODELS

\[ U_j^{(n)} = \sum_{i=1}^{j} V_i^{(n)}, \]
where \( \{V_i^{(n)}\}_{i=1}^{n} \) denote the temporal variates in the given model associated with \( \Delta t = T/n \). Given \( \{V_i^{(n)}\}_{i=1}^{n} \) in the additive temporal model, since \( X_j \equiv U_j^{(n)} \), it is apparent that if the distribution function of \( U_j^{(n)} \) is identifiable as \( F_{U_j^{(n)}} \), then this is the distribution function of \( X_j \), perhaps with a location change for \( X_0 \). In the multiplicative model, if the distribution function of \( U_j^{(n)} \) is identifiable as \( F_{U_j^{(n)}} \), then since \( X_j \equiv \exp U_j^{(n)} \), the distribution function of \( X_j \) is recognizable as a log-\( U_j^{(n)} \) distribution as defined in 7.28 and 7.29, again perhaps with a scale change for \( X_0 \).

So assume that for \( j = 1, \ldots, n \), that \( U_j^{(n)} \) is defined as above with temporal \( \{V_i^{(n)}\}_{i=1}^{n} \) that are independent and identically distributed. Dropping the superscripts for convenience, the distribution functions of \( U_j \) and \( V_i \) are related by iteration of 2.14:

\[ F_{U_j} = \ast_{i=1}^{j} F_{V_i}. \]

This simplified notation is intended to represent the \( j \)-fold convolution:

\[ \ast_{i=1}^{j} F_{V_i} = F_{V_1} \ast F_{V_2} \ast \cdots \ast F_{V_j}. \]

When the \( f_V \) density function exists we have that by 2.13:

\[ f_{U_j} = \ast_{i=1}^{j} f_{V_i}. \]

While true in theory, this result will rarely be useful in identifying \( f_{U_j} \) because of the intractability of convolution calculations except in rarest cases, or for relatively small \( j \).

More useful in practice is that by independence and 6.32:

\[ C_{U_j}(t) = \prod_{i=1}^{j} C_{V_i}(t), \]

and when moment generating functions exist, by 3.35 of book 4:

\[ M_{U_j}(t) = \prod_{i=1}^{j} M_{V_i}(t). \]

As seen below, there are many examples of models that satisfy criterion 1 of the above definition as verified using characteristic or moment generating function methods.
8.4.2 Scalability of the Temporal Distribution

For the second criterion of scalability, we first note a simple characterization of perfect scalability that may well have been anticipated.

**Proposition 8.13** A temporal model is perfectly scalable if and only if the period $V^{(n)}$ random variable is **infinitely divisible** (recall definition 7.22). Equivalently, a temporal model is perfectly scalable if and only if the $U^{(n)}$ random variable is infinitely divisible.

**Proof.** If $V^{(n)}_i$ is infinitely divisible, then for all $m$ there exists independent and identically distributed variates which we index as $\{V^{(mn)}_{k(i-1)m+1}\}_{k=1}^{m}$ so that:

$$\sum_{k=(i-1)m+1}^{im} V^{(mn)}_k = d V^{(n)}_i.$$ (\textit{(*)})

Further, these collections can be defined with independent $\{V^{(mn)}_{k(i-1)m+1}\}_{k=1}^{m}$ for each $j$:

$$\sum_{i=1}^{mj} V^{(mn)}_i = \sum_{i=1}^{j} \left( \sum_{k=(i-1)m+1}^{im} V^{(mn)}_k \right),$$

an independent $j$-sum since $\{V^{(n)}_i\}_{i=1}^n$ are independent by assumption. The distribution function of this $j$-sum is given by the convolution of underlying distribution functions. By (\textit{*}), the resulting convolution is also the distribution function of $\sum_{i=1}^{j} V^{(n)}_i$ and thus 8.12 is satisfied. Conversely, 8.12 implies that for all $m$ there exists $\{V^{(mn)}_{k(i-1)m+1}\}_{k=1}^{m}$ so that $\sum_{i=1}^{m} V^{(mn)}_k = d V^{(n)}_1$, and thus $V^{(n)}_1$ is infinitely divisible. But then all $V^{(n)}_i$-variates are infinitely divisible since identically distributed. For the statement on $U^{(n)}_i$, infinite divisibility of the period $V^{(n)}_i$ random variables implies infinite divisibility of $U^{(n)}_j$ for all $j$ by 2 of proposition 7.29.

Conversely if $U^{(n)}_n = \sum_{i=1}^{n} V^{(n)}_i$ is infinitely divisible, then for any $m$ there exists i.i.d. $\{V^{(n)}_k\}_{k=1}^{m}$ so that $U^{(n)}_n = \sum_{k=1}^{m} V^{(n)}_k$. Letting $m = n$ and evaluating the characteristic function of both representations of $U^{(n)}_i$ produces $[C_{V}(t)]^n = [C_{V}(t)]^n$ and the result that $V^{(n)}_k = d V^{(n)}_i$ follows by uniqueness of characteristic functions (proposition 6.25). Now letting $m = nl$ similarly obtains that $[C_{V}(t)]^{nl} = [C_{\Sigma V}(t)]^{nl}$ where $\Sigma V$ is shorthand for the sum of the $V^{(n)}_k$-variates in groups of 1. Thus each $V^{(n)}_i = d \sum_{k=(i-1)l+1}^{il} V^{(n)}_k$ and $V^{(n)}_i$ is infinitely divisible for all $i$. ■
The next result proves that scalability on either basis is preserved under an arbitrary affine transformation of a given set of temporal variates \( \{V_j^{(n)}\}_{j=1}^n \).

**Proposition 8.14** Given arbitrary \( a^{(1)}, b^{(1)} \), if a temporal model defined by \( \{V_j^{(n)}\}_{j=1}^n \) is perfectly or approximately scalable, then so too is the model defined by \( \{a^{(n)}V_j^{(n)} + b^{(n)}\}_{j=1}^n \) where
\[
a^{(mn)} \equiv a^{(n)}, \quad b^{(mn)} \equiv b^{(n)}/m.
\] (8.14)

**Proof.** Given independent \( \{V_i^{(n)}\}_{i=1}^n \) associated with \( \Delta t \equiv T/n \) and any \( m \), there exists \( \{V_i^{(mn)}\}_{i=1}^n \) so that for \( 1 \leq j \leq n \), either 8.12 is satisfied, or 8.13 is satisfied for \( k \leq k_0 \). For \( \{a^{(n)}V_j^{(n)} + b^{(n)}\}_{j=1}^n \) to be analogously scalable we require the existence of i.i.d. \( \{V_i^{(nm)}\}_{i=1}^n \) so that for \( 1 \leq j \leq n \):

- **Perfectly Scalable:**
  \[
  \sum_{i=1}^{mj} \tilde{V}_i^{(mn)} = d a^{(n)} \sum_{i=1}^j V_i^{(n)} + j b^{(n)},
  \]

- **Approximately Scalable:**
  \[
  E \left[ \left( \sum_{i=1}^{mj} \tilde{V}_i^{(mn)} \right)^k \right] = E \left[ \left( a^{(n)} \sum_{i=1}^j V_i^{(n)} + j b^{(n)} \right)^k \right], \quad k \leq k_0.
  \]

Define \( \tilde{V}_i^{(nm)} \equiv a^{(mn)}V_i^{(nm)} + b^{(mn)} \) with \( V_i^{(nm)} \) as above. Then if 8.14 is satisfied these requirements reduce to:

- **Perfectly Scalable:**
  \[
  a^{(n)} \sum_{i=1}^{mj} V_i^{(mn)} + j b^{(n)} = d a^{(n)} \sum_{i=1}^j V_i^{(n)} + j b^{(n)},
  \]

- **Approximately Scalable:**
  \[
  E \left[ \left( a^{(n)} \sum_{i=1}^{mj} V_i^{(mn)} + j b^{(n)} \right)^k \right] = E \left[ \left( a^{(n)} \sum_{i=1}^j V_i^{(n)} + j b^{(n)} \right)^k \right], \quad k \leq k_0.
  \]

The result now follows from the respective assumption on \( \{V_j^{(n)}\}_{j=1}^n \). 

The final result addresses the following question. Can we build a harmonious model using affine transformations of a fixed random variable? In other words, can we build such a model by defining
\[
V^{(n)} \equiv a^{(n)}V + b^{(n)}
\] (8.15)
for appropriate \( \{a^{(n)}, b^{(n)}\} \)?
CHAPTER 8 DISCRETE TIME ASSET MODELS IN FINANCE

Proposition 8.15 Given arbitrary $a^{(1)}$, $b^{(1)}$, the temporal model defined in 8.15 by $\{a^{(n)}V_j + b^{(n)}\}_{j=1}^n$ is perfectly scalable if $V$ is stable (recall definition 7.23), and with $c_m > 0$ and $d_m$ given in 7.18 if:

$$a^{(mn)} = \frac{a^{(n)}}{c_m}, \quad b^{(mn)} = \frac{1}{m} \left( b^{(n)} - a^{(n)} \frac{d_m}{c_m} \right). \quad (8.16)$$

If $V$ has two moments with $E[V] = 0$, then the temporal model defined in 8.15 by $\{a^{(n)}V_j + b^{(n)}\}_{j=1}^n$ is approximately scalable to order $k_0 = 2$ if

$$a^{(mn)} = \frac{a^{(n)}}{\sqrt{m}}, \quad b^{(mn)} = \frac{b^{(n)}}{m}. \quad (8.17)$$

Proof. Let $\{a^{(n)}V_j + b^{(n)}\}_{j=1}^n$ and $m$ be given. Since $V$ is stable we have by 7.18 that there exists real $c_m > 0$ and $d_m$ so that:

$$\sum_{k=1}^m V_{jk} = d_m V_j + d_m,$$

where $\{V_{jk}\}$ are independent and have the distribution of $V$. Since $c_m > 0$ this obtains that

$$V_j = d_m \frac{1}{c_m} \sum_{k=1}^m \left( V_{jk} - \frac{d_m}{m} \right),$$

and thus for $1 \leq j \leq n$ :

$$\sum_{i=1}^j \left( a^{(n)}V_j + b^{(n)} \right) = \sum_{i=1}^j \left[ a^{(n)} \sum_{k=1}^m \left( V_{jk} - \frac{d_m}{m} \right) + b^{(n)} \right]$$

$$= \sum_{k=1}^m \sum_{i=1}^j \left[ a^{(n)} \sum_{k=1}^m V_{jk} + \frac{1}{m} \left( b^{(n)} - a^{(n)} \frac{d_m}{c_m} \right) \right]$$

$$= \sum_{k=1}^m \left[ a^{(mn)} V_k + b^{(mn)} \right],$$

and thus $\{a^{(n)}V_j + b^{(n)}\}_{j=1}^n$ is perfectly scalable.

Again let $\{a^{(n)}V_j + b^{(n)}\}_{j=1}^n$ and $m$ be given. Since $E[V] = 0$ it follows from $b^{(mn)} = b^{(n)}/m$ that:

$$E \left[ a^{(mn)} \sum_{i=1}^m V_i + m j b^{(mn)} \right] = E \left[ a^{(n)} \sum_{i=1}^j V_i^{(n)} + j b^{(n)} \right].$$

Equating second moments is equivalent to equating variances as noted in remark 8.12, and thus ignoring the constants:

$$Var \left[ a^{(mn)} \sum_{i=1}^m V_i \right] = Var \left[ a^{(n)} \sum_{i=1}^j V_i^{(n)} \right],$$

if $a^{(mn)} = a^{(n)}/\sqrt{m}$. ■
**Remark 8.16** The parameterizations in 8.14 and 8.17 are verifiably invariant to the sequencing of interval divisions. That is, the parameters in $a^{(mn)}V_j + b^{(mn)}$ for $\Delta t \equiv T/mn$ are the same if derived in one division of $[0,T]$ into $mn$ intervals, or if derived in two steps, first dividing into $n$ intervals, then subdividing these into $m$ subintervals.

This invariance is less apparent for the parametrization in 8.16. For the $a$-parameter, invariance requires that

$$\frac{a^{(n)}}{c_m} = \frac{a^{(1)}}{c_{mn}}.$$  

For this to make sense we must justify that $c_m/c_{mn}$ is independent of $m$, and this is true by 7.19. For the $b$-parameter, invariance requires:

$$\frac{1}{m} \left( b^{(n)} - a^{(n)} \frac{d_m}{c_m} \right) = \frac{1}{mn} \left( b^{(1)} - a^{(1)} \frac{d_{mn}}{c_{mn}} \right). \quad (*)$$

Now 8.16 also provides $b^{(n)} = \frac{1}{n} \left( b^{(1)} - a^{(1)} \frac{d_n}{c_n} \right)$, and when substituted along with $a^{(n)} = \frac{c_m}{c_{mn}} a^{(1)}$ we obtain:

$$\frac{1}{m} \left( b^{(n)} - a^{(n)} \frac{d_m}{c_m} \right) = \frac{1}{mn} \left( b^{(1)} - a^{(1)} \left( \frac{d_n}{c_m} + \frac{nd_m}{c_{mn}} \right) \right).$$

That this expression equals that on the right in $(*)$ again follows from 7.19.

**Example 8.17** An example of the result on approximate scalability was seen in example 8.6, where

$$Z_n = a^{(1)} = \sigma \sqrt{T/n}, \quad b^{(1)} = T/n,$$

with $\{b_n\}_{n=1}^\infty$ independent and binomially distributed to equal $\pm 1$ with probability $1/2$. Hence $E[b_n] = 0$, and with $\Delta t \equiv T/n$ the parameterization there is:

$$a^{(n)} = \sigma \sqrt{T/n}, \quad b^{(n)} = \mu T/n,$$

consistent with 8.17 with $a^{(1)} = \sigma \sqrt{T}$ and $b^{(1)} = \mu T$.

### 8.4.3 Harmonious Additive Temporal Models

Assume that we are given a spatial asset price vector $X = (X_1, X_2, ..., X_n)$ and independent, identically distributed additive temporal variates $Y = (Y_1, Y_2, ..., Y_n)$ related as in 8.5:

$$X_j = \sum_{i=1}^{j} Y_i.$$
When needed for clarity in the scalability discussion below we superscript variates as:
\[ X_j^{(n)} = \sum_{i=1}^{j} Y_i^{(n)}, \]
Recall that the significance of \( n \) was in defining the size of the time step \( \Delta t \equiv T/n \) where \([0, T]\) is the modeling period, and the convention noted above that \( X_0 \) is suppressed and hence \( X_n \) equals the \( n \)-period change in this variate. The actual value of the asset at time \( n \) is then \( X_0 + X_n \).

**Exercise 8.18** Note that in all cases, while \( \{Y_i\}_{i=1}^{n} \) are assumed to be independent, \( \{X_i\}_{i=1}^{n} \) can never be independent by construction. Show that if the variance \( \sigma_Y^2 \) of \( Y_j \) exists, the covariance of these variates is given by:
\[ \text{Cov}(X_j, X_k) = \min(j, k)\sigma_Y^2. \tag{8.18} \]
Without assuming the existence of \( \sigma_Y^2 \), use characteristic functions to show that if \( X_j \) and \( X_k \) are independent for \( j \neq k \), then the distribution function of \( Y \) is the delta function of 6.13 with \( x_0 = 0 \). This requires the uniqueness result of proposition 6.25. Thus \( \{Y_i\}_{i=1}^{n} \) cannot be independent, a contradiction.

The first criterion for a harmonious model, identifiability, is satisfied if the distribution family of the \( Y \)-variates is related under addition to the distribution family of the \( X \)-variates, as discussed in section 7.2. Recalling such examples, we summarize results below.

**Example 8.19 (Identifiability)** In practice, it is uncommon to abandon the assumption that the \( Y \)-variates are identically distributed since this also complicates the scalability discussion below. That said, we present the more general results below since they apply to the current discussion on identifiability.

1. If \( \{Y_i\}_{i=1}^{j} \) are independent **binomial with parameters** \( p \) and \( n = 1 \), then \( X_j \) is **binomial with parameters** \( p \) and \( n = j \). It is also the case that if \( Y_i \) has parameters \( p \) and \( n_i \), then \( X_j \) is binomial with parameters \( p \) and \( \sum_{i=1}^{j} n_i \).

2. If \( \{Y_i\}_{i=1}^{j} \) are independent **negative binomial with parameters** \( p \) and \( k_i \), then \( X_j \) is **negative binomial with parameters** \( p \) and \( k = \sum_{i=1}^{j} k_i \). Also, if all \( k_i = 1 \), then \( \{Y_i\}_{i=1}^{j} \) have **geometric distributions** and then \( X_j \) is **negative binomial with parameters** \( p \) and \( j \).
3. If \( \{Y_i\}_{i=1}^j \) are independent \textbf{Poisson with parameters} \( \lambda_i \), then \( X_j \) is \textbf{Poisson with parameter} \( \lambda = \sum_{i=1}^j \lambda_i \).

4. If \( \{Y_i\}_{i=1}^j \) are independent \textbf{gamma with parameters} \( \lambda \) and \( \alpha_i \), then \( X_j \) is \textbf{gamma with parameters} \( \lambda \) and \( \alpha = \sum_{i=1}^j \alpha_i \). If all \( \alpha_i = 1 \), then \( \{Y_i\}_{i=1}^j \) have \textit{exponential distributions} and then \( X_j \) is \textbf{gamma with parameters} \( \lambda \) and \( j \).

5. If \( \{Y_i\}_{i=1}^j \) are independent \textbf{Cauchy variates with parameters} \( \gamma_i \) and \( x_i \), then \( X_j \) is \textbf{Cauchy with parameters} \( x_0 = \sum_{i=1}^j x_i \) and \( \gamma = \sum_{i=1}^j \gamma_i \).

6. If \( \{Y_i\}_{i=1}^j \) are independent \textbf{normal variates with parameters} \( \mu_i \) and \( \sigma_i^2 \), then \( X_j \) is \textbf{normal with parameters} \( \mu = \sum_{i=1}^j \mu_i \) and \( \sigma^2 = \sum_{i=1}^j \sigma_i^2 \).

7. If \( \{Y_i\}_{i=1}^j \) are independent \textbf{lognormal variates with parameters} \( \mu \) and \( \sigma^2 \), then \( X_j \) has an \textit{unknown distribution}, although from 3.68 of book 4 it can be concluded that the \( k \)th moment of \( X_j \) satisfies:

\[
\mu_k^j(X_j) = je^{k\mu + (k\sigma)^2/2}.
\]

As was the case in remark 3.30 of book 4, the power series \( \sum_{k=0}^\infty \mu_k^j(X_j)t^k/k! \) does not converge absolutely on \((-t_0,t_0)\) for any \( t_0 > 0 \), and thus proposition 3.57 of book 4 can not be employed to even assert that these moments uniquely determine the distribution of \( X_j \). So this distribution is not only unknown, but may not even be uniquely defined. Indeed, the lognormal distribution possesses the moments of infinitely many distributions as illustrated in example 3.53 of book 4.

8. If \( \{Y_i\}_{i=1}^j \) are independent \textbf{compound Poisson variates with parameters} \( \lambda_i \) and \textit{driving variate in 7.20 denoted} \( W \) (rather than \( X \) to avoid notational confusion), then \( X_j \) is \textbf{compound Poisson with parameters} \( \lambda = \sum_{i=1}^j \lambda_i \) and \textit{driving variate} \( W \).

**Example 8.20 (Scalability)** Reviewing the above distributions in order, but now restricting \( \{Y_i\}_{i=1}^j \) to be identically distributed, we conclude the following.

1. If \( X_j \) is \textbf{binomial with parameters} \( p \) and \( j \), then \( M_{X_j}(t) = (1 + p(e^t - 1))^j \) and the distribution function of \( X_j \) is not \textit{infinitely divisible} by 5 of proposition 7.29. Hence this model is \textbf{not perfectly}
scalable by proposition 8.13. However given \( m \) and the desire to approximately represent each \( X_j = \sum_{i=1}^{m_j} Y_i^{(mn)} \):

\[
[M_{X_j}(t)]^{1/m_j} = (1 + p(e^t - 1))^{1/m} \\
\approx 1 + \frac{1}{m} p(e^t - 1) + O(m^{-1}) .
\]

Thus the binomial model is potentially scalable with \( Y_i^{(mn)} \) binomial with parameters \( \frac{p}{m} \) and \( n = 1 \). With the variate \( Y_i^{(mn)} \) defined this way, we achieve approximate scalability to order \( k_0 = 1 \) but not \( k_0 = 2 \):

\[
E[hX_j^{(mn)}] = j p = E\left[\sum_{i=1}^{j} Y_i^{(mn)}\right] ,
\]

\[
\text{Var}[\sum_{i=1}^{m_j} Y_i^{(mn)}] = j p \left(1 - \frac{p}{m}\right) \neq j p (1 - p) = \text{Var}\left[\sum_{i=1}^{j} Y_i^{(mn)}\right] .
\]

In addition, the discrepancy in the variance estimate worsens with \( m \).

By using an affine transformation of a fixed binomial for \( \{Y_i\}_{i=1}^{j} \) a better result is produced. Analogous to the classical binomial model for equity prices noted in example 8.6 above, define \( X_j = \sum_{i=1}^{j} \left(\mu \Delta t + \sigma \sqrt{\Delta t} b_i\right) \),

with \( b_i \) independent and binomially distributed to equal 1 or -1 with probability \( p = 1/2 \). As in the above model, \( \Delta t = T/n \) for some fixed horizon time \( T \) and so \( X_j \) denotes the change in price up to time \( j \Delta t \), and the final value \( X_n \) denotes the change in price up to time \( T = n \Delta t \).

Rewriting with superscripts:

\[
X_j^{(n)} = \sum_{i=1}^{j} \left(\mu T/n + \sigma \sqrt{T/n} b_i\right) ,
\]

and thus \( Y_i^{(n)} = c^{(n)} b_i + d^{(n)} \) with \( c^{(n)} = \sigma \sqrt{T/n} \) and \( d^{(n)} = \mu T/n \).

Given \( m \), we now model \( X_j^{(n)} \) with binomial \( X_{mj}^{(mn)} \) defined by:

\[
X_{mj}^{(mn)} = \sum_{i=1}^{m_j} \left(\mu T/mn + \sigma \sqrt{T/mn} b_i\right) ,
\]

in effect by simply redefining \( \Delta t' = T/mn \). So the binomial \( X_j^{(n)} \)-distribution is approximated by the binomial \( X_{mj}^{(mn)} \)-distribution, \( X_j^{(n)} \approx X_{mj}^{(mn)} \), where now \( Y_i^{(mn)} = c^{(mn)} b_i + d^{(mn)} \) with \( c^{(mn)} = \sigma \sqrt{T/mn} \) and \( d^{(mn)} = \mu T/mn \). With this parametrization we have approximate
8.4 HARMONIOUS ASSET MODELS

scalability to order $k_0 = 2$, meaning that with $\Delta t = T/n$:

$$
E\left[ X^{(n)}_j \right] = \mu_j \Delta t = E\left[ X^{(mn)}_{mj} \right],
$$

$$
\text{Var}\left[ X^{(n)}_j \right] = \sigma^2_j \Delta t = \text{Var}\left[ X^{(mn)}_{mj} \right].
$$

That $X^{(n)}_j$ and $X^{(mn)}_{mj}$ are actually different binomial random variables can be appreciated by evaluating the respective moment generating functions.

Note that this model can be generalized for $p \neq 1/2$, but this requires re-defining $c^{(n)}$ and $d^{(n)}$ above in order to maintain the original calibrated values of $E\left[ X^{(n)}_j \right]$ and $\text{Var}\left[ X^{(n)}_j \right]$. See chapter 8 of the Reitano (2010) reference for more on this calibration.

2. If $X_j$ is negative binomial with parameters $p$ and $jk$, then the distribution function of $X_j$ is infinitely divisible and hence this model is perfectly scalable with $Y^{(mn)}_i$ independent negative binomial with parameters $p$ and $k/m$.

3. If $X_j$ is Poisson with parameter $j\lambda$, then the distribution function of $X_j$ is infinitely divisible and thus perfectly scalable with $Y^{(mn)}_i$ independent Poisson with parameter $\lambda/m$.

4. If $X_j$ is gamma with parameters $\lambda$ and $j\alpha$, then the distribution function of $X_j$ is infinitely divisible and thus perfectly scalable with $Y^{(mn)}_i$ independent gamma with parameters $\lambda$ and $\alpha/m$.

5. If $X_j$ is Cauchy with parameters $jx_0$ and $j\gamma$, then the distribution function of $X_j$ is infinitely divisible and thus perfectly scalable with $Y^{(mn)}_i$ independent Cauchy with parameters $x_0/m$ and $\gamma/m$.

6. If $X_j$ is normal with parameters $\mu_j$ and $\sigma_j^2$, then the distribution function of $X_j$ is infinitely divisible and thus perfectly scalable with $Y^{(mn)}_i$ independent normal with parameters $\mu/m$ and $\sigma^2/m$.

7. If $X_j$ has the unknown distributions associated with lognormal $Y_i$ with parameters $\mu$ and $\sigma^2$, then $X_j$ cannot be perfectly scalable with lognormal variates since it is then not even possible to reproduce the moments of $X_j$. Specifically, if $Y^{(mn)}_i$ are independent lognormal
with parameters $\mu(m)$ and $\sigma^2(m)$, then with $X_j^{(mn)} = \sum_{i=1}^{m_j} Y_i^{(mn)}$, the $k$th moment is given by:

$$\mu_k'(X_j^{(mn)}) = m_j \exp \left[ k\mu(m) + k^2\sigma^2(m)/2 \right].$$

Comparing this to the expression above for $\mu_n'(X_j)$, at most one moment can be matched for $j$ fixed with $\mu(m) = \mu + \frac{1}{k}\ln j$, $\sigma^2(m) = \sigma^2$, but this parametrization does not work simultaneously for all $j$, and thus this model is also not approximately scalable.

8. If $X_j$ is compound Poisson with parameters $\lambda_j$ and driving variate $W$, then the distribution function of $X_j$ is infinitely divisible and thus perfectly scalable with $Y_i^{(mn)}$ independent compound Poisson with parameters $\lambda/m$ and driving variate $W$.

### 8.4.4 Harmonious Multiplicative Temporal Models

Assume we are given spatial $X = (X_1, X_2, ..., X_n)$ and temporal returns $Z = (Z_1, Z_2, ..., Z_n)$, where these variates are related as in 8.8 by:

$$X_j = \prod_{i=1}^{j} \exp Z_i = \exp \left[ \sum_{i=1}^{j} Z_i \right].$$

As in the additive temporal model above, when needed for notational clarity in the scalability discussion below we superscript variates as:

$$X_j^{(n)} = \exp \left[ \sum_{i=1}^{j} Z_i^{(n)} \right],$$

recalling that the significance of $n$ was in defining the size of the time step, $\Delta t \equiv T/n$, where $[0, T]$ is the modeling period.

The first criterion for a harmonious model of identifiability is satisfied if the distribution family of the $Z$s is related under addition to the distribution family of $W$s where $W_j \equiv \sum_{i=1}^{j} Z_i$. Thus the models in example 8.20 apply equally well in this context, substituting $Z$ for $Y$ and $W$ for $X$. Then if $W_j \equiv \sum_{i=1}^{j} Z_i$ has distribution function $F_{w_j}$, then $X_j = \exp \left[ \sum_{i=1}^{j} Z_i \right]$ has a log-$F_{w_j}$ distribution as defined in 7.28 or 7.29.

For the second criterion of scalability, if the distribution function $F_{W_n}$ for $W_n$ is infinitely divisible, then so too is $F_{W_j}$ for all $j$ and hence this model is perfectly scalable. That is, for any $m$ there are independent, identically distributed variates $\{Z_i^{(mn)}\}_{i=1}^{m}$ so that with $W_{mj}^{(mn)} \equiv \sum_{i=1}^{m_j} Z_i^{(mn)}$:

$$W_j^{(n)} \overset{d}{=} W_{mj}^{(mn)}.$$
8.4 HARMONIOUS ASSET MODELS

Defining:

\[ X_{m_j}^{(mn)} = \exp W_{m_j}^{(nm)} , \]

it then follows that the notion of perfect scalability also applies to \( X \) in that for all \( j \):

\[ X_j^{(n)} =_d X_{m_j}^{(mn)} . \]

Thus the perfectly scalable models of example 8.20 work within the multiplicative framework as well, and the resulting asset prices also satisfy this scalability criterion.

If the \( F_W \)-distributions are only approximately scalable, meaning that the first \( k_0 \) moments of \( \sum_{i=1}^{j} Z_i^{(n)} \) match those of \( \sum_{i=1}^{m_j} Z_i^{(mn)} \), then by definition the resulting multiplicative temporal models are approximately scalable. Thus additive temporal models give rise to multiplicative temporal models by definition. However, unlike the perfectly scalable case, the implication for the moments of \( X_j \) is generally less favorable as seen below. But first a result.

**Proposition 8.21** Assume that \( W_j \equiv \sum_{i=1}^{j} Z_i \) has distribution function \( F_{W_j} \) and that the moments of \( X_j \equiv \exp W_j \) exist up to order \( k \). Then \( M_{W_j}(t) \) exists for \( |t| \leq k \), and for \( l \leq k \):

\[ E \left[ (X_j)^l \right] = M_{W_j}(l). \]

**Proof.** With \( F_X \) the distribution function of \( X_j \) and \( l \leq k \):

\[ E \left[ (X_j)^l \right] = \int_0^\infty x^l dF_X, \]

expressed as a Lebesgue-Stieltjes integral. By proposition 1.7, \( X_j \) is defined on some probability space \( (\mathcal{S}, \mathcal{E}, \lambda) \), so \( X_j : (\mathcal{S}, \mathcal{E}, \lambda) \rightarrow (\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), \mu_{F_X}) \) with \( \mu_{F_X} \) the Borel measure induced by \( F_X \). Define the transformation

\[ T : (\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), \mu_{F_X}) \rightarrow (\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+), (\mu_{F_X})_T) \]

by \( Tx = \ln x \), where \((\mu_{F_x})_T\) is the Borel measure induced by \( T \) as in definition 3.9 of book 5. Defining \( W \equiv \ln X \), a calculation obtains that \((\mu_{F_X})_T = \mu_{F_W}\). Then proposition 3.14 of book 5 provides with \( g(x) \equiv e^x \) and \( g(Tx) = x \):

\[ \int_0^\infty x^l dF_X = \int_{-\infty}^\infty e^{lw} dF_W = M_{W_j}(l). \]

Thus if the \( l \)th moment of \( X_j \) exists, so too does the moment generating function of \( W_j \) for \( t = l \), and hence \( M_{W_j}(t) \) exists for all \( |t| \leq l \). This last
conclusion follows from the proof of proposition 3.22 of book 4, that existence of \( MW_j(l) \) assures integrability of \( e^{lw} \), which then assures integrability of \( e^{tw} \) for \( |t| \leq l \).

If \( F_W_j \) is approximately scalable with \( W_j^{(mn)} = \sum_{i=1}^{m_j} Z_i^{(mn)} \) having distribution function \( F_W_j^{(mn)} \), and if \( X_j^{(mn)} \equiv \exp W_j^{(mn)} \) has a \( k \)th moment, then by proposition 8.21:

\[
E \left[ (X_j^{(mn)})^k \right] = MW_j^{(mn)}(k).
\]

Hence if \( E \left[ (W_j^{(n)})^k \right] = E \left[ (W_j^{(mn)})^k \right] = \mu_k \) for \( k \leq k_0 \), the associated moment generating functions agree to the \( k_0 \)th term and so applying proposition 3.24 of book 4:

\[
MW_j(t) = \sum_{i=0}^{k_0} i^t \mu'_i / i! + \sum_{i=k_0+1}^{\infty} i^t \mu'_i (W_j^{(n)}) / i!,
\]

\[
MW_j^{(mn)}(t) = \sum_{i=0}^{k_0} i^t \mu'_i / i! + \sum_{i=k_0+1}^{\infty} i^t \mu'_i (W_j^{(mn)}) / i!.
\]

Thus for \( k \leq k_0 \):

\[
E \left[ (X_j)^k \right] - E \left[ (X_j^{(mn)})^k \right] = \sum_{i=k_0+1}^{\infty} k^i \left[ \mu'_i(W_j) - \mu'_i(W_j^{(mn)}) \right] / i!.
\]

Hence if \( F_W_j \) is approximately scalable to order \( k_0 \), even lower order moments of \( X_j \) and \( X_j^{(mn)} \) will in general not agree. However, a useful harmonious model is still achievable.

**Example 8.22** Recall example 8.10 above of log-binomial \( X_j \equiv \exp W_j \) for \( 1 \leq j \leq n \) with binomial \( W_j \) so that:

\[
X_j = \exp \left[ \sum_{i=1}^{j} (a^{(n)}b_i + b^{(n)}) \right].
\]

The parameters of this example are defined so that with fixed horizon time \( T \) and initial \( \Delta t = T/n \), the price \( X_j = X_j^{(n)} \) represents the price at time \( j\Delta t = jT/n \). Also, \( b_i \) is a binomial distributed to equal 1 or \(-1\) with probability \( p = 1/2 \), while \( a^{(n)} = \sigma \sqrt{T/n} \) and \( b^{(n)} = \mu T/n \). The distribution function of \( W_j^{(n)} = \sum_{i=1}^{j} (a^{(n)}b_i + b^{(n)}) \) is approximately scalable, and it was seen that with \( W_j^{(mn)} = \sum_{i=1}^{m_j} (a^{(mn)}b_i + b^{(mn)}) \), \( a^{(mn)} = \sigma \sqrt{T/mn} \) and \( b^{(mn)} = \)
that the distribution functions of $W_{nj}$ and $W_{mnj}$ agreed to two moments.

By proposition 8.21 we can compare the moments of $X_{nj} \equiv \exp W_{nj}$
and $X_{mnj} \equiv \exp W_{mnj}$ by evaluating the moment generating functions of
$W_{nj}$ and $W_{mnj}$. As a sum of independent random variables:

$$M_{W_{nj}}(k) = \left( \exp \left[ \left( \mu T/n + \sigma \sqrt{T/n} \right) k \right] + \exp \left[ \left( \mu T/n - \sigma \sqrt{T/n} \right) k \right] \right]^j / 2^j$$

$$= \exp [\mu k j T/n] (\exp [\sigma k \sqrt{T/n}] + \exp [-\sigma k \sqrt{T/n}])^j / 2^j,$$

and similarly:

$$M_{W_{mnj}}(k) = \exp [\mu k j T/n] (\exp [\sigma k \sqrt{T/mn}] + \exp [-\sigma k \sqrt{T/mn}])^{mj} / 2^{mj}.$$ 

If $M_{W_{nj}}(k) = M_{W_{mnj}}(k)$ for some $k$, this would require that with $c \equiv \exp [\sigma k \sqrt{T/n}] > 1$ and $m > 1$:

$$c + c^{-1} = \left( \frac{c^{1/\sqrt{m}} + c^{-1/\sqrt{m}}}{2} \right)^m.$$

However, this is not possible since the function on the right is increasing for
$m \geq 1$, and so equality occurs only when $c = 1$. Thus $M_{W_{nj}}(k) \neq M_{W_{mnj}}(k)$
for any $k > 0$, and so by the above discussion $E \left[ (X_{nj}^{(n)})^k \right] \neq E \left[ (X_{mnj}^{(mn)})^k \right]$ for any $k$.

Thus $X_j$ is not "approximately scalable" with this model in terms of the
matching of lower moments, illustrating the general result above. However,
it is an exercise to show that for any real number $t$, $M_{W_{mnj}}^{(mn)}(t)$ has a well
deﬁned limit as $m \to \infty$, and speciﬁcally:

$$M_{W_{mnj}}^{(mn)}(t) \to \exp \left[ \mu (jt/n) t + \frac{1}{2} \sigma^2 (jt/n) t^2 \right]. \quad (\star)$$

This implies that for $m$ large, that $M_{W_{mnj}}^{(mn)}(t)$ is nearly independent of $m$,
and hence, so too is $E \left[ (X_{mnj}^{(mn)})^k \right]$. 

Remark 8.23 As an introduction to the topic of the next section, note that
(\ast) implies that for all \( j,n \) with \( 0 \leq j \leq n \), the limiting distribution of
\( W_{j}^{(mn)} \) as \( m \to \infty \) is normal with mean \( \mu jT/n \) and variance \( \sigma^2 jT/n \). This
follows because by 3.66 of book 4 the function in the limit is the moment
generating function of this variate, while corollary 3.58 of book 4 assures
that the normal variate is uniquely determined by this moment generating
function. This conclusion can also be derived by corollary 3.74 of book 4.

It then follows by the above discussion that for \( t = k \), the expression
on the right in (\ast) is the value of the limit of \( E \left[ \left( X_{mj}^{(mn)} \right)^k \right] \) as \( m \to \infty \),
while the lognormal distribution with parameters \( \mu jT/n \) and \( \sigma^2 jT/n \) also
has these same moments by 3.68 of book 4. However, example 3.54 of book
4 illustrates that the lognormal distribution is not uniquely determined by
these moments. Thus based only on (\ast) we can only say that as \( m \to \infty \) the
limiting distribution of the price \( X_t \) for \( t = jT/n \) is consistent with that of
the lognormal distribution with these parameters.

But we can say more by noting that:

\[
\Pr \left[ X_{mj}^{(mn)} \leq x \right] = \Pr \left[ \ln X_{mj}^{(mn)} \leq \ln x \right] = \Pr \left[ W_{mj}^{(mn)} \leq \ln x \right].
\]

Thus by the first paragraph:

\[
\Pr \left[ X_{j}^{(mn)} \leq x \right] \to \Phi \left( \frac{\ln x - \mu (jT/n)}{\sqrt{\sigma^2 jT/n}} \right),
\]

where \( \Phi \) is the distribution function of the standard normal in 3.2. This
then confirms that the limiting distribution of the price \( X_t \) is lognormal.

8.5 Limiting Distributions of Harmonious Asset
Models

In this section we discuss limiting distributions of harmonious asset models
as \( m \to \infty \). As above, in the typical application one has a fixed future time
horizon denoted \( T \), time-steps initially defined by \( \Delta t = T/n \) for some
\( n \geq 1 \), and refinements of these time-steps obtained by dividing \( \Delta t \) by \( m \).
Thus the distributional properties of the model as \( m \to \infty \) can be
equivalently framed in terms of the distributional properties as
\( \Delta t' = T/mn \to 0 \).
After quickly addressing perfectly scalable harmonious models, for which there is no question on limiting distributions, the following result will derived. Under the hypothesis that the model is approximately scalable to order \( k_0 \geq 2 \), there are only two possible limiting distributions for harmonious asset models which satisfy a regularity condition on the temporal variates:

1. Additive Temporal models ⇒ Normal distribution
2. Multiplicative Temporal models ⇒ Lognormal distribution

Subject to this regularity condition, these results are independent of the subinterval distributions of the \( Y_i^{(n)} \) and \( Z_i^{(n)} \) variates, and not surprisingly reflect an application of the central limit theorems of section 7.1.

The significance of this result is that in order to achieve other spatial distributions, a more general approach is needed for the specification of the associated temporal distributions. Specifically, one must either abandon independence of these variates, or abandon the assumption that these variates are identically distributed, or both. This then introduces the question of how such distributions are specified, and what are their properties, and this general investigation is the topic of the next three books.

### 8.5.1 Perfectly Scalable Harmonious Models

We can dispense with this class of models quickly because by definition, these models are independent of \( m \) and hence there is no limit to consider. Specifically, by definition of perfectly scalable, \( X_{mj}^{(mn)} = d X_j^{(n)} \) for all \( m \) in the additive model. In the multiplicative model, \( W_{mj}^{(mn)} = d W_j^{(n)} \) for all \( m \) where \( W_j^{(n)} = \sum_{i=1}^{j} Z_i^{(n)} \), and thus since \( X = \exp W \) it follows again that \( X_{mj}^{(mn)} = d X_j^{(n)} \) for all \( m \). Recalling that for given \( n \) and \( \Delta t = T/n \) that \( X_j^{(n)} \) denotes the asset price at time \( t = j\Delta t \), this implies that in a perfectly scalable model the distribution of \( X_t \) is independent of partitioning of the interval \([0, T]\) for all \( t = jT/n \), for all \( j,n \).

We illustrate this with a common example of an additive temporal model, but as noted the result is perfectly general, indeed by definition.

**Example 8.24** Assume spatial \( X = (X_1, X_2, ..., X_n) \) and additive temporal \( Y = (Y_1, Y_2, ..., Y_n) \) are related by:

\[
X_j = \sum_{i=1}^{j} Y_i,
\]
where \(\{Y_i\}\) are independent and normally distributed with mean 0 and variance \(\Delta t = T/n\) for some fixed \(n\). Then all \(X_j\) are normally distributed with mean 0 and variance \(\frac{\sigma^2}{m}\) for some fixed \(m\). Then all \(X_j\) are normally distributed with mean 0 and variance \(\frac{\sigma^2}{m}\) and hence perfectly scalable. For any \(m\) define \(Y^{(mn)} = (Y_1^{(mn)}, Y_2^{(mn)}, ..., Y_m^{(mn)})\), where \(Y_i^{(mn)}\) are independent and normally distributed with mean 0 and variance \(\Delta t/m = T/nm\). Then with

\[ X_j^{(mn)} = \sum_{i=1}^{mj} Y_i^{(mn)}, \]

we have by construction that \(X_j^{(mn)}\) has the same distribution as \(X_j\), which is normal with parameters 0 and \(j\Delta t\).

Moreover, it is an exercise in the multivariate normal distribution and change of variables that the joint distribution of the original \((X_1, X_2, ..., X_n)\) is the same as the joint distribution of \((X_m^{(mn)}, X_2^{(mn)}, ..., X_n^{(mn)})\).

### 8.5.2 Approximately Scalable Harmonious Models

For approximately scalable models, the question of limiting distributions is of apparent interest in cases where one desires to investigate properties of such models as \(m \to \infty\), or equivalently, as \(\Delta t' = T/mn \to 0\). First note, however, that limiting distributions need not exist nor be useful.

**Example 8.25** Modifying the binomial model in example 8.20 somewhat, we begin with \(X_j^{(n)} = \sum_{i=1}^{j} (\mu \Delta t + \sigma \sqrt{\Delta t} b_i)\) with \(b_i\) independent and binomially distributed to equal 1 or \(-1\) with probability \(p = 1/2\). As always, \(\Delta t = T/n\) for some fixed horizon time \(T\), and hence \(X_j^{(n)}\) is the value of the change in the spatial variate to time \(j\Delta t = jT/n\). Given \(m\), we now approximate \(X_j^{(n)}\) with binomial \(X_{mj}^{(mn)}\) defined by:

\[ X_{mj}^{(mn)} = \sum_{i=1}^{mj} \left( \mu T/mn + \sigma^m \sqrt{T/mnb_i} \right). \]

With this parametrization, the model is approximately scalable to order \(k_0 = 1\), since:

\[ E\left[ X_j^{(n)} \right] = \mu jT/n = E\left[ X_{mj}^{(mn)} \right]. \]

However, since \(\text{Var}\left[ X_j^{(n)} \right] = \sigma^2 jT/n\), and:

\[ \text{Var}\left[ X_{mj}^{(mn)} \right] = \sigma^2 m jT/n, \]

scalability to order 2 occurs only if \(\sigma = 1\). Further:
8.5 LIMITING DISTRIBUTIONS OF HARMONIOUS ASSET MODELS

1. If $0 < \sigma < 1$ then $\text{Var} \left[ X_{mnj}^{(mn)} \right] \to 0$ as $m \to \infty$, and hence by Chebyshev’s inequality (proposition 3.32 of book 4), $X_{mnj}^{(mn)} \Rightarrow \mu jT/n$. In other words, $X_{mnj}^{(mn)}$ converges weakly to the degenerate distribution with value $\mu jT/n = E \left[ X_j^{(n)} \right]$.

2. If $\sigma > 1$ then $\text{Var} \left[ X_{mnj}^{(mn)} \right] \to \infty$ as $m \to \infty$, and the same is true for all even central moments:

$$E \left[ \left( X_{mnj}^{(mn)} - \mu jT/n \right)^{2k} \right] \to \infty,$$

while all odd central moments are 0 by symmetry.

3. For any $\sigma > 0$, recall Lyapunov’s condition in 7.12. Letting $\delta > 0$, then as $m \to \infty$:

$$\frac{1}{(\sigma^2 m jT/n)^{1+\delta/2}} \sum_{i=1}^{m_j} E \left[ \left( \sigma^m \sqrt{T/mnb_i} \right)^{2+\delta} \right] = \frac{1}{(m_j)^{\delta/2}} \to 0.$$

Thus by Lyapunov’s central limit theorem of proposition 7.15, as $m \to \infty$:

$$\frac{X_{mnj}^{(mn)} - \mu jT/n}{\sqrt{\sigma^2 m jT/n}} \Rightarrow Z,$$

the standard normal variate $Z$. Thus for any $a > 0$:

$$\Pr \left[ \left| X_{mnj}^{(mn)} - \mu jT/n \right| < a \right] \approx \Phi \left( a/\sqrt{\sigma^2 m jT/n} \right) - \Phi \left( -a/\sqrt{\sigma^2 m jT/n} \right),$$

and this probability converges to 0 for $\sigma > 1$, and to 1 for $0 < \sigma < 1$.

While potentially interesting as a cautionary tale, it is unlikely that the limiting distributions of $X_{mnj}^{(mn)}$ would be useful in practice. Thus approximately scalable models do not necessarily yield useful results as $\Delta t \to 0$.

Additive Temporal Models

The prior example shows what can go wrong in an approximately scalable, additive temporal model. For the current analysis note that for any fixed $n$ and $j$ with $1 \leq j \leq n$, the collections of random variables $\{Y_i^{(mn)}\}_{i=1}^{m_j}$ for $m = 1, 2, \ldots$, form a triangular array as introduced in section 7.1.2. Thus the question of the limiting distribution of $X_{mnj}^{(mn)} \equiv \sum_{i=1}^{m_j} Y_i^{(mn)}$ is precisely the question addressed by the Lindeberg and Lyapunov central limit theorems of propositions 7.6 and 7.15.
For the following result, assume that we are given time horizon $T$, spatial $X^{(n)} = (X_1^{(n)}, X_2^{(n)}, \ldots, X_n^{(n)})$, and additive temporal $Y^{(n)} = (Y_1^{(n)}, Y_2^{(n)}, \ldots, Y_n^{(n)})$ related by:

$$X_j^{(n)} = X_0 + \sum_{i=1}^{j} Y_i^{(n)}.$$ 

Here $X_0$ denotes the initial asset value, $\{Y_i^{(n)}\}_{i=1}^{n}$ are independent and identically distributed on some space $(\mathcal{S}, \mathcal{E}, \lambda)$ and with distribution function appropriate for time-step $\Delta t \equiv T/n$ for some $n \geq 1$. Assume that this temporal model is approximately scalable to order $k_0$, so that given $m$ there are independent, identically distributed $\{Y_{i^{(mn)}}\}_{i=1}^{m}$ so that for $k \leq k_0$ and $1 \leq j \leq n$:

$$E \left[ \left( X_j^{(m)} \right)^k \right] = E \left[ \left( X_{mj}^{(mn)} \right)^k \right],$$

where:

$$X_{mj}^{(mn)} \equiv X_0 + \sum_{i=1}^{mj} Y_i^{(mn)}.$$ 

Note that the criterion for approximate scalability is not changed by the presence of the additive factor $X_0$.

**Proposition 8.26 (Additive Temporal Models)** With the above notation, define $\bar{Y}_m \equiv (Y^{(mn)} - \mu^{(mn)})/\sigma^{(mn)}$ where $\mu^{(mn)} \equiv E \left[ Y^{(mn)} \right]$ and $(\sigma^{(mn)})^2 \equiv \text{Var} \left[ Y^{(mn)} \right]$, and assume that for all $t > 0$:

$$\lim_{m \to \infty} \int_{|\bar{Y}_m| > t \sqrt{m}} \bar{Y}_m^2 d\lambda = 0. \quad (8.19)$$

Assume that either:

1. The additive temporal model is approximately scalable to order $k_0 \geq 2$, or more generally:

2. As $m \to \infty$:

$$m \mu^{(mn)} \to E \left[ X_1^{(n)} \right], \quad m \left( \sigma^{(mn)} \right)^2 \to \text{Var} \left[ X_1^{(n)} \right]. \quad (8.20)$$

Then as $m \to \infty$, for all $j$ with $1 \leq j \leq n$:

$$\frac{X_{mj}^{(mn)} - E \left[ X_j^{(n)} \right]}{\sqrt{\text{Var} \left[ X_j^{(n)} \right]}} \Rightarrow Z, \quad (8.21)$$
8.5 LIMITING DISTRIBUTIONS OF HARMONIOUS ASSET MODELS

with \( Z \) the standard normal variate. In other words, \( F_{X^{(mn)}_{mj}} \) converges in distribution to the normal distribution with mean \( E[X^{(n)}_j] \) and variance \( \text{Var}[X^{(n)}_j] \).

**Proof.** For the limiting distribution result for given \( j \), we recall Lindeberg’s central limit results applied to the limiting distribution of:

\[
\frac{\sum_{i=1}^{m_j} (Y_i^{(mn)} - E[Y_i^{(mn)}])}{\sqrt{\sum_{i=1}^{m_j} \text{Var}[Y_i^{(mn)}]}}.
\]

If the triangular array \( \{\{Y_i^{(mn)}\}_{i=1}^{m_n}\}_{n=1}^{\infty} \) satisfies the Lindeberg condition in 7.7 for any \( j \), then 8.21 will be proved by proposition 7.15 in case 1 since by definition of approximate scalability:

\[
\frac{\sum_{i=1}^{m_j} (Y_i^{(mn)} - E[Y_i^{(mn)}])}{\sqrt{\sum_{i=1}^{m_j} \text{Var}[Y_i^{(mn)}]}} = \frac{X_{mj}^{(mn)} - E[X^{(n)}_j]}{\sqrt{\text{Var}[X^{(n)}_j]}}.
\]

Similarly, this will prove the more general case 2 by corollary 7.9, since \( \{Y_i^{(mn)}\}_{i=1}^{m_n} \) are independent and identically distributed for each \( m \), and so 8.20 assures that for all \( j \):

\[
m_j \mu^{(mn)} \rightarrow E[X^{(n)}_j], \quad m_j (\sigma^{(mn)})^2 \rightarrow \text{Var}[X^{(n)}_j].
\]

Now since \( \{Y_i^{(mn)}\}_{i=1}^{m_n} \) are i.i.d. for each \( m \), it follows from exercise 7.8 that \( \{Y_i^{(mn)}\}_{i=1}^{m_n} \) satisfies Lindeberg’s condition if in the current notation:

\[
\lim_{m \rightarrow \infty} \int_{|\tilde{Y}_m| > t \sqrt{m_j}} \tilde{Y}_m^2 d\lambda = 0,
\]

for all \( t > 0 \), and with \( \tilde{Y}_m \) defined above. This now follows from 8.19 and the proposition is proved.

**Remark 8.27** The condition in 8.19 may perhaps seem to be easily satisfied. Of course it cannot be satisfied for perfectly scalable models, since then \( X^{(n)}_j \) is infinitely divisible by proposition 8.13 and the distribution functions of \( X_{mj}^{(mn)} \) and \( X^{(n)}_j \) agree. Thus the limiting result in 8.21 would be impossible unless \( X^{(n)}_j \) is normally distributed.
In the general case, Chebyshev’s inequality (proposition 3.32, book 4) and \( \text{Var}(\bar{Y}_m) = 1 \) obtains only that:

\[
\int_{|\bar{Y}_m| > t\sqrt{m}} \bar{Y}_m^2 d\lambda \leq mt^2 \text{Pr}[|\bar{Y}_m| > t\sqrt{m}] \leq 1.
\]

Indeed, one can identify approximately scalable models where this upper bound is achieved.

**Example 8.28** Assume that \( E[Y^{(n)}] = \mu^{(n)} \) and \( \sqrt{\text{Var}[Y^{(n)}]} = \sigma^{(n)} \), and define:

\[
Y^{(mn)} = \begin{cases} 
-\sqrt{m}\sigma^{(n)} + \mu^{(n)}/m, & p = 1/2m^2, \\
\mu^{(n)}/m, & p = 1 - 1/m^2, \\
\sqrt{m}\sigma^{(n)} + \mu^{(n)}/m, & p = 1/2m^2.
\end{cases}
\]

A calculation obtains that \( E[Y^{(mn)}] = \mu^{(n)}/m \) and \( \sqrt{\text{Var}[Y^{(mn)}]} = \sigma^{(n)}/\sqrt{m} \), and thus this model is approximately scalability to order \( k_0 = 2 \). Also:

\[
\bar{Y}^{(mn)}_t = \begin{cases} 
-m, & p = 1/2m^2, \\
0, & p = 1 - 1/m^2, \\
m, & p = 1/2m^2,
\end{cases}
\]

and letting \( t = 1 \) we see that for all \( m \):

\[
\int_{|\bar{Y}_m| > \sqrt{m}} \bar{Y}_m^2 d\lambda = 1.
\]

In addition, the limit in 8.19 is 1 for all \( t \).

**Corollary 8.29 (Additive Temporal Models)** With the above notation, assume that for some \( \delta > 0 \) that \( \bar{Y}_m^{2+\delta} \) is integrable and that:

\[
\lim_{m \to \infty} \frac{1}{m^{\delta/2}} \int |\bar{Y}_m|^{2+\delta} d\lambda = 0. \tag{8.22}
\]

If the additive temporal model is approximately scalable to order \( k_0 \geq 2 \), or more generally 8.20 holds, then 8.21 is satisfied.
8.5 LIMITING DISTRIBUTIONS OF HARMONIOUS ASSET MODELS

Proof. This is Lyapunov’s condition restated as in exercise 7.14. But directly:

\[
\int_{|\mathcal{Y}_m|>t\sqrt{m}} \frac{\hat{Y}_m^2}{d\lambda} \leq \int_{|\mathcal{Y}_m|>t\sqrt{m}} \frac{1}{(t\sqrt{m})^\delta} d\lambda \leq \frac{1}{t^\delta m^{\delta/2}} \int |\hat{Y}_m|^{2+\delta} d\lambda,
\]

and thus 8.22 implies 8.19.

Remark 8.30 It should be noted that there is nothing special about the initial collection of variates \((X_1^{(n)}, X_2^{(n)}, ..., X_n^{(n)})\), reflecting asset model prices at times \(j\Delta t\) for \(1 \leq j \leq n\). Arbitrarily choosing \(m_0\), the above results automatically apply to provide convergence results relative to initial prices \((X_1^{(m_0n)}, X_2^{(m_0n)}, ..., X_n^{(m_0n)})\). This follows because the triangular array of temporal variates needed for this result, \(\{Y_i^{(k_m n)}\}_{i=1}^{k_m} \}_{k_m=1}^{\infty}\), is a sub-array of \(\{Y_i^{(mn)}\}_{i=1}^{\infty} \) \(m=1\). Thus if either 8.22 or 8.19 is satisfied for the original array, it will also be satisfied for this sub-array.

Consequently, this proposition’s result applies to the distribution of prices at any time point \(t\) for which \(t = jT/mn\) for integers \(0 < j \leq mn\), which represents a dense set in the interval \([0, T]\).

The next result identifies a class of approximately scalable models for which the conclusion of the above proposition always holds.

Proposition 8.31 Every approximately scalable model parametrized as in proposition 8.15, \(Y_i^{(n)} = a^{(n)}V_i + b^{(n)}\), satisfies the requirement of proposition 8.26, and thus 8.21 is satisfied.

Proof. Recall that if \(V\) is a random variable with two moments and \(E[V] = 0\), the parametrization given by \(\{Y_i^{(n)}\}_{i=1}^{n} = \{a^{(n)}V_i + b^{(n)}\}_{i=1}^{n}\) obtains an approximately scalable model to order \(k_0 = 2\) if as in 8.13:

\[
a^{(mn)} = a^{(n)}/\sqrt{m}, \quad b^{(mn)} = b^{(n)}/m.
\]

Hence in the above notation \(\mu^{(mn)} = b^{(mn)}, \sigma^{(mn)} = a^{(mn)}\sigma_V\) and:

\[
\hat{Y}_m = \frac{V}{\sigma_V}. \tag{(*)}
\]

Thus \(\hat{Y}_m^2\) is integrable by existence of \(\sigma_V^2\), and because \(\hat{Y}_m^2\) is independent of \(m\) we see that 8.19 is always satisfied.
Corollary 8.32 If $V^{2+\delta}$ is integrable for some $\delta > 0$, with $V$ as in proposition 8.31, then 8.21 is satisfied.

Proof. Because $Y_m^{2+\delta}$ is integrable and independent of $m$ by ($\ast$), 8.22 is satisfied and corollary 8.29 applies.

Example 8.33 (Additive Binomial Temporal Model) For models parameterized as in proposition 8.31 and $X_{mj}^{(mn)} = X_0 + \sum_{i=1}^{mj} Y_i^{(mn)}$:

$$\frac{X_{mj}^{(mn)} - (X_0 + jb^{(n)})}{\sqrt{ja^{(n)}\sigma_V}} \Rightarrow Z$$

as $m \to \infty$, or in common notation:

$$X_{mj}^{(mn)} \to N(X_0 + jb^{(n)}, ja^{(n)}\sigma_V^2).$$

An example of such a model is seen in example 8.20, where in the current notation:

$$Y_i^{(n)} \equiv \mu \Delta t + \sigma \sqrt{\Delta t}b_i,$$

where $\{b_i\}_{i=1}^{\infty}$ are independent and binomially distributed to equal $\pm 1$ with probability $1/2$. Hence $E[b_i] = 0$, and with $\Delta t \equiv T/n$ the parameterization for $\{Y_i^{(n)}\}_{i=1}^{n}$ is:

$$a^{(n)} = \sigma \sqrt{\frac{T}{n}}, \quad b^{(n)} = \mu T/n.$$

This approximately scalable additive temporal model with $\sigma_b^2 = 1$ has the property that:

$$X_{mj}^{(mn)} \to N(X_0 + j\mu T/n, j\sigma^2 T/n).$$

This weak convergence of $X_{mj}^{(mn)}$ can also be demonstrated directly with moment generating function techniques.

Multiplicative Temporal Models

We next turn to multiplicative temporal models, the conclusions of which are simplified by the Mann-Wald theorem of proposition 4.21. For the following result, assume that we are given time horizon $T$, spatial $X^{(n)} = (X_1^{(n)}, X_2^{(n)}, ..., X_n^{(n)})$ and multiplicative temporal $Z^{(n)} = (Z_1^{(n)}, Z_2^{(n)}, ..., Z_n^{(n)})$ related by:

$$X_j^{(n)} = X_0 \exp \left( \sum_{i=1}^{j} Z_i^{(n)} \right).$$
8.5 LIMITING DISTRIBUTIONS OF HARMONIOUS ASSET MODELS

Here \( X_0 \) denotes the initial asset value, \( \{Z_i^{(n)}\}_{i=1}^n \) are independent and identically distributed on \((\mathcal{S}, \mathcal{E}, \lambda)\) and with distribution function appropriate for time-step \( \Delta t \equiv T/n \) for some \( n \geq 1 \).

We assume that this temporal model is approximately scalable to order \( k_0 \), so that given \( m \) there are independent, identically distributed \( \{Z_i^{(mn)}\}_{i=1}^{mn} \)
so that \( E \left[ \left( W_j^{(n)} \right)^k \right] = E \left[ \left( W_m^{(mn)} \right)^k \right] \) for \( k \leq k_0 \) and \( 1 \leq j \leq n \) where:

\[
W_j^{(n)} = \sum_{i=1}^{j} Z_i^{(n)}, \quad W_m^{(mn)} = \sum_{i=1}^{mj} Z_i^{(mn)}. 
\]

Then as above we define:

\[
X_j^{(mn)} = X_0 \exp \left( \sum_{i=1}^{mj} Z_i^{(mn)} \right). 
\]

**Proposition 8.34 (Multiplicative Temporal Models)** With the above notation, let \( \bar{Z}_m \equiv (Z^{(mn)} - \mu^{(mn)})/\sigma^{(mn)} \) where \( \mu^{(mn)} \equiv E[Z^{(mn)}] \) and \( (\sigma^{(mn)})^2 \equiv \text{Var}[Z^{(mn)}] \), and assume that for all \( t > 0 \):

\[
\lim_{m \to \infty} \int_{|\bar{Z}_m| > t\sqrt{m}} \bar{Z}_m^2 d\lambda = 0. \quad (8.23)
\]

Assume that either:

1. The multiplicative temporal model is approximately scalable to order \( k_0 \geq 2 \), or more generally,

2. As \( m \to \infty \):

\[
m\mu^{(mn)} \to E[W_1^{(n)}], \quad m\left(\sigma^{(mn)}\right)^2 \to \text{Var}[W_1^{(n)}]. \quad (8.24)
\]

Then as \( m \to \infty \), for all \( j \) with \( 1 \leq j \leq n \):

\[
\frac{\ln \left[ X_{mj}^{(mn)}/X_0 \right] - E[W_j^{(n)}]}{\sqrt{\text{Var}[W_j^{(n)}]}} \Rightarrow Z, \quad (8.25)
\]

with \( Z \) the standard normal variate. In other words,

\[
X_{mj}^{(mn)} \Rightarrow \exp \left( E[W_j^{(n)}] + \ln X_0 + Z\sqrt{\text{Var}[W_j^{(n)}]} \right), \quad (8.26)
\]
250 CHAPTER 8 DISCRETE TIME ASSET MODELS IN FINANCE

and $F_{X^{(mn)}_{m,j}}$ converges in distribution to the lognormal distribution with parameters $\mu = E[W_j^{(n)}] + \ln X_0$ and $\sigma^2 = \text{Var}[W_j^{(n)}]$.

**Proof.** By definition $\ln \left( \frac{X^{(mn)}_{m,j}}{X_0} \right) = \sum_{i=1}^{mj} Z_i^{(mn)}$ and thus by 8.23 and either proposition 7.6 or corollary 7.9, 8.25 follows. The Mann-Wald theorem of proposition 4.21 and $h(x) = \exp \left( x \sqrt{\text{Var}[W_j^{(n)}]} + E[W_j^{(n)}] + \ln X_0 \right)$ then obtains 8.26. ■

**Corollary 8.35 (Multiplicative Temporal Models)** Assume that for some $\delta > 0$ that $\tilde{Z}_m^{2+\delta}$ is integrable and that:

$$
\lim_{m \to \infty} \frac{1}{m^{\delta/2}} \int \tilde{Z}_m^{2+\delta} d\lambda = 0. (8.27)
$$

If the multiplicative temporal model is approximately scalable to order $k_0 \geq 2$, or more generally 8.24 holds, then 8.25 is satisfied.

**Proof.** Left as an exercise using exercise 7.14. ■

We state the example of the following result as a proposition because of the importance of this model in finance and for the next section.

**Proposition 8.36 (Multiplicative Binomial Temporal Model)** Let

$$
X_j^{(n)} = X_0 \exp \left[ \sum_{i=1}^j Z_i^{(n)} \right]
$$

where $Z_i^{(n)} = \mu \Delta t + \sigma \sqrt{\Delta t} b_i$ and $b_i$ are independent and binomially distributed to equal 1 or $-1$ with probability $p = 1/2$. As above $\Delta t \equiv T/n$ for some fixed horizon time $T$, and thus $X_j^{(n)}$ is the price variate at time $t = jT/n$ for any integer $0 < j \leq n$. Then given $m$, define the refined price variate at time $t = jT/n$:

$$
X_{m,j}^{(mn)} = X_0 \exp \left[ \sum_{i=1}^{mj} \left( \frac{\mu T}{mn} + \sigma \sqrt{T/mn} b_i \right) \right].
$$

Then as $m \to \infty$,

$$
\frac{\ln[X_j^{(mn)}/X_0] - \mu t}{\sigma \sqrt{t}} \Rightarrow Z, \quad (8.28)
$$

where $Z$ denotes the standard normal variate.
Equivalently, $X_j^{(mn)}$ converges in distribution to a lognormal distribution with parameters $\mu t + \ln X_0$ and $\sigma^2 t$, and denoting by $X_t$ as this limiting lognormal variate with $t = jT/n$:

$$X_t = X_0 \exp \left[ \mu t + \sigma \sqrt{t} Z \right]. \quad (8.29)$$

**Proof.** The proof is a direct application of proposition 8.34 above since this model is approximately scalable to order $k_0 = 2$ by example 8.20. ■
Chapter 9

Pricing of Financial Derivatives

9.1 Discrete Time Pricing on a Binomial Lattice

The binomial multiplicative temporal model of example 8.22 above is often referred to as the binomial return model of equities or other assets in the real world measure. This means that the implied distribution of asset prices as determined by $\mu$, $\sigma$ and $\Delta t$, and associated probabilities as determined by $p$, represent a model for the actual, observable, future outcomes. Under the assumption that this asset trades without costs or other "frictions" in the market, and that both "long" and "short" positions are possible, it turns out that given any such binomial model, the payoffs at maturity of a European option or other derivative on this asset can be achieved by an initial portfolio of the underlying asset and the risk-free asset. To be successful, one also needs to appropriately rebalance between the asset and the risk-free asset at the end of each time-step. Such an option is then said to be replicated by this initial portfolio, and it then follows by the so-called law of one price that the price of the option must agree with the price of this initial replicating portfolio.

It is really this last conclusion where the mathematics of replication and option pricing becomes the realities of a financial market, and for this conclusion, the critical assumption is that the modeled asset trades freely. The reason for this is that market enforces the law of one price by risk-free arbitrage, whereby a market trader is able to take long and short positions in portfolios with identical payoffs, but different prices. Logically shorting the higher priced portfolio and going long the lower priced, the trader makes
a risk free profit of the price differential since a risk free exit is assured by the assumed equality of the portfolios' payoffs. As financial derivatives and risk free assets trade freely, the critical assumption needed for the enforcement of the law of one price is then the assumption that the asset itself trades freely. In most models one ignores the actual trading costs and other frictions like taxes that exist in the market.

It then turns out that the price of this option or other derivative contract on the underlying asset, which is the price of this replicating portfolio, then equals the risk free present value of the expected derivative payoff. For this calculation however, this expectation is not taken with respect to the original real world probabilities \( p \), but instead with respect to a new set of probabilities which we now investigate.

### 9.1.1 Binomial Lattice Pricing of European Derivatives

Recall that a European derivative security on an asset \( X \) is a financial contract with a specified term \( T \), which provides a payoff at time \( T \) which is a function of the value of \( X \) at that time. European put and call options are specific examples of such derivative securities with payoff functions defined in 9.29 and 9.28 below, which conventionally make nonnegative payments to the buyer of the option who is said to be long the option. However, there is nothing about the mathematics of derivatives pricing below that requires this, and indeed, the pricing approach below applies equally well to forward or other derivative contracts for which the payoff, now better called a "settlement," may be positive or negative. When the payoff is strictly nonnegative, the seller or "writer" of the contract is said to be short the contract, and makes payments at time \( T \) as contractually required. When the settlement can be positive or negative, the long is the party that profits when \( X \) increases, while the short profits when \( X \) decreases.

To simplify language, we will mostly use the term "option" in the development below, but it will be clear from the development that other than the explicit form of the Black-Scholes-Merton option pricing formulas for put and call options, that all of the development below applies equally well to general European derivative securities on \( X \).

As noted above, the law of one price is predicated on the idea that if the prices of assets or portfolios with identical payoffs did not agree, investors could create a risk-free arbitrage by going long the cheaper alternative, and shorting the more expensive alternative. The positions are then settled
9.1 DISCRETE TIME PRICING ON A BINOMIAL LATTICE

at maturity with perfectly offsetting contractual payoffs, and the investor keeps the initial profit without having taken any risk. These buy and sell pressures then adjust prices back to an equilibrium level which ultimately obeys this law. In the real world, "back to equilibrium" need not mean "exactly equal," but only to the point at which the arbitrage is no longer profitable, which is to say that the price discrepancy is within trading costs. So this law is a reasonable assumption in any market that allows both long and short positions in various contracts, and this is typically the case for equities, tradable indexes, currencies and other so-called investment assets, as well as options and futures on such underlying assets.

Exercise 9.1 A risk free asset over a period \([0, \Delta t]\) is defined as a security which can be acquired at \(t = 0\) for \(B(0)\), and the maturity value at time \(t = \Delta t\) is known at that time with certainty to be \(B(\Delta t)\). Define the implied risk-free interest rate in given units by \(B(\Delta t) = B(0)e^{r_c\Delta t}\), or \(B(\Delta t) = B(0)(1 + r_a\Delta t)\), etc. Prove by an arbitrage argument that if risk-free assets can be acquired or shorted without friction costs, that the risk-free rate is uniquely defined for any period.

As noted above, we now show that the price of the initial replicating portfolio equals the present value of the option’s expected payoff, where this expected value is calculated with the given binomial model but with a new probability, whereby the real world \(p\) is replaced by the risk neutral world \(q\).

To this end, given \(T\) and \(n\) with \(\Delta t = T/n\), and \(\mu\) and \(\sigma\), the above multiplicative binomial temporal model in proposition 8.36 produces \(n + 1\) asset prices at time \(T\):

\[
\left\{ X_0e^{ku(\Delta t)}e^{(n-k)d(\Delta t)} \right\}_{k=0}^{n}
\]

with associated probabilities:

\[
\left\{ \binom{n}{k}p^k(1-p)^{n-k} \right\}_{k=0}^{n}
\]

It is common to use such a model to represent the price of one unit of \(X\), so \(X\) could be the price of one share of a common stock, or one unit of a foreign currency, or the futures price on a given futures contract, which in turn typically identifies the futures price of one unit of the contract’s underlying asset.
When \( p = 1/2 \), the "up return" and "down return" are given by:

\[
 u(\Delta t) = \mu \Delta t + \sigma \sqrt{\Delta t}, \quad d(\Delta t) = \mu \Delta t - \sigma \sqrt{\Delta t},
\]

and produce the real world calibrated results for returns:

\[
 E[\ln(X_T/X_0)] = \mu T, \quad Var[\ln(X_T/X_0)] = \sigma^2 T.
\]

The same calibration of moments can be achieved with any other real world \( p \) with \( 0 < p < 1 \) by properly redefining \( u(\Delta t) \) and \( d(\Delta t) \). Details are left as an exercise, or can be found in Reitano (2010) along with other details noted below.

**Exercise 9.2** Given \( p \) with \( 0 < p < 1 \), show that (9.2) is satisfied with:

\[
 u(\Delta t) = \mu \Delta t + a \sigma \sqrt{\Delta t}, \quad d(\Delta t) = \mu \Delta t - a^{-1} \sigma \sqrt{\Delta t},
\]

where \( a = \sqrt{(1-p)/p} \).

Prove by an arbitrage argument that with \( r_c \equiv r_c(\Delta t) \) given in exercise 9.1, that it must be the case in any model that precludes risk-free arbitrage that:

\[
 d(\Delta t) < r_c(\Delta t) < u(\Delta t).
\]

**Hint:** Consider proof by contradiction.

**Notation 9.3** Asset prices often need to be identified by "time" and "state." In this model, asset prices are defined at times \( t = j \Delta t \), for \( j = 1, ..., n \), but then at any time \( j \) there are also \( j + 1 \) possible "states" for this price. A common notational scheme is to denote this time-state price as \( X_{k,j} \), defined by:

\[
 X_{k,j} = X_0 e^{ku(\Delta t)} e^{(j-k)d(\Delta t)},
\]

where \( 0 \leq j \leq n \) and \( 0 \leq k \leq j \). Of course \( X_{0,0} = X_0 \), while the \( j+1 \) state prices at time \( j \) are \( \{X_0 e^{ku(\Delta t)} e^{(j-k)d(\Delta t)}\}_{k=0}^j \).

The notation \( X_{k,j} \) is reminiscent of notation for matrix components under the usual convention of \( X_{\text{row,col}} \), where the rows of the array relate to states, and the columns defines time values. But this is not a rectangular matrix of course, but a triangular matrix. In this notational scheme, each \( X_{k,j} \) is connected by the \( u(\Delta t) \) and \( d(\Delta t) \) return variates to two time \( j+1 \)
9.1 DISCRETE TIME PRICING ON A BINOMIAL LATTICE

prices, schematically:

\[
X_{k+1,j+1} = u(\Delta t) X_{k,j} \quad \text{and} \quad X_{k,j+1} = d(\Delta t) X_{k,j}
\]

\[\text{(9.6)}\]

Thus \(k\) denotes the number of \(u(\Delta t)\) returns over all prior periods.

If the above diagram is completed from \(t = 0\) to \(t = T\) the schematic created is called a **binomial lattice of asset prices**, and provides a handy visual aid in the current development. The above visual representation is compelling since often \(d(\Delta t) < 0\) and \(u(\Delta t) > 0\) in a model, though this is not necessary by 9.4.

In any case it is often easier to program this lattice as an upper or lower triangular matrix. As an upper triangular matrix the top row is \(\{X_{k,k}\}_{k=0}^n \equiv \{X_0 e^{ku(\Delta t)}\}_{k=0}^n\), while in the lower triangular matrix the bottom row is \(\{X_{0,k}\}_{k=0}^n \equiv \{X_0 e^{kd(\Delta t)}\}_{k=0}^n\).

**Proposition 9.4** Let \(O_T [X_0 e^{j(\Delta t)} e^{(n-j)d(\Delta t)}]\) denote the payoff of a European derivative security at time \(T = n\Delta t\) expressed as a function of the then prevailing price of an asset \(X\). Assume that investors can take long or short positions in \(X\) with no bid-ask spread and no costs, that \(X\) pays no income, and the risk-free interest rate \(r\) is fixed over the period.

Then the price at time \(0\) of a replicating portfolio for this derivative security is given by:

\[
O_0^{(n)} [X_0] = e^{-rT} \sum_{j=0}^n \binom{n}{j} q^j (1-q)^{n-j} O_T \left[ X_0 e^{j(\Delta t)} e^{(n-j)d(\Delta t)} \right]. \quad (9.7)
\]

In words, \(O_0^{(n)} [X_0]\) equals the risk free present value of the expected payoffs at time \(T = n\Delta t\), where this expectation reflects the original binomial distribution of returns but with parameters \(q \equiv q(\Delta t)\), the **risk neutral probability**, defined by:

\[
q(\Delta t) = \frac{e^{r\Delta t} - e^{d(\Delta t)}}{e^{u(\Delta t)} - e^{d(\Delta t)}}. \quad (9.8)
\]

**Proof.** The details of the derivation of this option pricing result is an interesting exercise with the following steps.
1. Show that given any of the $j+1$ asset prices $X_{k,j} \in \{X_0 e^{k(j-1)d\Delta t} e^{(j-k)d\Delta t}\}$ at time $j\Delta t$, and given any payoff function $f(X_{k,j+1})$ defined at time $(j+1)\Delta t$, the following holds. There are time-state constants $a_{k,j}$ and $b_{k,j}$, so that an initial portfolio of $a_{k,j}$ units of $X_{k,j}$, and $b_{k,j}$ units invested at rate $r$, exactly replicates the payoff values of $f(X_{k,j+1})$ and $f(X_{k+1,j+1})$ at time $(j+1)\Delta t$. Hint: Look for two equations in two unknowns. Note that it is in this step that the assumption on $X$ paying no income is used. Obtain:

$$a_{k,j} = \frac{f(e^{d\Delta t} X_{k,j}) - f(e^{d\Delta t} X_{k,j})}{(e^{u\Delta t} - e^{d\Delta t}) X_{k,j}}, \quad (9.9)$$

and,

$$b_{k,j} = \left(\frac{e^{u\Delta t} f(e^{d\Delta t} X_{k,j}) - e^{d\Delta t} f(e^{u\Delta t} X_{k,j})}{e^{u\Delta t} - e^{d\Delta t}}\right) e^{-r\Delta t}. \quad (9.10)$$

2. The price of this portfolio at time $j\Delta t$, is notationally $a_{k,j} X_{k,j} + b_{k,j}$. Show that the law of one price applies, and thus $a_{k,j} X_{k,j} + b_{k,j}$ is the market price at time $j\Delta t$ of the time $(j+1)\Delta t$ payoffs of $f(X_{k,j+1})$ and $f(X_{k+1,j+1})$. This requires you to be able to construct risk-free arbitrages if the market price exceeds, or is below this calculated price, which in turn will require the assumption that you can go long or short $X$.

3. The market price at time $j\Delta t$ of these time $(j+1)\Delta t$ payoffs, $a_{k,j} X_{k,j} + b_{k,j}$, can be algebraically manipulated into the form:

$$P_{j\Delta t}(X_{k,j}) = e^{-r\Delta t} \left[ q f(X_{k+1,j+1}) + (1 - q) f(X_{k,j+1}) \right] = e^{-r\Delta t} \left[ q f(e^{u\Delta t} X_{k,j}) + (1 - q) f(e^{d\Delta t} X_{k,j}) \right],$$

with $q \equiv q(\Delta t)$ defined in 9.8. By 1 and the law of one price, these market prices unambiguously equal the market values of this option at that time, and thus we label these prices as $O_{j\Delta t}[X_{k,j}].$

4. The formula in 3 can be applied iteratively starting at time $T = n\Delta t$ with the given European payoff function $O_T[X_{k,n}]$, deriving the market prices $O_{T-n\Delta t}[X_{k,n-1}]$ for all $k$. Then by induction it can be shown that 9.7 is satisfied in the general form:

$$O_{(n)\Delta t}[X_{k,n}] = e^{-r(n-j)\Delta t} \sum_{i=0}^{n-j} \binom{n-j}{i} q^i (1-q)^{n-j-i} O_T \left[ X_{k,j} e^{iu\Delta t} e^{(n-j-i)d\Delta t} \right].$$
9.1 DISCRETE TIME PRICING ON A BINOMIAL LATTICE

Here, \( O^{(n)}_{j\Delta t} \) denotes the market price for this option at time \( t = j\Delta t \), and not a payoff value since a European option makes no payoff until time \( t = T \).

5. Observe that this general formula produces 9.7 when \( j = 0 \).

Remark 9.5 For general \( p \neq 1/2 \), the risk neutral probability \( q \equiv q(\Delta t) \) is still given by 9.8 even though the definitions of \( u(\Delta t) \) and \( d(\Delta t) \) must reflect \( p \) as noted in exercise 9.2 above.

Exercise 9.6 1. Assume that \( X \) is the price of one unit of foreign currency, and that 9.6 is used to model this price. Repeat the steps in the above proof to show that 9.7 is still valid, but now:

\[
q^{\text{Curr}}(\Delta t) = \frac{e^{(r-r_f)\Delta t} - e^{d(\Delta t)}}{e^{u(\Delta t)} - e^{d(\Delta t)}},
\]

(9.11)

where \( r_f \) denotes the assumed constant risk-free interest rate paid on the foreign currency. Hint: Think about how step 1 above changes if you hold foreign currency in the replicating portfolio, since you would presumably invest it.

Note: This approach is also used to approximate the dividends paid by the stocks in an equity index by a continuous dividend rate of \( \delta \) say, assuming that dividends are paid frequently and homogeneously enough to justify such an approximation. All results below then apply with \( \delta \) replacing \( r_f \).

2. Assume that \( X \) is a futures price, and that 9.6 is used to model this price. Repeat the steps in the above proof to show that 9.7 is still valid, but now:

\[
q^{\text{Fut}}(\Delta t) = \frac{1 - e^{d(\Delta t)}}{e^{u(\Delta t)} - e^{d(\Delta t)}},
\]

(9.12)

Hint: Think about how step 1 above changes if you hold futures contracts in the replicating portfolio. What is the value of this position if the futures price changes from \( X_{k,j} \) to \( X_{k+1,j+1} \), say.

3. Assume that \( X \) is the price of a commodity, and that 9.6 is used to model this price. Verify that the above derivation fails because:
(a) Step 1 works mathematically, but holding commodities long often requires payment of storage costs. Thus if \( a_{k,j} > 0 \), this problem is equivalent to negative income and can be accommodated as for currency but with negative \( r_c \).

(b) If the calculated \( a_{k,j} < 0 \) in step 1, this requires a short position in \( X \) which is generally not possible (gold is the exception). Thus step 1 can be pushed through for derivatives with \( a_{k,j} > 0 \), meaning derivatives with payoffs that are positively correlated with asset prices (calls for example).

(c) Step 2 always fails, since the law of one price requires that you can take long or short positions in \( X \), and shorts are generally not possible. For derivatives with \( a_{k,j} > 0 \), step 2 will provide only that \( a_{k,j}X_{k,j} + b_{k,j} \) is an upper bound to price of the end of period payoffs.

Note: One exception to the conclusion in 3 is for gold, which though a commodity, has the somewhat unique distinction of being a commodity that investors often hold in inventories for its appreciation potential. Thus it is possible to short gold, and the original set-up is valid with the small adjustment that there is a cost to borrowing gold, called the "lease rate," which in effect one earns if long gold, and pays if short, much like a currency.

**Notation 9.7** The European derivatives price in 9.7 can also be written in the suggestive notation:

\[
O_0^{(n)}[X_0] = E_q \left[ e^{-rT}O_T[X_T] \right], \tag{9.13}
\]

where \( E_q \) denotes the expectation of the option payoffs, \( O_T[X_T] \), relative to the risk neutral probabilities. Specifically:

\[
Pr_q[X_{k,n}] = \binom{n}{k} q^k (1 - q)^{n-k}.
\]

where \( n \Delta t = T \) and \( q(\Delta t) \) is defined in 9.8.

This notation is suggestive because it raises one’s expectation that the limiting price of this option as \( n \to \infty \) will be given by the same formula, but where the \( E_q \)-expectation is taken with respect to the limiting distribution of these binomials as \( n \to \infty \). Indeed, we will see below that this expectation is confirmed.
9.1 DISCRETE TIME PRICING ON A BINOMIAL LATTICE

Example 9.8 As a very simple example, let the European derivative security be a long forward contract on $X$ with price $F$. Then $O_T[X_T] = X - F$, and thus:

$$F_0^{(n)}[X_0] = e^{-rT} \sum_{j=0}^{n} \binom{n}{j} q^j (1 - q)^{n-j} \left[ X_0 e^{ju(\Delta t)} e^{(n-j)d(\Delta t)} - F \right]$$

$$= e^{-rT} \left[ X_0 \sum_{j=0}^{n} \binom{n}{j} (qe^{u(\Delta t)})^j [(1 - q)e^{d(\Delta t)}]^{n-j} - F \right],$$

since $\sum_{j=0}^{n} \binom{n}{j} q^j (1 - q)^{n-j} = 1$. Now since $qe^{u(\Delta t)} + (1 - q)e^{d(\Delta t)} = e^{r\Delta t}$, one obtains:

$$\sum_{j=0}^{n} \binom{n}{j} (qe^{u(\Delta t)})^j [(1 - q)e^{d(\Delta t)}]^{n-j} = [qe^{u(\Delta t)} + (1 - q)e^{d(\Delta t)}]^n = e^{rn\Delta t},$$

and thus as expected:

$$F_0^{(n)}[X_0] = X_0 - Fe^{-rT}. $$

The probabilities $q$ in 9.8, 9.11 or 9.12 are called **risk neutral probabilities** for the respective securities because defining $O_T(x) = x$:

$$X_0 = e^{-rT} \sum_{j=0}^{n} \binom{n}{j} q^j (1 - q)^{n-j} \left[ X_0 e^{ju(\Delta t)} e^{(n-j)d(\Delta t)} \right]. \quad (9.14)$$

It is an exercise to verify this identity algebraically. Stated another way:

$$X_0 = E_q \left[ e^{-rT} X_T \right]. \quad (9.15)$$

Thus, the current price of the asset $X_0$, is the risk-free present value of its expected price at time $T$, where this expectation is calculated using $q$. In a risk neutral world, which is characterized by a linear utility function (see Reitano (2010), section 9.8.8), all risky assets are priced in terms of the risk-free present value of expected payoffs. Thus $q$ is the unique probability that would apply in a risk neutral world in order to justify the current price of $X_0$. The real world is not risk neutral certainly, but it turns out that as a corollary to the replicating portfolio construction that many financial derivatives can be priced as if we lived in a risk neutral world, because many such derivatives can be replicated. If a derivative can be replicated, then by the law of one price the derivative’s price cannot depend on risk preferences of investors.
CHAPTER 9 PRICING OF FINANCIAL DERIVATIVES

This same formula prices the risk free security since this identity is true for any $q$. In this case, with an initial investment of $B_0$ we have that $X_{k,n} = B_0 e^{rT}$ for all $k$, and so:

$$B_0 = e^{-rT} \sum_{j=0}^{n} \binom{n}{j} q^j (1-q)^{n-j} [B_0 e^{rT}].$$

Algebraically the right hand side collapses for any value of $q$, not just the risk neutral probabilities above.

Remark 9.9 It is tempting to think that the risk neutral pricing approach above is an inessential convenience. Indeed, within the context of the real world binomial model it is in theory also possible to represent $X_0$ in terms of the real world probability by:

$$X_0 = E_p [e^{-r_{RW}T} X_T],$$

where $r_{RW}$ is now the return demanded by investors to reflect risk. Since $p = 1/2$ for the given definitions of $u(\Delta t)$ and $d(\Delta t)$, a calculation produces:

$$X_0 = e^{-r_{RW}T} \left( \left[ e^{u(\Delta t)} + e^{d(\Delta t)} \right] / 2 \right)^n X_0,$$

and thus $r_{RW}$ can be mechanically calculated from this formula. But this is then a dead end.

While such $r_{RW}$ may be the appropriate rate to discount for the risk of $X_T$ in the real world, the risk of $O_T [X_T]$ will in general be quite different, and indeed may be negatively correlated with the risk of $X_T$. The significance of this is that whatever the appropriate rate $r'_{RW}$ would be to assure that:

$$O_0^{(n)} [X_0] = E_p \left[ e^{-r'_{RW}T} O_T [X_T] \right],$$

this rate is unknowable, will in theory differ for every derivative, and need not even be positive.

Thus the power of risk neutral probabilities is found in the simple fact that they allow expected present value pricing of derivatives with risk-free rates, which are unambiguously observable.

9.1.2 Binomial Lattice Pricing of American Derivatives

An American option or other derivative provides for a formulaic payment at time $T$ as does a European option, but also provides the long position an early exercise option, meaning that at times $0 < t < T$, and
sometimes $0 < t_0 \leq t < T$, the long can elect to exercise the option and demand payment based on a payoff function appropriate at that time. For traditional American puts and calls the payoff function is typically the same at all times $t \leq T$, reflecting the asset price at that time, but nothing in the mathematics below requires this.

It is apparent that the market price for an American option must equal or exceed its then current exercise value. This observation reflects the assumption made in derivatives pricing that the writer of the option, the short, exposes the long to no credit risk, and all payments are certain to be made. This assumption is realized for options traded on exchanges due to the margin account mechanism, but also in the over-the-counter options markets when traders use similar collateral structures.

Hence, the pricing of European options can be adapted to American options by appropriately adjusting the one period iterative formulas to:

$$
O^{(n)}_{j+1} [X_{k,j}] = \max \left( e^{-r\Delta t} \left[ qO^{(n)}_{j+1} \Delta t(X_{k+1,j+1}) + (1-q)O^{(n)}_{j+1} \Delta t(X_{k+1,j}) \right], P_{j+1} \Delta t [X_{k,j}] \right).
$$

In other words, the market price of the option is the greater of the current exercise value, $P_{j+1} \Delta t [X_{k,j}]$, and the value of all future exercise opportunities as reflected in the expected present value. In contrast to the European option formula, this pricing formula is necessarily iterative and does not reduce to a simple formula at $t = 0$ as does the European price in 9.7.

It is an exercise to confirm that the pricing approach in 9.16 is consistent with the replicating portfolio approach initiated with European options above. Specifically, one first recalls that

$$
e^{-r\Delta t} \left[ qO^{(n)}_{j+1} \Delta t(X_{k+1,j+1}) + (1-q)O^{(n)}_{j+1} \Delta t(X_{k+1,j}) \right]
$$

equals the cost of a portfolio of $X$-assets and risk free assets that has value $O^{(n)}_{j+1} \Delta t(X_{k+1,j+1})$ when $X_{k+1,j+1}$ prevails, and value $O^{(n)}_{j+1} \Delta t(X_{k+1,j})$ when $X_{k+1,j}$ prevails. Thus if this value exceeds the current exercise payoff $P_{j+1} \Delta t [X_{k,j}]$, the price of the American option in this time-state equals that of the portfolio that replicates the next period’s values. If $P_{j+1} \Delta t [X_{k,j}]$ exceeds this calculated value, then the value of the American option exceeds that needed for further replication of values, but this is not a problem because then the option is exercised and there are no further values to replicate.

In summary, American option pricing provides a portfolio at time 0 that can be rebalanced to the optimal exercise date. This date might be that...
of an early exercise or at final maturity. That is, this replicating portfolio replicates the values of this option throughout its existence.

9.2 Limiting Asset Distribution: Risk Neutral Measure

Given the above derivative pricing results, a logical next question now becomes: What is the limiting distribution of \( X_T \), or \( X_T^{(n)} \) in the notation of the prior chapter, as \( n \to \infty \)? We cannot simply apply the result of the previous chapter on multiplicative binomial temporal models because here the binomial probability is a function of \( \Delta t \), \( q \equiv q(\Delta t) \), or equivalently, \( q \) is a function of \( n \) since \( \Delta t = T/n \). However, an immediate corollary of the next result is that

\[
\lim_{\Delta t \to 0} q(\Delta t) = \frac{1}{2},
\]

and this generalizes to \( \lim_{\Delta t \to 0} q(\Delta t) = p \) when \( u(\Delta t) \) and \( d(\Delta t) \) are redefined to reflect \( p \neq 1/2 \). So it is natural to expect that the limiting distribution under the risk neutral \( q \) probabilities agrees with the limiting distribution under the real world \( p \) probabilities, but this expectation is soon to be dashed. It turns out that \( q(\Delta t) \) approaches \( 1/2 \) (or \( p \)) relatively slowly, and indeed, slowly enough to change the mean of the limiting distribution, but not the variance.

9.2.1 Analysis of Risk Neutral Binomial Probability \( q \)

To investigate the limiting distribution of assets under \( q(\Delta t) \) will require a more informative representation of this probability as a function of \( \Delta t \). To this end we prove the following, and assign a generalization for \( p = 1/2 \) as an exercise. Alternatively, see proposition 9.155 in Reitano (2010).

**Remark 9.10** Recall that the big-O error term in 9.17, \( O(\Delta t^{3/2}) \), means that as \( \Delta t \to 0 \):

\[
\left( q(\Delta t) - \frac{1}{2} - \frac{r - \mu - \sigma^2/2}{2\sigma} \sqrt{\Delta t} \right) / \Delta t^{3/2} \to C < \infty.
\]

**Proposition 9.11** With \( q \equiv q(\Delta t) \) as in 9.8, and \( u(\Delta t) \) and \( d(\Delta t) \) as in 9.1:

\[
q(\Delta t) = \frac{1}{2} + \frac{r - \mu - \sigma^2/2}{2\sigma} \sqrt{\Delta t} + O(\Delta t^{3/2}).
\] (9.17)
9.2 LIMITING ASSET DISTRIBUTION: RISK NEUTRAL MEASURE

Proof. By a Taylor series analysis:

\[
q(\Delta t) = \frac{\exp \left[ \sigma \sqrt{\Delta t} + (r - \mu) \Delta t \right] - 1}{\exp \left[ 2\sigma \sqrt{\Delta t} \right] - 1} = \frac{\sigma \sqrt{\Delta t} + \left[ (r - \mu) + \frac{\sigma^2}{2} \right] \Delta t + O(\Delta t^{3/2})}{2\sigma \sqrt{\Delta t} + 2\sigma^2 \Delta t + O(\Delta t^{3/2})}.
\]

The result in 9.17 is derived by long division, a derivation made easier by the notational substitution: \( \sqrt{\Delta t} \to x \). One the other hand, verification of 9.17 can now be made by polynomial multiplication.

Exercise 9.12 Verify that with \( q \) as in 9.11 or 9.12, that without any more work than redefining \( r \) in 9.8 that:

\[
q_{\text{curr}}(\Delta t) = \frac{1}{2} + \frac{r - r_f - \mu - \frac{\sigma^2}{2}}{2\sigma} \sqrt{\Delta t} + O(\Delta t^{3/2}),
\]

\[
q_{\text{fut}}(\Delta t) = \frac{1}{2} + \frac{-\mu - \frac{\sigma^2}{2}}{2\sigma} \sqrt{\Delta t} + O(\Delta t^{3/2}).
\]

The most important consequence of the slow convergence of \( q(\Delta t) \) to \( 1/2 \), now apparently caused by the \( \sqrt{\Delta t} \) term in 9.17, is that the mean of asset prices is shifted from \( \mu t \) to \( (r - \sigma^2/2)t + O(\Delta t^{1/2}) \), while the effect on variance is minimal.

Proposition 9.13 Given \( T \) and \( n \) with \( \Delta t = T/n \), if \( X_t \) denotes the random variable of asset prices at time \( t = j\Delta t \), then under the \( q \)-probabilities:

\[
E_q [\ln[X_t/X_0]] = (R(r) - \sigma^2/2)t + O(\Delta t^{1/2}),
\]

\[
Var_q [\ln[X_t/X_0]] = \sigma^2 t + O(\Delta t^{1/2}),
\]

where \( R(r) = r \), \( R(r) = r - r_f \) or \( R(r) = 0 \) in the respective definitions of \( q \) in 9.8, 9.11 or 9.12.

Proof. Note that \( X_t/X_0 = \prod_{k=1}^{j} X_{k\Delta t}/X_{(k-1)\Delta t} \), so we derive the result for \( \ln \left[ X_{k\Delta t}/X_{(k-1)\Delta t} \right] \). To this end,

\[
\ln \left[ X_{k\Delta t}/X_{(k-1)\Delta t} \right] = \begin{cases} 
\mu \Delta t + \sigma \sqrt{\Delta t}, & \text{Pr} = q, \\
\mu \Delta t - \sigma \sqrt{\Delta t}, & \text{Pr} = 1 - q.
\end{cases}
\]

With some algebra applied with the expansion of \( q(\Delta t) \) above:

\[
E_q [\ln[X_{k\Delta t}/X_{(k-1)\Delta t}]] = (R(r) - \sigma^2/2)\Delta t + O(\Delta t^{3/2}),
\]

\[
Var_q [\ln[X_{k\Delta t}/X_{(k-1)\Delta t}]] = \sigma^2 \Delta t + O(\Delta t^{3/2}).
\]
Now note that \( \{X_{k\Delta t}/X_{(k-1)\Delta t}\}_{k=1}^j \) is a collection of independent random variables, since this equals \( \{e^{u\Delta t+\sigma\sqrt{\Delta t}b_k}\}_{k=1}^j \) where \( b_k = 1 \) for \( u(\Delta t) \) and \( b_k = -1 \) for \( d(\Delta t) \). Since \( \{b_k\}_{k=1}^j \) are independent by assumption, so too are these exponentiated expressions by proposition 3.56 of book 2. The result in 9.18 then follows by addition, noting that \( nO(\Delta t^{3/2}) \) is at worst \( nO(\Delta t^{3/2}) = O(\Delta t^{1/2}) \).

**Remark 9.14** The expressions in 9.18 make clear that even without deriving the limiting distribution of \( X_t \) under the risk neutral measure, that the mean and variance of \( \ln[X_t/X_0] \) must converge to \( (R(t) - \sigma^2/2)t \) and \( \sigma^2t \), respectively.

### 9.2.2 Limiting Asset Distribution Under \( q \)

Based on 9.17, 9.18 and proposition 8.36, it would be a reasonable to expect that the limiting distribution of \( X_t/X_0 \) is lognormal with parameters \( E_q[\ln[X_t/X_0]] = (R(t) - \sigma^2/2)t \) and \( Var_q[\ln[X_t/X_0]] = \sigma^2t \), or equivalently, that \( \ln[X_t/X_0] \) is normally distributed with mean \( (R(t) - \sigma^2/2)t \) and variance \( \sigma^2t \). But note that while this asset model is approximately scalable to order \( k_0 = 2 \) in the \( p \) probabilities, is not approximately scalable under \( q \) because by 9.18 the first two moments depend on \( \Delta t \). But a result is still derivable with the more general statement in proposition 8.34 because 9.18 implies that as \( n \to \infty \):

\[
nE_q[\ln[X_{k\Delta t}/X_{(k-1)\Delta t}]] \to (R(t) - \sigma^2/2)T, \quad nVar_q[\ln[X_{k\Delta t}/X_{(k-1)\Delta t}]] \to \sigma^2T.
\]

**Remark 9.15** Perhaps the most compelling and surprising implication of the following result is that the limiting distribution of \( X_t \) is independent of \( \mu \) in the definition of \( u(\Delta t) \) and \( d(\Delta t) \) in 9.1. The same result occurs with the more general model of 9.3 as seen in proposition 9.139 of Reitano (2010). Thus in applications, nothing is lost for the pricing of financial derivatives if one simply assumes \( \mu = 0 \) in the definition of \( u(\Delta t) \) and \( d(\Delta t) \), despite the fact that this would then not yield a realistic model for asset prices based on 9.2.

This is a common practice in derivative pricing, however, because we can change the real world \( p \) probabilities to fix the real world results. Specifically, if one sets \( \mu = 0 \) in the definition of \( u(\Delta t) \) and \( d(\Delta t) \) in 9.1 and redefines the real world probability of \( u(\Delta t) \) by:

\[
p' = \frac{1}{2}\left(1 + \frac{\mu}{\sigma\sqrt{\Delta t}}\right),
\]
9.2 LIMITING ASSET DISTRIBUTION: RISK NEUTRAL MEASURE

An appropriate real world model is produced. A calculation produces:

\[ E_p \ln[X_T/X_0] = \mu T, \quad \text{Var}_p \ln[X_T/X_0] = \sigma^2 T + O(\Delta t), \quad (9.19) \]

which is comparable with the real world results in 9.2.

For generality, we do not make this assumption on \( \mu \) in the following.

**Proposition 9.16** Let \( X_j^{(n)} = X_0 \exp \left[ \sum_{i=1}^{j} Z_i^{(n)} \right] \), where \( Z_i^{(n)} = \mu \Delta t + \sigma \sqrt{\Delta t} b_i^{(n)} \), with \( b_i^{(n)} \) independent and binomially distributed to equal 1 or \(-1\) with respective probabilities \( q_n \) and \( 1 - q_n \). Here \( q_n \) is defined by 9.8, 9.11 or 9.12, using \( t = T/n \) for some fixed horizon time \( T \).

Thus \( X_j^{(n)} \) is the price variate at time \( t = jT/n \) for integer \( 0 < j \leq n \). Given \( m \), define the refined price variate at time \( t = jT/n \):

\[
X_{mj}^{(mn)} = X_0 \exp \left[ \sum_{i=1}^{mj} \left( \mu (T/mn) + \sigma \sqrt{T/mn} b_i^{(mn)} \right) \right],
\]

where \( b_i^{(mn)} \) is defined as above relative to \( q_{mn} \) and \( \Delta t_m \equiv T/mn \).

Then for all \( t \equiv jT/n \), as \( m \to \infty \):

\[
\frac{\ln[X_{mj}^{(mn)}/X_0] - (R(r) - \sigma^2/2)t}{\sigma \sqrt{t}} \Rightarrow Z, \quad (9.20)
\]

where \( Z \) denotes the standard normal variate, and \( R(r) = r, R(r) = r - r_f \) or \( R(r) = 0 \) in the respective definitions of \( q \) in 9.8, 9.11 or 9.12.

Equivalently, \( X_{mj}^{(mn)} \) converges in distribution to the lognormal distribution with parameters \( \ln X_0 + (R(r) - \sigma^2/2)t \) and \( \sigma^2 t \). That is, denoting this limiting lognormal variate by \( X_t \) for \( t \equiv jT/n \):

\[
X_{mj}^{(mn)} \Rightarrow X_t \equiv X_0 \exp [Z_t], \quad (9.21)
\]

where \( Z_t \) has a normal distribution with mean \( (R(r) - \sigma^2/2)t \) and variance \( \sigma^2 t \).

**Proof.** For given \( t = jT/n \), define a triangular array indexed by \( m \) by

\[
\{Z_{m,i}\}_{i=1}^{mj} \equiv \{\mu \Delta t_m + \sigma \sqrt{\Delta t_m} b_i^{(mn)}\}_{i=1}^{mj},
\]

where as above, \( \Delta t_m \equiv T/mn \). Then \( \ln \left[ X_{mj}^{(mn)}/X_0 \right] = \sum_{i=1}^{mj} Z_{m,i} \), and the mean and variance of \( Z_m \), denoted \( \mu_m \) and \( \sigma_m^2 \) respectively, are given in the proof of proposition 9.13:

\[
\mu_m = (R(r) - \sigma^2/2) \Delta t_m + O(\Delta t_m^3/2), \quad \sigma_m^2 = \sigma^2 \Delta t_m + O(\Delta t_m^3/2). \quad (*)
\]
Thus defining $\tilde{Z}_m \equiv (Z_m - \mu_m)/\sigma_m$:

$$
\tilde{Z}_m = \left( \frac{\mu - (R(r) - \sigma^2/2)) \Delta t_m + \sigma \sqrt{\Delta t_m \beta_i^{(mn)}} - O(\Delta t_m^{3/2})}{\sqrt{\sigma^2 \Delta t_m + O(\Delta t_m^{3/2})}} \right)
$$

$$
= \frac{(\mu - (R(r) - \sigma^2/2)) \sqrt{\Delta t_m} + \sigma \beta_i^{(mn)} - O(\Delta t_m)}{\sigma \sqrt{1 + O(\Delta t_m^{1/2})}}.
$$

To apply proposition 8.34, we must show that for all $t > 0$:

$$
\lim_{m \to \infty} \int_{|\tilde{Z}_m| > t \sqrt{m}} \tilde{Z}_m^2 d\lambda = 0.
$$

But this follows because as $m \to \infty$:

$$
|\tilde{Z}_m| \to 1,
$$

and thus $|\tilde{Z}_m|$ is bounded for $m$ large and 8.23 is satisfied.

Now by (*), as $m \to \infty$:

$$
m\mu_m = (R(r) - \sigma^2/2) \Delta t + O(\Delta t \Delta t_m^{1/2}) \to (R(r) - \sigma^2/2) \Delta t = \mu_1,$$

$$
m\sigma^2_m = \sigma^2 \Delta t + O(\Delta t \Delta t_m^{1/2}) \to \sigma^2 \Delta t = \sigma_1^2,
$$

and thus 8.24 is satisfied. The result in 9.20 now follows by proposition 8.34, noting that $t$ in 9.20 is given by $t = j\Delta t$.

### 9.3 Limiting Price of European Derivatives

We now have all the ingredients for a final limiting formula for the price of a European financial derivative defined relative to an underlying security which is actively traded. To utilize the above results, we assume that this security can be traded both long and short and with no costs, and that the future price of this security at any time $T$ has a lognormal distribution with parameters $(r - \sigma^2/2)T + \ln X_0$ and $\sigma^2 T$, and thus by proposition 9.16 can be approximated by an appropriately parametrized binomial price distribution.

In detail, given exercise date $T$ and $n$, define $\Delta t \equiv T/n$ and binomial returns as in 9.1. This trading assumption allowed the construction of a replicating portfolio which contains the underlying security and a risk free
security, and has the following replication property. This portfolio can be rebalanced between security positions at the end of each \( \Delta t \)-period in a self-financing way, meaning without changing the value of the portfolio, so that at time \( T \), the portfolio replicates the financial derivative’s payoff in any of the \( n + 1 \) possible price states. Assuming that the market does not allow risk-free arbitrage, the law of one price then assures that the price of this financial derivative at time 0 must agree with the price of this replicating portfolio.

The price of this portfolio can then be formulaically expressed in terms of the financial derivative’s payoffs in these \( n + 1 \) states, and indeed, this price equals the present value of the expected payoff as seen in 9.7. This expectation was calculated based on the originally assumed binomial distribution of prices at time \( T \), but with a new "risk neutral" binomial probability assumption denoted \( q \equiv q(\Delta t) \), in contrast to the originally assumed "real world" binomial probability of \( p \). The formula for the price of the derivative as given in 9.13 notationally links price to an expected present value under \( q \), and suppresses the binomial distribution assumption. Now 9.13 is valid for every \( n \), and by proposition 9.16 above we have that the binomial distribution of underlying asset prices converges weakly to a lognormal distribution as \( n \to \infty \). Thus it seems natural to expect that the limiting price of such financial derivatives again satisfies this expectations formula, but with the limiting lognormal distribution used for the expected value in place of the binomial.

We now prove this result. The essence of the proof is as follows. With \( X_T^{(n)} \) denoting the time \( T \) binomial model prices given \( n \), we have by proposition 9.16 that \( X_T^{(n)} \Rightarrow X_T \), or in words, that \( X_T^{(n)} \) converges weakly as \( n \to \infty \) to the lognormal variate \( X_T \) with parameters \( \ln X_0 + (R(r) - \sigma^2/2)t \) and \( \sigma^2 t \). Now \( O_T \left[ X_T^{(n)} \right] \) is a random variable for any Borel measurable payoff function \( O_T \), and thus we must prove weak convergence, \( O_T \left[ X_T^{(n)} \right] \Rightarrow O_T \left[ X_T \right] \), as well as convergence of the expected values of these variates, \( E \left[ O_T \left[ X_T^{(n)} \right] \right] \to E \left[ O_T \left[ X_T \right] \right] \). Weak convergence of \( O_T \left[ X_T^{(n)} \right] \) follows readily from prior results, but to prove convergence of expectations will require an additional assumption.

**Remark 9.17** 1. Letting \( j = n \) and thus \( t = T \), 9.21 states that \( X_T^{(n)} \Rightarrow X_T \equiv X_0 \exp \left[ Z_T \right] \) as \( n \to \infty \), simplifying notation. If \( F_{(n)} \) and \( F_L \) are the associated binomial and lognormal distribution functions, then by definitions 5.19 and 8.2 of book 2 (see also definition 4.6 above),
this can be equivalently stated \( F(n) \Rightarrow F_L \), or in terms of the induced probability measures, \( \mu^{(n)} \Rightarrow \mu_L \). By part 2 of proposition 4.30 of book 3, \( O_0^{(n)} [X_0] \) in 9.7 can be expressed as a Riemann-Stieltjes integral for continuous \( O_T(x) \):

\[
O_0^{(n)} [X_0] = e^{-rT} \int O_T(x) dF(n),
\]

while by proposition 2.56 of book 5 this Riemann-Stieltjes integral is equal to the associated Lebesgue-Stieltjes integral, and thus:

\[
O_0^{(n)} [X_0] = e^{-rT} \int O_T(x) d\mu^{(n)}.
\]

The portmanteau theorem of proposition 4.4 now assures that for every bounded, continuous \( O_T(x) \), that:

\[
e^{-rT} \int O_T(x) d\mu^{(n)} \to e^{-rT} \int O_T(x) d\mu_L.
\]

This latter integral is again equal to the Riemann-Stieltjes integral with respect to \( F \), and then by the change of variables of proposition 3.6 of book 5:

\[
e^{-rT} \int O_T(x) d\mu = e^{-rT} \int O_T(x) f_L(x) dx,
\]

where \( f_L(x) \) is the lognormal density function and this is now a Riemann integral. A final change of variables allows a restatement of this Riemann integral in terms of the normal density \( f_N \), and combining the above obtains:

\[
O_0^{(n)} [X_0] \to e^{-rT} \int_{-\infty}^{\infty} O_T(X_0e^x) f_N(x) dx.
\]

In summary, the results of the following proposition in 9.24 are derivable with only the portmanteau theorem and a change of variables if restricted to bounded continuous \( O_T(x) \). The need for a proof below is to generalize this result to Borel measurable payoff functions with finite second moments.

2. A consequence of this generalization to Borel measurable payoff functions is that the integral in 9.24 must be understood as a Lebesgue integral. By proposition 2.64 of book 3, if \( O_T(y) \) is absolutely Riemann integrable, then the integral in 9.24 can also be evaluated as a
9.3 LIMITING PRICE OF EUROPEAN DERIVATIVES

Riemann integral. Finally, by proposition 10.22 of Reitano (2010), \( \text{O}_T(y) \) will be absolutely Riemann integrable on any bounded interval \([a,b]\) if \( |\text{O}_T(y)| \) is continuous almost everywhere. Thus subject to the existence of the improper integral, 9.24 can also be interpreted as a Riemann integral for payoff functions that are continuous almost everywhere. Note: Recalling remark 1.1 of book 3, Proposition 10.22 is the Lebesgue existence theorem for the Riemann integral, due to Henri Lebesgue (1875 – 1941), and proved sometime after Riemann’s death in 1866.

**Proposition 9.18 (General Black-Scholes-Merton)** With \( T > 0 \) fixed and \( \Delta t \equiv T/n \) let \( X_j^{(n)} \equiv X_0 \exp \left( \sum_{i=1}^n \left( \mu \Delta t + \sigma \sqrt{\Delta t} b_i \right) \right) \) for \( 0 < j \leq n \), where \( b_i \) are independent and binomially distributed to equal 1 or \(-1\) with respective probabilities \( q \) and \( 1 - q \), where \( q \) is defined in 9.8, 9.11 or 9.12, and 9.1. Let \( O_0^{(n)} [X_0] \) be given as in 9.7, representing the price of a European financial derivative on \( X \) with Borel measurable payoff function \( \text{O}_T \left[ X_T^{(n)} \right] \) at time \( T \), where \( X_T^{(n)} \equiv X_n^{(n)} \). Assume that the discontinuity set of \( \text{O}_T \) has Lebesgue measure zero, and that:

\[
\sup_n E_q \left( \left( \text{O}_T \left[ X_T^{(n)} \right] \right)^2 \right) < \infty, \quad (9.22)
\]

where this expectation is defined relative to the binomial distribution of \( X_T^{(n)} \).

Then as \( n \to \infty \):

\[
O_0^{(n)} [X_0] \to O_0 [X_0], \quad (9.23)
\]

where letting \( X_T \) denote the limit variate of proposition 9.16:

\[
O_0 [X_0] \equiv e^{-rT} E \left( \text{O}_T \left( X_T \right) \right).\]

Thus:

\[
O_0 [X_0] = e^{-rT} \int_0^\infty \text{O}_T (y) f_L(y) dy \quad (9.24)
\]

where \( f_L(y) \) denotes the lognormal density function with parameters \( \ln X_0 + (R(r) - \sigma^2/2)T \) and \( \sigma^2 T \), and the integral in 9.24 is understood as a Lebesgue integral. Equivalently:

\[
O_0 [X_0] \equiv e^{-rT} \int_{-\infty}^{\infty} \text{O}_T \left( X_0 e^x \right) f_N(x) dx, \quad (9.25)
\]

where \( f_N(x) \) is the density function of the normal distribution with mean \((R(r) - \sigma^2/2)T\) and variance \( \sigma^2 T \).
CHAPTER 9 PRICING OF FINANCIAL DERIVATIVES

As above, \( R(r) = r, R(r) = r - r_f \) or \( R(r) = 0 \) in the respective definitions of \( q \) in 9.8, 9.11 or 9.12.

**Proof.** Recall that \( O_0^{(n)} [X_0] \equiv e^{-rT}E_q \left( O_T \left[ X_T^{(n)} \right] \right) \) by 9.13, with expectation defined relative to the binomial distribution of \( X_T^{(n)} \equiv X_n^{(n)} \), and thus 9.23 is equivalent to

\[
E_q \left( O_T \left[ X_T^{(n)} \right] \right) = E \left( O_T \left( X_T \right) \right).
\]

In order to utilize earlier results we should note that it can be assumed that the random variables \( \{X_T^{(n)}, X_T\} \) are defined on some common probability space \((S, \mathcal{E}, \lambda)\). This follows from Skorokhod’s representation theorem of proposition 8.30 of book 2. Specifically, as noted in remark 9.17 \( X_T^{(n)} \Rightarrow X_T \) implies that for the associated probability measures, \( \mu^{(n)} \Rightarrow \mu_L \). Here \( \mu^{(n)} \) is the binomial measure defined above, and \( \mu_L \) the lognormal measure parametrized as in proposition 9.16 above. By Skorokhod’s theorem there exists random variables \( \{Y_T^{(n)}, Y_T\} \) defined on a common probability space \((S, \mathcal{E}, \lambda)\), with the given probability measures \( \{\mu^{(n)}, \mu_L\} \), and such that \( Y_T^{(n)}(s) \rightarrow Y_T(s) \) for all \( s \in S \). By proposition 5.21 of book 2, pointwise convergence assures that \( Y_T^{(n)} \Rightarrow Y_T \). In summary, \( \{Y_T^{(n)}, Y_T\} \) are defined on a common probability space and have the same distributions as \( \{X_T^{(n)}, X_T\} \), and thus this proof proceeds with these random variables as a technical simplification.

Using the notation above, proposition 8.37 of book 2 assures that if \( Y_T^{(n)} \Rightarrow Y_T \) then \( O_T \left[ Y_T^{(n)} \right] \Rightarrow O_T \left[ Y_T \right] \) for any Borel measurable function

\[ O_T : (\mathbb{R}, \mathcal{B}(\mathbb{R}), m) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}), m), \]

such that \( m(D_{O_T}) = 0 \), where \( D_{O_T} \) denotes the discontinuity set of the \( O_T \). Since \( m(D_{O_T}) = 0 \) by assumption we conclude that the payoff random variable \( O_T \left[ Y_T^{(n)} \right] \) converges weakly to \( O_T \left[ Y_T \right] \).

The final step is to prove convergence of expectations:

\[
E \left[ O_T \left[ Y_T^{(n)} \right] \right] \rightarrow E \left[ O_T \left[ Y_T \right] \right],
\]

using proposition 5.8. By definition:

\[
E \left[ \left( O_T \left[ Y_T^{(n)} \right] \right)^2 \right] = \int \left( O_T \left[ Y_T^{(n)} \right] \right)^2 d\lambda.
\]
To apply proposition 3.14 of book 5 to change variables, note that \( Y^{(n)}_T \), the measure on \( \mathbb{R} \) induced by \( Y^{(n)}_T \), is by example 3.11 of that book given by \( \lambda^{(n)}_T \equiv \mu^{(n)}_T \). Thus by that proposition 3.14:

\[
\int \left( O_T \left[ Y^{(n)}_T \right] \right)^2 d\lambda = \int (O_T[y])^2 d\mu^{(n)}_T \equiv E_q \left( \left( O_T \left[ X^{(n)}_T \right] \right)^2 \right),
\]

where the last step comes from the observation above that \( Y^{(n)}_T \) and \( X^{(n)}_T \) have the same distribution. By 9.22 we obtain:

\[
\sup_n \int \left( O_T \left[ Y^{(n)}_T \right] \right)^2 d\lambda < \infty,
\]

and thus an application of proposition 5.8 proves that:

\[
\int O_T \left[ Y^{(n)}_T \right] d\lambda \to \int O_T [Y_T] d\lambda.
\]

Applying the same change of variables as above proves 9.24. In detail, \( \int O_T \left[ Y^{(n)}_T \right] d\lambda = E_q \left[ O_T \left[ X^{(n)}_T \right] \right] \) with expectation defined relative to \( d\mu^{(n)}_T \), the binomial distribution of \( X^{(n)}_T \), and \( \int O_T [Y_T] d\lambda = E \left[ O_T \left[ X_T \right] \right] \) with expectation defined relative to \( d\mu_L \), the lognormal distribution of \( X_T \) defined above. The final step is to note that:

\[
\int O_T [X_T] d\mu_L = \int_0^\infty O_T [y] f_L(y) dy
\]

by proposition 3.6 of book 5, where \( f_L \) is the lognormal density function with the above parameters.

The final expression in 9.25 is obtained by a change of variables again using proposition 3.14 of book 5, \( T : x \to X_0 e^x \), and identifying:

\[
\int_0^\infty O_T [y] f_L(y) dy \equiv \int O_T [y] d\mu_T.
\]

This requires that the measure \( \mu_T \) is that of the lognormal distribution function. But then by definition 3.9 of book 5 and using that \( T \) is invertible, the measure on the domain space \( \mu \) that induces \( \mu_T \) is defined on Borel sets \( A \subset \mathbb{R} \) by:

\[
\mu(A) = \mu_T(T(A)).
\]
Letting $A = (-\infty, x]$ and substitution in the resulting integral, the distribution function $F$ induced by $\mu$ is defined by:

$$F(x) = \int_0^{X_0 e^x} f_L(y)\,dy \equiv F_N(x),$$

the normal distribution function. The result in 9.25 is now given by proposition 3.14. □

**Remark 9.19** For the step in the above proof which concludes the convergence $E_q\left[ O_T X_T^{(n)} \right] \to E[O_T X_T]$, it is sufficient to require that $\{O_T X_T^{(n)}\}$ are uniformly integrable by proposition 5.6. By proposition 5.8 this conclusion is also obtained if $\sup_n E_q \left( |O_T X_T^{(n)}|^{1+\delta} \right) < \infty$ for some $\delta > 0$, so the assumption in 9.22 with $\delta = 1$ is a bit of overkill, but simplifies the presentation.

Alternatively, we can replace 9.22 by:

$$\sup_n E_q \left( |O_T X_T^{(n)}| \right) < \infty, \quad \sup_n \text{Var}_q \left( O_T X_T^{(n)} \right) < \infty. \quad (9.26)$$

This follows because

$$E_q \left( \left( O_T X_T^{(n)} \right)^2 \right) \leq \text{Var}_q \left( O_T X_T^{(n)} \right) + \left[ E_q \left( |O_T X_T^{(n)}| \right) \right]^2.$$

The advantage of the assumption that $\sup_n E_q \left( |O_T X_T^{(n)}|^{1+\delta} \right) < \infty$ for any $\delta > 0$ is that one obtains the uniform absolute mean bound for free, by Lyapunov’s inequality of corollary 3.51 of book 4.

While the restriction as in 9.22 is needed for the theoretical development of proposition 9.18, in the real world it can be argued that $\{O_T X_T^{(n)}\}$ are uniformly integrable without explicit additional assumptions. This follows because for any financial derivative, these payoffs are in fact uniformly bounded by global GDP say.

**Exercise 9.20** Show that 9.22 assures that $E \left( (O_T X_T)^2 \right) < \infty$ where this expectation is defined relative to the lognormal distribution of $X_T$. In other words:

$$\int_0^{\infty} \left( O_T y \right)^2 f_{LN}(y)\,dy < \infty. \quad (9.27)$$
9.3 LIMITING PRICE OF EUROPEAN DERIVATIVES

Hint: Define \( \{ Y^{(n)}_T, Y_T \} \) as in the above proof on a common probability space \((S, \mathcal{E}, \lambda)\), with the same distributions as \( \{ X^{(n)}_T, X_T \} \) and \( Y^{(n)}_T(s) \to Y_T(s) \) for all \( s \in S \). Recall Fatou’s lemma of proposition 2.18 of book 5 to show that:

\[
\int_S (O_T [Y_T])^2 \, d\lambda < \infty
\]

is implied by

\[
\sup_n \int_S (O_T [Y^{(n)}_T])^2 \, d\lambda < \infty.
\]

Justify the application of this proposition, and then justify all integral change of variables to the desired result.

Show that it is also the case that \( \text{Var} (O_T [y]) < \infty \).

**Example 9.21** The price of a long forward contract on an asset \( X \) which pays no income, or on a currency, is given by

\[
O_0 [X_0] = e^{R(r)T} X_0 - F e^{-rT},
\]

where \( F \) is the forward price. This equals the price of the replicating portfolio, which is a long position in \( X \) or \( e^{-rT} X \), respectively, and a loan at the risk-free rate which matures for \( F \).

This price is also produced with the formula above. To see this, note that \( O_T(x) = x - F \) and thus by 9.25:

\[
O_0 [X_0] \equiv e^{-rT} \int_{-\infty}^{\infty} (X_0 e^x - F) f_N(x) \, dx
\]

\[
= e^{-rT} \left[ X_0 \int_{-\infty}^{\infty} e^x f_N(x) \, dx - F \right].
\]

The integral is the moment generating function of this normal density evaluated at \( t = 1 \). Recalling 3.66 of book 4 this integral has value \( e^{R(r)T} \), and the result follows.

9.3.1 Black-Scholes-Merton Option Pricing Formulas

Applying proposition 9.18 above to European put and call options, one arrives at the famous **Black-Scholes-Merton formulas for the price of a European put and call option**. This result is named for **Fischer Black** (1938 -1995), **Myron S. Scholes** (b. 1941) and **Robert C. Merton** (b. 1944) for research published in 1973 in papers by Black and Scholes, and Merton, and for which Merton and Scholes received the 1997 Nobel Prize in Economics. Sadly, Black was deceased by that time and such awards are not made posthumously.
The payoff functions for these options at time $T$ are defined in reference to a fixed strike price $K$ by:

1. **Call Option:**
   \[ O_T^C = \max [X_T - K, 0] . \]  
   \[ (9.28) \]

2. **Put Option:**
   \[ O_T^P = \max [K - X_T, 0] . \]  
   \[ (9.29) \]

**Notation 9.22** One often sees these options referred to as vanilla European calls and vanilla European puts to distinguish them from the more exotic versions of such options discussed below, which are then labelled exotic European puts and calls. Analogously, one sees the American-style versions of these options with corresponding labels. That said, the terminology for financial derivatives is an evolving art-form, and it can be hazardous to make assumptions about contractual obligations.

Also, to simplify the above payoff expressions it is common to write these as:

\[ O_T^C = (X_T - K)^+ , \quad O_T^P = (K - X_T)^+ . \]

In order to justify the application of proposition 9.18, we must verify that 9.22 is satisfied, since continuity is apparent. The easiest way to do this is to note that

\[ E \left[ \left( O_T^C \left[ X_T^{(n)} \right] - O_T^P \left[ X_T^{(n)} \right] \right)^2 \right] = E \left[ \left( O_T^C \left[ X_T^{(n)} \right] \right)^2 \right] + E \left[ \left( O_T^P \left[ X_T^{(n)} \right] \right)^2 \right] , \]

since \( O_T^C \left[ X_T^{(n)} \right] O_T^P \left[ X_T^{(n)} \right] \equiv 0 \). Now:

\[ O_T^C \left[ X_T^{(n)} \right] - O_T^P \left[ X_T^{(n)} \right] \equiv X_T^{(n)} - K , \]

and so:

\[ E \left[ \left( O_T^C \left[ X_T^{(n)} \right] \right)^2 \right] + E \left[ \left( O_T^P \left[ X_T^{(n)} \right] \right)^2 \right] = E \left[ \left( X_T^{(n)} - K \right)^2 \right] . \]

To confirm that 9.22 is satisfied for both options we prove:

\[ \sup_n E \left[ \left( X_T^{(n)} - K \right)^2 \right] < \infty , \]
and this can be done by showing that \( \lim_n E \left[ \left( X_T^{(n)} - K \right)^2 \right] \) exists. To this end, since \( E \left[ X_T^{(n)} \right] = X_0 e^{rT} \) by 9.15:

\[
E \left[ \left( X_T^{(n)} - K \right)^2 \right] = E \left[ X_T^{(n)} \right]^2 - 2KX_0 e^{rT} + K^2
\]

\[
= X_0^2 e^{(2R(r) + \sigma^2)T} - 2KX_0 e^{rT} + K^2,
\]

and the next exercise provides the details on this \( \lim_n E \left[ \left( X_T^{(n)} \right)^2 \right] \) with \( R(r) \) defined as above.

**Exercise 9.23** Prove that:

\[
\lim_{n \to \infty} E \left[ \left( X_T^{(n)} \right)^2 \right] = X_0^2 e^{(2R(r) + \sigma^2)T},
\]

(9.30)

as follows:

1. First derive that

\[
E \left[ \left( X_T^{(n)} \right)^2 \right] = X_0^2 \left( q e^{2u(\Delta t)} + (1 - q) e^{2d(\Delta t)} \right)^n.
\]

2. Substitute for \( q \) using 9.8, 9.11 or 9.12 and then derive:

\[
E \left[ \left( X_T^{(n)} \right)^2 \right] = X_0^2 \left( e^{(\mu + R(r))\Delta t + \sigma \sqrt{\Delta t}} + e^{(\mu + R(r))\Delta t - \sigma \sqrt{\Delta t}} - e^{2\mu \Delta t} \right)^n,
\]

with \( R(r) = r, \) \( R(r) = r - r_f \) or \( R(r) = 0, \) respectively.

3. Focusing on the \( n \)th power and recalling that \( n = T/\Delta t \):

\[
\lim_{n \to \infty} \ln \left( e^{(\mu + R(r))\Delta t + \sigma \sqrt{\Delta t}} + e^{(\mu + R(r))\Delta t - \sigma \sqrt{\Delta t}} - e^{2\mu \Delta t} \right)^n
\]

\[
= T \lim_{\Delta t \to 0} \frac{1}{\Delta t} \ln \left( e^{(\mu + R(r))\Delta t + \sigma \sqrt{\Delta t}} + e^{(\mu + R(r))\Delta t - \sigma \sqrt{\Delta t}} - e^{2\mu \Delta t} \right)
\]

\[
= T \frac{df}{dx} \bigg|_{x=0}
\]

where

\[
f(x) = \ln \left( e^{(\mu + R(r))x + \sigma \sqrt{x}} + e^{(\mu + R(r))x - \sigma \sqrt{x}} - e^{2\mu x} \right).
\]

Justify that this function is differentiable at \( x = 0 \) by considering the Taylor series of \( e^{(\mu + R(r))x + \sigma \sqrt{x}} + e^{(\mu + R(r))x - \sigma \sqrt{x}} - e^{2\mu x}. \)
4. Evaluate this derivative, using a Taylor approximation for the evaluation at \( x = 0 \), and justify the expression for
\[
\lim_{n \to \infty} \left( e^{(\mu + R(r))\Delta t + \sigma \sqrt{\Delta t}} + e^{(\mu + R(r))\Delta t - \sigma \sqrt{\Delta t} - \mu \Delta t} \right)^n
\]
from this result.

**Remark 9.24** Note that the formula in 9.30 agrees with the formula for second moment of the lognormal distribution parameterized as in proposition 9.16, recalling these moment formulas from 3.68 of book 4. Thus the convergence in distribution in this model, \( X_T^{(n)} \Rightarrow X_T \), also obtains convergence of two moments:
\[
E \left[ (X_T^{(n)})^j \right] \to E \left[ (X_T)^j \right], \quad j = 1, 2.
\]

In proposition 9.159 of Reitano (2010) is proved that the moment generating function of \( \ln \left[ X_T^{(n)}/X_0 \right] \), which has a binomial distribution under \( q \) defined in 9.8, converges to the moment generating function of a normal distribution.

The final result for the **Black-Scholes-Merton formulas** follows.

**Proposition 9.25 (Black-Scholes-Merton)** For a **European call option**, the limiting price in 9.24 becomes:
\[
O^C_0(X_0) = e^{-(r-R(r))T}X_0\Phi(d_1) - e^{-rT}K\Phi(d_2), \quad (9.31a)
\]
while the result for a **European put option**:
\[
O^P_0(X_0) = e^{-rT}K\Phi(-d_2) - e^{-(r-R(r))T}X_0\Phi(-d_1), \quad (9.32)
\]
where \( \Phi \) denotes the distribution function of the standard normal, and
\[
d_1 = \frac{\ln (X_0/K) + (R(r) + \sigma^2/2)T}{\sigma \sqrt{T}}, \quad d_2 = \frac{\ln (X_0/K) + (R(r) - \sigma^2/2)T}{\sigma \sqrt{T}}. \quad (9.33)
\]

As above, \( R(r) = r \), \( R(r) = r - r_f \) or \( R(r) = 0 \), respectively, for options on assets with no income, currencies, and futures.

**Proof.** These formulas are assigned as exercises, and can be derived by an explicit evaluation of the associated Riemann integrals using change of variables in 9.24. \( \blacksquare \)
9.3 LIMITING PRICE OF EUROPEAN DERIVATIVES

Remark 9.26 The approach used by Black-Scholes and Merton was in spirit similar to that above, in the sense that they were able to "replicate" the option with a portfolio of $X$-assets and risk-free assets, and hence concluded that the option must therefore have a price equal to the price of this replicating portfolio. However, they used the advanced tools of stochastic calculus for this development which will be addressed in the next few books. The approach taken here, which used a binomial approximation to stock price movements, a discrete replication of the option, and then the evaluation of the limiting price as $\Delta t \to 0$ to the Black-Scholes-Merton formulas, is known as the Cox-Ross-Rubinstein binomial lattice model for option pricing. It was developed in a 1979 paper by John C. Cox, Stephen A. Ross and Mark Rubinstein.

9.3.2 Properties of Black-Scholes-Merton Formulas

Price Convergence to the Payoff

Rewriting somewhat more generally the price of a European derivative as given in 9.24, but now evaluated at time $t < T$:

$$O_t [X_t] \equiv e^{-r(T-t)} \int_{-\infty}^{\infty} O_T \left(X_t \exp \left[(R(r) - \sigma^2/2)(T-t) + x\sigma \sqrt{T-t}\right]\right) \phi(x) dx.$$  

(9.34)

As above, $R(r) = r$, $R(r) = r - r_f$ or $R(r) = 0$, respectively, for options on assets with no income, currencies, and futures. Also we use a change of variables in 9.24 to put the normal distribution parameters into the payoff function and thus $\phi(x) \equiv e^{-x^2/2}/\sqrt{2\pi}$, the unit normal density function of 3.2.

A natural question arises from 9.34. If $X_t \to X_T$ as $t \to T$, meaning convergence of the realized prices, is it the case that with $O_t [X_t]$ defined in 9.34 that $O_t [X_t] \to O_T (X_T)$? Of course this would in theory be enforced in the real world by arbitrage, but here this is a statement about the associated mathematics.

If the limit could be brought inside the integral in 9.34 then this result would certainly be obtained, so this question is really one of "integration to the limit" as developed in section 2.4 of book 5. As seen in that development there are several assumptions which allow the passing of the limit into the integrand, and the next result provides a general and easily applicable result. See also remark 9.28.
Proposition 9.27 Assume that the payoff function $O_T[x]$ at time $T$ is Borel measurable, and that its discontinuity set has Lebesgue measure zero. If $X_t \rightarrow X_T$ as $t \rightarrow T$, and for some $m$:

$$|O_T(x)| \leq c|x|^m. \quad (9.35)$$

then:

$$O_t[X_t] \rightarrow O_T(X_T). \quad (9.36)$$

Proof. We first show that 9.35 implies that for all $t \leq T$:

$$\int_{-\infty}^{\infty} O_T^2 \left(X_t \exp \left[ (R(r) - \sigma^2/2)(T-t) + x\sigma\sqrt{T-t} \right] \right) \phi(x)dx < C < \infty. \quad (\text{(**)})$$

This follows because the normal has a moment generating function $M_N(t)$ in 3.66 of book 4, and $X_t \rightarrow X_T$ as $t \rightarrow T$ implies this collection of prices is bounded, $|X_t| \leq k$ say. In detail:

$$\int_{-\infty}^{\infty} O_T^2 \left(X_t \exp \left[ (R(r) - \sigma^2/2)(T-t) + x\sigma\sqrt{T-t} \right] \right) \phi(x)dx$$

$$\leq c^2 \int_{-\infty}^{\infty} \left|X_t \exp \left[ (R(r) - \sigma^2/2)(T-t) + x\sigma\sqrt{T-t} \right] \right|^{2m} \phi(x)dx$$

$$\leq c^2 k^{2m} \exp \left[ 2m(R(r) - \sigma^2/2)(T-t) \right] M_N \left( 2m\sigma\sqrt{T-t} \right).$$

Now by proposition 3.6 of book 5,

$$O_t[X_t] \equiv e^{-r(T-t)} \int_{\mathbb{R}} O_T\left(X_t \exp \left[ (R(r) - \sigma^2/2)(T-t) + x\sigma\sqrt{T-t} \right] \right) d\mu_\phi(x),$$

where $\mu_\phi$ is the Borel measure induced by $\phi(x)$. Similarly, (**) can be restated:

$$\int_{-\infty}^{\infty} O_T^2 \left(X_t \exp \left[ (R(r) - \sigma^2/2)(T-t) + x\sigma\sqrt{T-t} \right] \right) d\mu_\phi(x) < C < \infty. \quad (\text{(***)})$$

By exercise 3.70 of book 4, restated relative to the measure $\mu_\phi$, (***) assures that

$$f_t(x) \equiv O_T\left(X_t \exp \left[ (R(r) - \sigma^2/2)(T-t) + x\sigma\sqrt{T-t} \right] \right)$$

is uniformly integrable relative to $\mu_\phi$ for $t \leq T$ (recall definition 2.50 of book 5). The uniform integrability convergence theorem of proposition 2.52 of book 5 now allows the passing of the limit under the integral, and 9.36 is obtained. $\blacksquare$
Remark 9.28 If \( O_T(x) \) is continuous and bounded, there is a direct proof of 9.36 using the portmanteau theorem of proposition 4.4. First rewrite 9.34:

\[
O_t [X_t] \equiv e^{-r(T-t)} \int_{-\infty}^{\infty} O_T (e^x) f_N(x) dx, \tag{9.37}
\]

where \( f_N(x) \) is the density function of the normal distribution with mean \( \mu \equiv \ln X_t + (R(r) - \sigma^2/2)(T-t) \) and variance \( \sigma^2 (T-t) \). Again applying proposition 3.6 of book 5:

\[
O_t [X_t] \equiv e^{-r(T-t)} \int_{\mathbb{R}} O_T (e^x) d\mu_N,
\]

where \( \mu_N \) is the Borel measure induced by \( f_N(x) \). By 6.22,

\[
C_N(s) = \exp \left( i \left[ \ln X_t + (R(r) - \sigma^2/2)(T-t) \right] s - \frac{1}{2} \sigma^2 (T-t) s^2 \right),
\]

and \( C_N(s) \to \exp(\pm \ln X_T) \) as \( t \to T \). This is the characteristic function of the degenerate distribution \( F_D(x) \) with \( x_0 = \ln X_T \) by 6.13, and thus by Lévy’s continuity theorem of proposition 6.16, \( \mu_N \Rightarrow \mu_D \), the measure induced by \( F_D(x) \). The portmanteau theorem of proposition 4.4 now obtains:

\[
O_t [X_t] = e^{-r(T-t)} \int_{\mathbb{R}} O_T (e^x) d\mu_N \to \int_{\mathbb{R}} O_T (e^x) d\mu_D = O_T (X_T).
\]

Put-Call Parity

It was noted in the prior section that at time \( T \), the payoff functions for European put and call options satisfy:

\[
O_T^C \left[ X_T^{(n)} \right] - O_T^P \left[ X_T^{(n)} \right] \equiv X_T^{(n)} - K.
\]

This identity has nothing to do with the binomial model of prices, and is true by definition for all values of \( X_T \). The interpretation of this is that if an investor has a portfolio with a long call option and a short put option at time \( t = 0 \), then at time \( T \) this investor will have a portfolio with value \( X_T - K \). This latter expression is the payoff for a long \( T \)-period forward contract on \( X \) with forward price \( K \). The actual contractual obligation of this forward to the long is that \( X \) must be purchased at time \( T \) for amount \( K \), and thus \( X_T - K \) is this contractholder’s profit and loss position at that time if the asset is immediately liquidated at price \( X_T \), or more generally the mark-to-market profit and loss at that time. Conversely, the contractual obligation for the short is to sell \( X \) for \( K \), and thus the short has a position that is equivalent to a long put and a short call.
Now if an investor creates a portfolio that is long one unit of \( X_0 \), now assumed to be an asset with no income, and short a risk-free security that matures for \( K \) at time \( T \), this portfolio will also have value \( X_T - K \) at time \( T \). By the law of one price and the assumption that \( X \) is actively traded, that these portfolios have exactly the same values at time \( T \) assures they must have identically the same value at time \( 0 \):

\[
O_0^C(X_0) - O_0^P(X_0) = X_0 - Ke^{-rT}. \quad (9.38)
\]

This relationship is known as put-call parity, and often expressed with conventional notation for European put and call premiums:

\[
c - p = X_0 - Ke^{-rT}.
\]

It states that independent of the individual pricing of puts and calls, it must be the case the difference between these prices equals the price of a long or short forward contract. Specifically, \( O_0^C(X_0) - O_0^P(X_0) \) equals the price of a long forward, and conversely. That \( X_0 - Ke^{-rT} \) is the price of a long forward again follows from the law of one price, that this is also the price of a portfolio of \( X \) and risk free assets which replicates the forward’s payoff.

The put-call parity identity has two interpretations:

1. It is an identity in the prices of European puts, European calls and forward contracts, or equivalently, an identity in the prices of European puts, European calls, the underlying asset, and risk-free bonds.

2. It is a replication recipe, which translates any algebraically equivalent version of this identity into a contract replication statement. For example:

\[
c = p + X_0 - Ke^{-rT},
\]

can be understood as:

\[
\text{long call} = \{\text{long put, long underlying asset, short RF asset}\}.
\]

Thus positive terms are long positions, negatives terms are short positions.

**Exercise 9.29** Derive put-call parity for options on currencies and futures using the above logic:

1. **Currency** (\( c, p \) denote option prices per unit of foreign currency, so \( X_0 \) is the currency price for one unit):

\[
c - p = X_0e^{-rT} - Ke^{-rT}. \quad (9.39)
\]
2. Futures \( (c, p \) denote options prices per futures contract on one unit of underlying assets, so here \( X_0 \) is the futures price): 

\[
c - p = X_0 e^{-rT} - Ke^{-rT} + \text{"long futures"}, \tag{9.40}
\]

where \( X_0 e^{-rT} \) is a risk free asset, and "long futures" is a long position in a futures contract on one unit of the underlying asset. As an identity in prices noted above, the long futures position disappears since initially a futures contract has a price of 0. Hint: Recall that when exercised, a call delivers cash of \( (X_T - X_0)^+ \) plus a long futures at price \( X_T \) to the long, while a put delivers cash of \( (X_0 - X_T)^+ \) plus a short futures at price \( X_T \) to the long.

### A Derivative’s "Greeks"

Various partial derivatives of the price of a financial derivative are called the "Greeks" of the derivative and are often denoted by Greek letters. This is a natural connection because "Delta," whether as upper case \( \Delta \) or lower case \( \delta \), has been codified in mathematical analysis as the abbreviation for "change," whether in variates like \( \Delta x \), or functions \( \Delta f \equiv f(x + \Delta x) - f(x) \), or combined in a Taylor series:

\[
\Delta f \approx f'(x) \Delta x + \frac{1}{2} f''(x) (\Delta x)^2 + ...
\]

The most common Greeks follow, although it should be noted that there are many more that are defined by mixed and higher partial derivatives, and inevitably the list is growing. When these derivatives exist, they are defined as follows:

1. Delta:

\[
\Delta \equiv \frac{\partial O_t [X_t]}{\partial X_t}.
\]

2. Gamma:

\[
\Gamma \equiv \frac{\partial^2 O_t [X_t]}{\partial X_t^2}.
\]

3. Vega:

\[
\nu \equiv \frac{\partial O_t [X_t]}{\partial \sigma}.
\]

Note that "vega" is not a Greek letter, but the letter commonly used for it is the Greek lower case "Nu," or simply the English lower case "Vee."
4. Rho:

\[ \rho = \frac{\partial O_t [X_t]}{\partial r} \]

5. Theta:

\[ \theta = \frac{\partial O_t [X_t]}{\partial t} \]

Note that as is standard practice, \( \partial / \partial t \) is evaluated holding all other variables constant, and thus there is no contribution from the dependence of \( X \) on \( t \).

The Greeks of a financial derivative are typically represented numerically, as are prices, but all are functions which depend on the various parameters which define the distribution of \( X_T \). For example, \( \Delta \equiv \Delta (X_t, \sigma, r, t) \) where the parameter \( t \) is primarily relevant because it identifies the important parameter of \( T-t \).

**Remark 9.30 (Differentiating "Under" the Integral)** Given a measure space \((S, \mathcal{E}, \lambda)\), \( E \in \mathcal{E} \), and open set \( O \subset \mathbb{R} \), let \( f : O \times E \rightarrow \mathbb{R} \) be a function and consider \( g : O \rightarrow \mathbb{R} \) defined by:

\[ g(x) = \int_E f(x, s) d\lambda(s). \]

In calculus with \((S, \mathcal{E}, \lambda) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), m)\) and \( E = [a, b] \) say, **Leibniz’s rule** provides conditions on \( f(x, s) \) so that \( g(x) \) as a Riemann integral is differentiable and

\[ g'(x) = \int_E f_x(x, s) d\lambda(s), \quad \text{(**)} \]

where \( f_x(x, s) \) denotes the \( x \)-partial derivative of \( f \). Leibniz’s rule is named for **Gottfried Leibniz** (1646 – 1716).

The general case can be addressed with the aid of various theorems on integrating to the limit in section 2.4 of book 5. To this end, note that:

\[ \frac{g(x + \Delta x) - g(x)}{\Delta x} = \int_E \frac{f(x + \Delta x, s) - f(x, s)}{\Delta x} d\lambda(s), \]

and so when this limit exists:

\[ g'(x) = \lim_{\Delta x \to 0} \int_E \frac{f(x + \Delta x, s) - f(x, s)}{\Delta x} d\lambda(s). \]

Fix \( x \) and assume that \( f_x(x, s) \) exists \( \lambda \)-a.e. Then \( f_n(x, s) \equiv \frac{f(x + \Delta x_n, s) - f(x, s)}{\Delta x_n} \rightarrow f_x(x, s) \) \( \lambda \)-a.e. for an arbitrary sequence \( \Delta x_n \to 0 \), and thus the general
version of the identity in (*) relies on a justification for bringing the limit inside the integral, a procedure addressed in various results of section 2.4 of book 5.

For example:

1. Assume that $\lambda[E] < \infty$ and for each $x$, $f_x(x,s)$ is $\lambda$-measurable and $|f_n(x,s)| \leq M$ for all $n$, $\lambda$-a.e. on $E$. Then the bounded convergence theorem of proposition 2.46 of book 5 justifies the result in (*) for each $x$. A special case of this obtains Leibniz’s rule where $E = [a,b]$ and $f_x(x,s)$ is assumed continuous. Then given $x$, $f_n(x,s) \to f_x(x,s)$ converges pointwise for $s \in E$, and since $E$ is compact this implies uniform convergence. Thus since $f_x(x,s)$ is bounded on $E$, uniform convergence implies that $|f_n(x,s)| \leq M$.

2. Assume that for each $x$ that $f_x(x,s)$ is $\lambda$-measurable, and there exists $h(x,s)$ with $\int_E h(x,s)d\lambda(s) < \infty$ such that $|f_n(x,s)| \leq h(x,s) \lambda$-a.e. for all $n$. Then Lebesgue’s dominated convergence theorem of proposition 2.43 of book 5 justifies the result in (*) for each $x$. A special case of this applies in the proof below, where for each $x$, $|f_x(x,s)| \leq k(x,s)$ with $k(x,s)$ bounded and uniformly integrable in $s$ over compact $x$-sets. Then by the mean value theorem,

$$|f_n(x,s)| \leq \max_I |f_x(x,s)| \leq \max_I k(x,s),$$

where this maximum is over the interval $I \equiv [x - \Delta x_n, x + \Delta x_n]$. By uniform integrability, $\max_I k(x,s) \leq h(x,s)$ with $h$ integrable. This approach applies to $g''(x)$ and higher derivatives by iteration.

It is remarkably difficult to establish general criteria for the existence of the Greeks for a general European financial derivative. The complication is that we seek to establish derivatives of a price function parametrized as in 9.34 or 9.37, and to do so we need to justify differentiating under the integral. This in turn requires conditions on $O_T(x)$ which allow the application of one of the "integration to the limit" results of section 2.4 of book 5, and also the messy calculation of such general derivatives. The parametrization is 9.34 is simpler to work with, but then we must require differentiability of $O_T(x)$, an assumption that fails even for simple puts and calls. The parametrization in 9.37 allows differentiation without smoothness assumptions on $O_T(x)$, but here the execution of derivatives quickly becomes unwieldy.

The next result will use the 9.34 parametrization, and thus the requirement on differentiability and growth of $O_T(x)$ to justify the manipulations.
CHAPTER 9 PRICING OF FINANCIAL DERIVATIVES

However, this result is true more generally. This result in 9.42 is known as the Black-Scholes-Merton partial differential equation.

Proposition 9.31 (Black-Scholes-Merton PDE) Assume that $O_T(x)$ is twice continuously differentiable, and:

$$\left| \frac{\partial^n O_T(x)}{\partial x^n} \right| \leq c_n |x|^{m_n} \quad (9.41)$$

for $n \leq 2$, where we define $\frac{\partial^n O_T(x)}{\partial x^n} \equiv O_T(x)$ for $n = 0$. Let $f(t, x) \equiv O_t[x]$ where the price function $O_t[X_t]$ as given in 9.34 or any of the equivalent forms. Then for $t < T$:

$$\frac{\partial f}{\partial t} + R(r)x \frac{\partial f}{\partial x} + \frac{1}{2} \sigma^2 x^2 \frac{\partial^2 f}{\partial x^2} = rf, \quad (9.42)$$

and:

$$f(T, x) = O_T(x).$$

Thus $\theta$, $\Delta$ and $\Gamma$ exist for $t < T$ and:

$$\theta + R(r)X_t\Delta + \frac{1}{2}\sigma^2 X_t^2 \Gamma = rO_t[X_t]. \quad (9.43)$$

Proof. The boundary condition $f(T, x) = O_T(x)$ is addressed in proposition 9.27 by 9.41, and thus we focus on 9.42. By 9.34:

$$f(t, x) \equiv e^{-r(T-t)} \int_{-\infty}^{\infty} O_T\left(x \exp \left[\left( R - \sigma^2/2 \right) (T-t) + y \sigma \sqrt{T-t} \right] \right) \phi(y) dy.$$  

We will differentiate under the integral formally, and then justify this by observing that the resulting integrands satisfy the conditions of 2 of remark 9.30. To simplify notation, let $Y_t \equiv (R - \sigma^2/2)(T-t) + y\sigma\sqrt{T-t}$. A calculation obtains:

$$\frac{\partial f}{\partial t} = rf - e^{-r(T-t)} \int_{-\infty}^{\infty} O_T'(x \exp Y_t) x \exp Y_t \left[ \left( R - \sigma^2/2 \right) + \frac{y \sigma}{2\sqrt{T-t}} \right] \phi(y) dy.$$  

Note that this integral is well defined since the expression preceding $\phi(y)$ is uniformly continuous and grows at worst as $e^{cy}$.Specifically, fixing $x$ and letting $I(t, y)$ denote this integrand, we have by the above growth bound and denoting $a \equiv (R - \sigma^2/2)$ to simplify notation:

$$|I(t, y)| \leq c \left| x \exp \left[ a (T-t) + y \sigma \sqrt{T-t} \right] \right|^{m_1} a + \frac{y \sigma}{2\sqrt{T-t}} e^{-y^2/2}.$$
9.3 LIMITING PRICE OF EUROPEAN DERIVATIVES

For \( t < T \) and \( t + \Delta t < T \) this bound is uniformly integrable in \( y \) over \([t - \Delta t, t + \Delta t]\), and thus remark 9.30 applies. We leave the details related to the other derivatives below to the reader.

Similarly:

\[
\frac{\partial f}{\partial x} = e^{-r(T-t)} \int_{-\infty}^{\infty} O'_T (x \exp Y_t) \exp Y_t \phi(y) dy,
\]

and thus

\[
\frac{\partial f}{\partial t} = rf - (R(r) - \sigma^2/2)x \frac{\partial f}{\partial x} e^{-r(T-t)} \left( \frac{\sigma x}{2\sqrt{T-t}} \right) \int_{-\infty}^{\infty} O'_T (x \exp Y_t) \exp Y_t \phi(y) dy.
\]

Also:

\[
\frac{\partial^2 f}{\partial x^2} = e^{-r(T-t)} \int_{-\infty}^{\infty} O''_T (x \exp Y_t) (\exp Y_t)^2 \phi(y) dy.
\]

Now

\[
\frac{\partial O'_T (x \exp Y_t)}{\partial y} = \sigma \sqrt{T-t} O''_T (x \exp Y_t) x \exp Y_t,
\]

and thus rewriting:

\[
\frac{\partial^2 f}{\partial x^2} = \frac{e^{-r(T-t)}}{x \sigma \sqrt{T-t}} \int_{-\infty}^{\infty} \frac{\partial O'_T (x \exp Y_t)}{\partial y} \exp Y_t \phi(y) dy.
\]

Performing integration by parts on the integral and noting that the boundary values are zero obtains:

\[
\frac{\partial^2 f}{\partial x^2} = -\frac{e^{-r(T-t)}}{x \sigma \sqrt{T-t}} \int_{-\infty}^{\infty} \frac{\partial \exp Y_t \phi(y)}{\partial y} O'_T (x \exp Y_t) dy
\]

\[
= \frac{e^{-r(T-t)}}{x \sigma \sqrt{T-t}} \int_{-\infty}^{\infty} O'_T (x \exp Y_t) \exp Y_t \phi(y) dy
\]

\[
- \frac{e^{-r(T-t)}}{x} \int_{-\infty}^{\infty} O'_T (x \exp Y_t) \exp Y_t \phi(y) dy,
\]

and using the above expression for \( \frac{\partial f}{\partial x} \):

\[
\frac{1}{2} \sigma^2 x^2 \frac{\partial^2 f}{\partial x^2} = e^{-r(T-t)} \left( \frac{\sigma x}{2\sqrt{T-t}} \right) \int_{-\infty}^{\infty} O'_T (x \exp Y_t) \exp Y_t \phi(y) dy
\]

\[
- \frac{1}{2} \sigma^2 x \frac{\partial f}{\partial x}.
\]

Substituting the integral in this expression in (\* \*) completes the proof. ■
Exercise 9.32  Show that given the assumptions of the above proposition, that \( v = \frac{\partial O_t[X_t]}{\partial \sigma} \) also exists.

Remark 9.33  The financial derivative price function specified in 9.34 satisfies the partial differential equation in 9.42 under much more general conditions on the payoff function \( O_T(x) \) than those specified above. One purely mathematical approach to this result is to transform the equation in 9.42 to the so-called heat equation using a change of variables from \( f(t,x) \) to \( u(\tau,y) \) to obtain:

\[
\frac{\partial u}{\partial \tau} = \frac{\partial^2 u}{\partial y^2}.
\]

This equation is solved on the \( \tau \)-time interval \( 0 \leq \tau < \infty \), but because \( \tau \approx -t \), time is reversed and the boundary condition is switched from \( f(T,x) = O_T(x) \) to an initial value condition \( u(0,y) = \tilde{O}_T(y) \) where \( \tilde{O}_T(y) \) denotes the transformed \( O_T(x) \). Now \( h(\tau,y) = \frac{1}{\sqrt{4\pi \tau}} \exp \left[-y^2/4\tau\right] \) solves this differential equation for \( \tau > 0 \), and it then turns out that \( u(\tau,y) = \int h(\tau,y-x)\tilde{O}_T(x)dx \) also solves this differential equation and the initial value condition. The conditions on \( \tilde{O}_T(x) \) needed to justify differentiation under the integral and taking the limit \( \tau \to 0 \) are relatively mild and include measurability and growth bounds.

See the chapter 5 exercises of Etheridge (2002) for the needed transformation, and any standard text on partial differential equations for more details on the heat equation and the fundamental solution of the heat equation, \( h(\tau,y) \). Note that \( u(\tau,y) \) above can be defined as:

\[
u(\tau,y) = h \ast_y \tilde{O}_T(\tau,y)
\]

where \( \ast_y \) denotes a convolution in the \( y \)-variate. See chapter 2 above and section 6.2 of book 5 for more on convolutions.

Example 9.34  By example 9.21, the price of a long forward contract on an asset \( X \) which pays no income or a currency is given by \( O_t[X_t] = e^{(R(r)-r)(T-t)}X_t - Fe^{-r(T-t)} \) where \( F \) is the forward price. Hence \( \theta = -(R(r)-r)e^{(R(r)-r)(T-t)}X_t - Fe^{-r(T-t)} \), while \( \Delta = e^{(R(r)-r)(T-t)} \) and \( \Gamma = 0 \). Thus:

\[
\theta + R(r)X_t \Delta + \frac{1}{2} \sigma^2 X_t^2 \Gamma = rO_t[X_t].
\]

In the case of European puts and call, the price functions of proposition 9.25 are seen to be infinitely differentiable in \( X, \sigma, r, \) and \( t \), and some of the above Greeks are calculated with the following exercise. These put and call prices also satisfy 9.42, though details are left to the reader.
9.3 LIMITING PRICE OF EUROPEAN DERIVATIVES

Exercise 9.35 Restate the formulas in proposition 9.25 to be evaluated at time \( t < T \) as follows. For a European call option:

\[
O_t^C(X_t) = e^{-(r-R)(T-t)}X_t\Phi(d_1) - e^{-r(T-t)}K\Phi(d_2), \tag{9.44}
\]

while the result for a European put option is:

\[
O_t^P(X_t) = e^{-r(T-t)}K\Phi(-d_2) - e^{-r-R}(T-t)X_t\Phi(-d_1), \tag{9.45}
\]

where \( \Phi \) denotes the distribution function of the standard normal, and:

\[
d_1 = \frac{\ln(X_t/K) + (R(r) + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, \quad d_2 = \frac{\ln(X_t/K) + (R(r) - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}. \tag{9.46}
\]

As above, \( R(r) = r \), \( R(r) = r - r_f \) or \( R(r) = 0 \), respectively, for options on assets with no income, currencies, and futures.

Derive the following, where \( \phi \) denotes the density function of the standard normal in 3.2. Note that 9.47 looks elementary, but remember that \( d_1 \) and \( d_2 \) are functions of \( X_t \):

\[
\Delta_t^C = e^{-(r-R)(T-t)}\Phi(d_1), \quad \Delta_t^P = -e^{-r-R}(T-t)\Phi(-d_1), \tag{9.47}
\]

\[
\Gamma_t^C = \Gamma_t^P = e^{-(r-R)(T-t)}\phi(d_1)/X_t\sigma\sqrt{T-t}. \tag{9.48a}
\]

If you enjoy differentiation, derive formulas for vega, rho and theta, and verify that the above put and call prices satisfy 9.42.

A Derivative’s "Greeks" - Lattice Approximations

Because \( O_t^{(n)}[X_t] \) is by the law of one price, equal to the price of the replicating portfolio for a European financial derivative at time \( t \), it is the case that:

\[
O_t^{(n)}[X_t] = a_t^{(n)}X_t + b_t^{(n)},
\]

where \( a_t^{(n)} \) denotes number of units of the underlying asset with market value \( a_t^{(n)}X_t \), and \( b_t^{(n)} \) is the market value of an investment in the risk free asset at rate \( r \). When the underlying asset is a futures contract, the market value of \( a_t^{(n)} \) units of the underlying is 0, and thus \( O_t^{(n)}[X_t] = b_t^{(n)} \). The signs of \( a_t^{(n)} \) and \( b_t^{(n)} \) determine whether the respective positions are long or short. Here, because the focus is on time \( t \) and the evaluation of how these parameters change as \( n \to \infty \), this notation differs from the notation in the proof of 9.7 where these were respectively denoted \( a_{0,0} \) and \( b_{0,0} \).
While it is tempting to think that this formula implies that the option price has a delta, meaning an $X_t$-calculus derivative, and moreover that at time $t$
\[
\frac{dO_t^{(n)}[X_t]}{dX_t} = a_t^{(n)},
\]
this calculation is not immediately justified. First off, the above development concluded only that $O_t^{(n)}[X_t]$ was equal to $a_t^{(n)}X_t + b_t^{(n)}$ under the assumption that $X_t$-prices were restricted within the binomial model as $X_t \rightarrow X_te^{u(\Delta t)}$ or $X_t \rightarrow X_te^{d(\Delta t)}$. But even within the context of this model, the above delta calculation implies that $a_t^{(n)}$ and $b_t^{(n)}$ are independent of $X_t$, and by 9.9 and 9.10 it is clear that they are not.

Recall that by 9.9:
\[
a_t^{(n)} = \frac{O_t^{(n)}[X_t - e^{u(\Delta t)}] - O_t^{(n)}[X_t - e^{d(\Delta t)}]}{X_te^{u(\Delta t)} - X_te^{d(\Delta t)}}.
\] (9.49)

Thus if $O_t^{(n)}(x)$ is indeed a differentiable function of $x$, this formula states that $a_t^{(n)}$ provides a central difference approximation to $dO_t^{(n)}/dx$ evaluated at the mean price defined by:
\[
\bar{X}_{t+\Delta t} = X_t \left( e^{u(\Delta t)} + e^{d(\Delta t)} \right) / 2.
\]
To see this, recall that the central difference approximation to $f'(x_0)$:
\[
f'(x_0) \approx \frac{f(x_0 + \Delta x) - f(x_0 - \Delta x)}{2\Delta x}.
\] (9.50)

The error of this approximation is $O[(\Delta x)^2]$ when $f(x)$ has 3 derivatives, as is justified with a Taylor series analysis. Here we use $O$ for the big-$O$ notation of remark 9.10 to avoid notational confusion with the function $O_t$.

In the current application, $a_t^{(n)}$ in 9.49 can be expressed in the form 9.50 with $x_0 = \bar{X}_{t+\Delta t}$ where:
\[
\bar{X}_{t+\Delta t} \equiv .5 \left( X_te^{u(\Delta t)} + X_te^{d(\Delta t)} \right),
\]
and
\[
\Delta x \equiv .5 \left( X_te^{u(\Delta t)} - X_te^{d(\Delta t)} \right).
\]
The error between $a_t^{(n)}$ and $dO_t^{(n)}/dx$ is thus $O[(\Delta x)^2]$ when $O_t^{(n)}(x)$ has 3 derivatives. Recalling 9.1, a calculation produces:
\[
\bar{X}_{t+\Delta t} = X_te^{\mu \Delta t} \left( 1 + \frac{1}{2}\sigma^2\Delta t + O[(\Delta t)^2] \right),
\]
9.3 LIMITING PRICE OF EUROPEAN DERIVATIVES

and:

\[ O \left[ (\Delta x)^2 \right] = O [\Delta t] . \]

Hence,

\[ a_t^{(n)} = \frac{dO^{(n)}_{t+\Delta t}(x)}{dx} \bigg|_{x=X_t+\Delta t} + O [\Delta t] , \]

and the binomial model \( a_t^{(n)} \)-coefficients approximate the delta of the financial derivative at time \( t + \Delta t \), when the asset price is \( X_t e^{\mu \Delta t} + O [\Delta t] \).

As noted in remark 9.15, it is common to set \( \mu = 0 \) for derivatives pricing and thus when \( \Delta t \) is small, \( a_t^{(n)} \) is a good approximation to the current delta, meaning with price \( X_t \), within the binomial model framework.

This approach also applies to the estimation of gamma using the lattice prices at time \( t + 2\Delta t \), transformed into deltas at time \( t + \Delta t \). For example, dropping the superscript \( (n) \) to make room for \( u/d; \) let:

\[ a_{t+\Delta t}^u = \frac{O_{t+2\Delta t}^{(n)} \left[ X_t e^{u(\Delta t)} \right] - O_{t+\Delta t}^{(n)} \left[ X_t e^{u(\Delta t)+d(\Delta t)} \right]}{X_t e^{u(\Delta t)} - X_t e^{u(\Delta t)+d(\Delta t)}} . \]

Then \( a_{t+\Delta t}^u \) is the asset position at that node, in the sense that \( O_{t+\Delta t}^{(n)} \left[ e^{u(\Delta t)} X_t \right] = a_t^u e^{u(\Delta t)} X_t + b_t^u \), and by the above analysis:

\[ a_{t+\Delta t}^u = \frac{dO_{t+2\Delta t}^{(n)} (x)}{dx} \bigg|_{x=X_t+2\Delta t} + O [\Delta t] , \]

where \( \bar{X}_{t+2\Delta t}^u \equiv \frac{1}{2} (X_t e^{2u(\Delta t)} + X_t e^{u(\Delta t)+d(\Delta t)}) \). Defining \( a_{t+\Delta t}^d \) analogously, gamma can be estimated:

\[ \frac{a_{t+\Delta t}^u - a_{t+\Delta t}^d}{.5 \left( X_{t+2\Delta t}^u - X_{t+2\Delta t}^d \right)} \approx \frac{d^2O_{t+2\Delta t}^{(n)} (x)}{dx^2} \bigg|_{x=X_t+2\Delta t} \]

with \( \bar{X}_{t+2\Delta t}^u \equiv \frac{1}{2} \left( X_{t+2\Delta t}^u + X_{t+2\Delta t}^d \right) \).

Of course one can always recalculate the lattice at prices \( X_t \pm \Delta X \) for some value of \( \Delta X \) and then apply 9.50 directly, and the analogous formula for \( f''(x_0) \):

\[ f''(x_0) \approx \frac{f(x_0 + \Delta x) + f(x_0 - \Delta x) - 2f(x_0)}{\Delta x^2} . \] (9.51)
CHAPTER 9 PRICING OF FINANCIAL DERIVATIVES

The error of this approximation is $O\left( (\Delta x)^2 \right)$ when $f(x)$ has 4 derivatives, as is justified with a Taylor series analysis.

Lattice recalculation are necessary for other Greeks, such as vega, theta and rho.

9.4 Limiting Price of American Derivatives

Though beyond the scope of the materials developed here, there are also a host of research results on the convergence as $\Delta t \to 0$ of American option prices derived from binomial models. We provide three earlier references. Using advanced tools related to stochastic processes that are studied in books 7-9, K. Amin and A. Khanna (1994) prove that if the sequence of binomial models converges weakly to the stochastic process (or, diffusion) of asset prices, then the corresponding sequence of binomial American option prices also converges to the American option price implied by that diffusion.

Dietmar P.J. Leisen (1998) evaluates the order of convergence for American put option prices under several variants of binomial models, and proposes a model that accelerates this convergence. Lastly, the paper of Lishang Jiang and Min Dai (1999) focuses on American call options and a generalization, and proves convergence of binomial prices using methods of numerical analysis.

9.5 Binomial Pricing of Path Dependent European Options

A path dependent European option is a financial contract with payoff function at time $T$ that depends not only on the value of the $X_T$, but also on the values of $X_t$ for $t < T$. In other words, the payoff at time $T$ not only depends on the ultimate value of the underlying asset, but also on the "path" of asset prices from $X_0$ to $X_T$. It is apparent that the pricing approach of the last section requires some adaptation, since working on a binomial lattice it is not even possible to complete the first step in the proof of proposition 9.4. Recall that in the proof of that proposition’s conclusion in 9.7, the first step is the valuation of derivative payoffs at expiry, $O_T \left[ X_0 e^{j u (\Delta t)} e^{(n-j) d(\Delta t)} \right]$. But for given $j$ the time $T$ asset price $X_0 e^{j u (\Delta t)} e^{(n-j) d(\Delta t)}$ has $\binom{n}{j}$ different paths which lead to this value, and
thus potentially \( \binom{n}{j} \) different payoffs for a path dependent derivative security.

In order to accommodate this application, we first restate the binomial pricing result for standard European derivatives in a path-based framework.

9.5.1 Binomial Path-Based Pricing

Standard European Options

In the framework developed for the binomial lattice pricing formula in 9.7, the price of an option \( O_0^{(n)}[X_0] \) equals the present value of the expected payoffs: \( O_T \left[ X_0 e^{u(\Delta t)} e^{(n-j)d(\Delta t)} \right] \) for \( j = 0, 1, \ldots, n \), where this expectation is calculated using a binomial distribution with parameters \( q(\Delta t) \) in 9.8, 9.11 or 9.12 and \( n = T/\Delta t \). Within this model there are \( n+1 \) final state prices:

\[
X_0 e^{u(\Delta t)} e^{(n-j)d(\Delta t)}
\]

for \( j = 0; 1; \ldots; n; \) where this expectation is calculated using a binomial distribution with parameters \( q(\Delta t) \) in 9.8, 9.11 or 9.12 and \( n = T/\Delta t \). Within this model there are \( n+1 \) final state prices:

\[
X_0 e^{u(\Delta t)} e^{(n-j)d(\Delta t)}
\]

for \( j = 0; 1; \ldots; n; \) with respective binomial probabilities:

\[
\binom{n}{j} q^j (1-q)^{n-j}
\]

We can formally restate the formula in 9.7 in terms of the \( 2^n \) paths implied by this model, where a path is defined by an \( (n+1) \)-tuple of prices:

\[
(X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{n\Delta t}),
\]

where for all \( j \), \( X_{(j+1)\Delta t} = e^{u(\Delta t)} X_{2\Delta t} \) or \( X_{(j+1)\Delta t} = e^{d(\Delta t)} X_{2\Delta t} \). Any such path can equally well be identified by the \( n \)-tuple of \( u(\Delta t) \) and \( d(\Delta t) \) values which generates it, or, by the \( n \)-tuple of binomial variates \( b_j = \pm 1 \) which generates these \( u(\Delta t) \) and \( d(\Delta t) \) values. The probability of any such path is \( q^j (1-q)^{n-j} \) when this path is defined by \( j \)-\( u(\Delta t) \)s and \( (n-j) \)-\( d(\Delta t) \)s. It is then possible to express 9.7 by:

\[
O_0^{(n)}[X_0] = e^{-rT} \sum_{2^n} q^j (1-q)^{n-j} O_T [(X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{n\Delta t})],
\]

where this summation is over all \( 2^n \) \( n \)-tuples defined in terms of all sequences of returns or binomial variates. As above, \( j \) then equals the number of \( u(\Delta t) \)s or \( +1 \)s in the given \( n \)-tupel. For any \( j \) there are \( \binom{n}{j} \) such \( n \)-tuples, and hence when \( O_T [(X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{n\Delta t})] = O_T [X_{n\Delta t}] \), this formula simply reduces to the earlier formula.

While it is tempting to simply apply this reformulated version of the option pricing formula to path-dependent options, now simply inserting payoff functions which formulaically reflect the price path \( (X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{n\Delta t}) \), we cannot justify this without further analysis. Indeed, the formula in 9.7
was derived from a replicating portfolio argument. This formula provided the market value of a portfolio of underlying assets and risk free securities at time 0, which if appropriately rebalanced each period would ultimately replicate the \( n + 1 \) payoffs at time \( T \) for a standard European option. In order to support the conclusion that the formula in \((*)\) provides the price of a path-dependent option, we must demonstrate that this again equals the price of such a portfolio, which after rebalancing will replicate the \( 2^n \) path-based payoffs implied by such an

**Path-Dependent European Options**

In this section we demonstrate that \( O_0^{(n)}[X_0] \) given in \((*)\) is indeed the price of a portfolio of underlying assets and risk free securities which can be rebalanced through time so that at time \( T \), the portfolio will have value \( O_T[(X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{n\Delta t})] \) when \( (X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{n\Delta t}) \) is the price path realized. In general, such payoff functions do not explicitly reflect the value of \( X_0 \), since this value is known on the pricing date at \( t = 0 \), but we reflect this value notationally because the rest of the path depends on this value.

The demonstration of the following result reflects a relatively minor adaptation of the proof above for standard European options, and thus details are left as an exercise. The conclusion in 9.52 is the same as it was for 9.7 and notationally expressed in 9.13. The derivative security price today, \( O_0^{(n)}[X_0] \), equals the risk free present value of the expected payoff at time \( T = n\Delta t \), where this expectation reflects the distribution of \( 2^n \) paths with risk neutral probability \( q \equiv q(\Delta t) \) given by 9.8, 9.11 or 9.12 and \( n = T/\Delta t \).

**Proposition 9.36** If \( O_T[(X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{n\Delta t})] \) denotes the payoff of a path-dependent European derivative security at time \( T \) as a function of the realized asset price path, the price at time 0 of a replicating portfolio is given by:

\[
O_0^{(n)}[X_0] = e^{-rT} \sum_{2^n} q^j (1 - q)^{n-j} O_T[(X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{n\Delta t})]. \tag{9.52}
\]

Here \( r \) denotes the risk free interest rate, and the summation is over all \( 2^n \) \( n \)-tuples of binomial variates where for each such \( n \)-tuple, \( j \) denotes the number of binomial \( b = +1 \) variates. Here as above, \( q \equiv q(\Delta t) \) is given by 9.8, 9.11 or 9.12 is the risk-neutral probability that \( b = +1 \), and \( n = T/\Delta t \).

**Proof.** The derivation of this option pricing result requires a small reorganization of the previous derivation for proposition 9.4, and in fact the first two steps differ only in notational representation.
9.5 BINOMIAL PRICING OF PATH DEPENDENT EUROPEAN OPTIONS

1. At time $j\Delta t$, given any of the $2^j$ asset price paths $(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{j\Delta t})$, and given any payoff function values $f(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{j\Delta t}, X_{(j+1)\Delta t})$, there are real constants $a_j, b_j$, so that an initial portfolio of $a_j$ units of $X_{j\Delta t}$, and $b_j$ units invested at rate $r$ exactly replicates the $2^j$ values of $f(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{j\Delta t}, X_{(j+1)\Delta t})$ at time $(j + 1)\Delta t$, where $X_{(j+1)\Delta t} = X_{j\Delta t}e^{u(\Delta t)}$ or $X_{(j+1)\Delta t} = X_{j\Delta t}e^{d(\Delta t)}$. Specifically, denoting $(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{j\Delta t}) \equiv \bar{X}$ for notational simplicity:

$$a_j = \frac{f(\bar{X}, e^{u(\Delta t)}X_{j\Delta t}) - f(\bar{X}, e^{d(\Delta t)}X_{j\Delta t})}{(e^{u(\Delta t)} - e^{d(\Delta t)}) X_{j\Delta t}}, \quad (9.53)$$

and,

$$b_j = \left(\frac{e^{u(\Delta t)}f(\bar{X}, e^{d(\Delta t)}X_{j\Delta t}) - e^{d(\Delta t)}f(\bar{X}, e^{u(\Delta t)}X_{j\Delta t})}{e^{u(\Delta t)} - e^{d(\Delta t)}}\right) e^{-r\Delta t}. \quad (9.54)$$

2. For any given $(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{j\Delta t})$, the price of this portfolio at time $j\Delta t$, and hence the value at time $j\Delta t$ of the associated payoffs, can be algebraically manipulated into the form:

$$f(\bar{X}) = e^{-r\Delta t} \left[ qf(\bar{X}, X_{j\Delta t}e^{u(\Delta t)}) + (1 - q)f(\bar{X}, X_{j\Delta t}e^{d(\Delta t)}) \right].$$

3. The formula in 2 can be applied starting at time $T = n\Delta t$, at which time each of the $2^n$ price paths provide the final payoff values,

$$O_T [(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{n\Delta t})].$$

The result in 2 is then applied by pairing final values, where each pair shares the common price path $(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{(n-1)\Delta t})$, producing the time $(n - 1)\Delta t = T - \Delta t$ payoff function:

$$f(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{(n-1)\Delta t}) \equiv O_{(n-1)\Delta t}^{(n)} (X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{(n-1)\Delta t}).$$

Each of these $2^{n-1}$ values equals the price at time $T - \Delta t$ of a portfolio that replicates the associated pairs of time $T = n\Delta t$ payoff values. These $2^{n-1}$ values at time $T - \Delta t$ are then paired in the same way, where each pair shares the common price path $(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{(n-2)\Delta t})$, producing $2^{n-2}$ values of $O_{(n-2)\Delta t}^{(n)} (X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{(n-2)\Delta t})$. Again, each such value equals the price of a portfolio at time $T - 2\Delta t$ that
replicates the associated time $T - \Delta t$ payoff values, which in turn provide for replicating portfolios for the time $T$ payoffs. Continuing in this way, we derive by induction:

$$O^{(n)}_{j\Delta t}(X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{j\Delta t}) = e^{-r(n-j)\Delta t} \sum_{q=0}^{n-j} q^i (1 - q)^{n-j-i} O_T(X_0, X_{\Delta t}, \ldots, X_{j\Delta t}, X_{(j+1)\Delta t}, \ldots, X_{n\Delta t}),$$

where this summation is over all $2^{n-j}, (n-j)$-tuples of price paths defined by $(X_{(j+1)\Delta t}, \ldots, X_{n\Delta t})$.

4. This general formula produces 9.7 when $j = 0$.

\[ \square \]

**Example 9.37** Examples of common path dependent options are as follows:

1. **Asian Options:** These put and call options have payoff functions which reflect the average price of the underlying assets. Such averages can be defined as an arithmetic average or geometric average:

$$X^A = \frac{1}{m} \sum_{j=1}^{m} X_{jT/m}, \quad X^G = \left( \prod_{j=1}^{m} X_{jT/m} \right)^{1/m},$$

where the averaging frequency implied by $m$ can be defined in various ways, such as daily, weekly, etc.

The average price of the underlying assets can then be used in the time $T$ payoff function as the strike price or the reference price. For example, **Asian strike** puts or calls are defined with the standard reference price of $X_T$, but with strike $K = X^A$ or $K = X^G$. **Asian price** puts and calls are defined in terms of the reference price of $X^A$ or $X^G$, and fixed strike price of $K$.

2. **Lookback Options:** These put and call options reflect the maximum or minimum of underlying asset prices:

$$X^{\text{max}} = \max_j \{X_{jT/m}\}; \quad X^{\text{min}} = \min_j \{X_{jT/m}\},$$

with $m$ again defined in various ways. Such options are again labelled based on whether the maximum or minimum is used as the reference price or strike price.

A **floating lookback call** is defined with a reference price of $X_T$ and strike $K = X^{\text{min}}$, while a **floating lookback put** is defined with
reference $X_T$ and strike $K = X^{\text{max}}$. Analogously, a fixed lookback call is defined with reference price $X^{\text{max}}$ and fixed strike $K$, while a fixed lookback put is defined with reference $X^{\text{min}}$ and fixed strike $K$. Thus in all cases, the option is defined to maximize the potential payoffs to the long position.

3. Barrier Options: In one respect, barrier put and call options have conventional payoff functions defined at time $T$ with the usual formulas and a fixed strike $K$. What makes barrier options path dependent is that the applicability of this payoff function depends on whether the price $X_t$ for $t < T$ reaches a given barrier $B$. These options are classified into knock-in options, for which the payoff function become applicable only if the barrier is reached, and knock-out options, for which the payoff function is cancelled if the barrier is reached.

Knock-in options are then categorized as down-and-in, or up-and-in, meaning respectively that the barrier $B$ is attained from above or below, and is applicable respectively when $X_0 > B$ or $X_0 < B$. Correspondingly, one has the categories of down-and-out, or up-and-out for the knock-out options. Each of these 4 types of barrier options can then be defined as puts or calls.

Remark 9.38 (Limiting Path-Based Price) It is certainly tempting to speculate about the "limit" of the price $O_0^{(n)} [X_0]$ in 9.52 as $n \to \infty$. As for the price of standard European options, $O_0^{(n)} [X_0]$ equals the present value of the expectation of $O_T [(X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{n\Delta t})]$, but now with probabilities assigned to each path $(X_0, X_{\Delta t}, X_{2\Delta t}, \ldots, X_{n\Delta t})$ as defined above. But here, to the extent that these random $n$-vectors converge weakly to some random vector $X$, it is clear that such $X$ is infinite dimensional and we do not have the tools to contemplate assigning a probability measure to such a space. This may seem surprising given chapter 9 of book 1 which develops infinite dimensional probability spaces, but note that this chapter contemplated spaces with countably many component variables which could be ordered.

In the current context, while it could be argued that we "only" need a probability structure for prices $X_t$ for $t = kX_0/n$ for all $n$ and $1 \leq k \leq n$, of which there are countably many, this collection of prices has no natural ordering which would make the book 1 construction feasible. If ordered in some arbitrary way, we would lose the natural connection between paths and probabilities. In book 7 we will return to such constructions, taking a more general, continuous time point of view.
Once we have such a probability space, we could investigate the extent to which $(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{n\Delta t}) \Rightarrow X$ implies that $O_T[(X_0, X_{\Delta t}, X_{2\Delta t}, ..., X_{n\Delta t})] \Rightarrow O_T[X]$ for Borel measurable functions $O_T$ say, and then contemplate conditions on $O_T$ that ensure that expectations converge, and thus $O^{(n)}_0[X_0] \rightarrow O^{(\infty)}_0[X_0]$. In this analysis, we might even imagine a representation for $O^{(\infty)}_0[X_0]$.

But we are way ahead of ourselves. At the moment we can only say that if the payoff function is bounded, say $|O_T| \leq c$, then for all $n$:

$$\left| O^{(n)}_0[X_0] \right| \leq c e^{-rT},$$

since $\sum_{2^n} q^j (1-q)^{n-j} = 1$ by definition. As a bounded sequence, it follows that there exists a subsequence so that

$$O^{(n_k)}_0[X_0] \rightarrow O^{(\infty)}_0[X_0]$$

for some $O^{(\infty)}_0[X_0]$. However, we cannot then assert that this limit is unique, nor provide a representation for its value.

### 9.5.2 Approximate Path-Based Pricing

The formula in 9.52 provides an exact valuation of the replicating portfolio for path dependent European options under the assumption of binomial movements in the price of the underlying asset. It is apparent that this exact approach becomes difficult for $n$ large since the summation requires $2^n$ terms, while for vanilla options, the summation reflects only $n+1$ terms. However, we expect that if a large sampling of binomial paths is generated, the distribution of such paths and their payoffs will be close to that predicted in theory, and hence the price of the option based on these sampled paths will be "close" to the theoretical price.

Of course a given path dependent payoff function may well group paths into far less than $2^n$ payoff "buckets," since for example many will result in no payoff as is also true for vanilla options. But we will ignore these special cases in the general development for notational simplicity. The model below will then be generally applicable since it will not be assumed in the derivation that different paths produce different payoffs.

Thus we focus on the distribution of paths generated by the model and not on the distribution of payoffs. The distribution of paths is governed by the multinomial distribution, which is developed first.
9.5 BINOMIAL PRICING OF PATH DEPENDENT EUROPEAN OPTIONS

Multinomial Distribution

Generalizing the binomial theorem in 1.8 of book 4, one obtains the so-called multinomial theorem:

\[
\left( \sum_{i=1}^{R} a_i \right)^n = \sum_{n_1,n_2,...,n_R} \frac{n!}{n_1! n_2! ... n_R!} a_1^{n_1} a_2^{n_2} ... a_R^{n_R}, \tag{9.55}
\]

where this summation is over all distinct \( R \)-tuples \((n_1,n_2,..n_R)\) with \( n_j \geq 0 \) and \( \sum_{j=1}^{R} n_j = n \). The combinatorial coefficient \( n!/(n_1!n_2!...n_R!) \) is seen to represent the number of times the given factor \( a_1^{n_1} a_2^{n_2} ... a_R^{n_R} \) arises in this product. This follows because given any \( n \) terms \( \{a_{ij}\}_{j=1}^{n} \), where "any" means allowing for repetitions, there are \( n! \) possible orderings, while the divisions by the \( n_k! \) factors eliminate the multiple counts associated with the subsets where \( a_{ij} = a_k \).

Given \( \{p_j\}_{j=1}^{R} \) with \( 0 < p_j < 1 \) and \( \sum_{j=1}^{R} p_j = 1 \), and fixed \( n \in \mathbb{N} \), the multinomial probability function or density function \( f_M \) with parameters \( \{p_j\}_{j=1}^{R} \) and \( n \) is defined on every integer \( R \)-tuple \((n_1,n_2,..,n_R)\) with \( n_j \geq 0 \) and \( \sum n_j = n \) by:

\[
f_M(n_1,n_2,...,n_R) = \frac{n! p_1^{n_1} p_2^{n_2} ... p_R^{n_R}}{n_1! n_2! ... n_R!}, \tag{9.56}
\]

This is indeed a probability function since \( \sum_{j=1}^{R} p_j = 1 \), and thus by the multinomial theorem in 9.55:

\[
\sum_{n_1,n_2,...,n_R} f_M(n_1,n_2,...,n_R) = 1,
\]

where the summation is over all \( R \)-tuples \((n_1,n_2,..,n_R)\) as defined above.

Example 9.39 There are a number of applications for this distribution.

1. At a country fair, a girl with \( n \) balls is throwing down-field at \( R-1 \) baskets of different sizes with probability \( p_j \) of hitting the \( j \)th basket, and probability \( p_R = 1 - \sum_{j=1}^{R-1} p_j \) of missing them all. The sample space is then the collection of all \( R \)-tuples of results, where each \( n_j \) denotes the number of balls entering the respective basket, where \( j = R \) denotes the number hitting the ground.

2. Recall the binomial model and assume that \( M \) sequences of \( n \) coin flips are to be generated with probability of heads \( q \) given in 9.8, 9.11
or 9.12. One question could be, for any non-negative \((n + 1)\)-tuple \((M_0, M_1, M_2, \ldots, M_n)\) with \(\sum M_j = M\), what is the probability that exactly \(M_j\) sequences will have \(j\) Heads for all \(j\)?

This is the model for vanilla options because payoff buckets are defined by the final underlying asset price, which in turn is defined by the number of \(u(\Delta t)\) returns which like "heads," occur with probability \(q\). In this application, the probability of a path ending in the \(j\)-heads bucket is given by \(p_j = \binom{n}{j} q^j (1 - q)^{n-j}\), \(j = 0, 1, \ldots, n\). From 9.56 it then follows that:

\[
f_M(M_0, M_1, \ldots, M_n) = \frac{(n + 1)! P_0^{M_0} P_1^{M_1} P_2^{M_2} \cdots P_n^{M_n}}{M_0! M_1! M_2! \ldots M_n!}.
\]

3. Generalizing model 2 for the path dependent option valuation, assume that \(M\) sequences of \(n\) coin flips are to be generated, with heads probability \(q\) given again in 9.8, 9.11 or 9.12. The question now becomes, for any non-negative \(2^n\)-tuple \((M_1, M_2, \ldots, M_{2^n})\) with \(\sum M_j = M\), what is the probability that exactly \(M_j\) sequences will equal the \(j\)th of \(2^n\) possible sequences, assuming these are ordered in some manner?

This is the model for the most general path dependent option because the payoff buckets are potentially defined by the entire sequence of asset prices. This sequence in turn is defined by the exact order of the sequence of \(u(\Delta t)\) and \(d(\Delta t)\) returns, which occur with respective probabilities \(q\) and \(1 - q\). In this application the probability of a path equaling the \(j\)th sequence is given by \(p_j = q^{k_j} (1 - q)^{n-k_j}\), where \(k_j\) denotes the number of \(u(\Delta t)\) returns in the \(j\)th sequence. Thus by 9.56 it then follows that with \(N = 2^n\):

\[
f_M(M_1, \ldots, M_N) = \frac{N! P_1^{M_1} P_2^{M_2} \cdots P_N^{M_N}}{M_1! M_2! \ldots M_N!}.
\]

Some properties of the multinomial distribution needed below and assigned as an exercise are summarized in the following proposition. See also section 7.6.4 of Reitano (2010).

**Proposition 9.40** Given the multinomial probability density with parameters \(\{p_j\}_{j=1}^{R}\) and \(n\) defined in 9.56:

1. For any \(j\), the marginal probability function \(f(n_j)\) is binomial with parameters \(p_j\) and \(n\):

\[
f(n_j) = \binom{n}{n_j} p_j^{n_j} (1 - p_j)^{n-n_j}.
\]
and hence
\[
E[N_j] = np_j, \quad \text{Var}[N_j] = np_j(1 - p_j).
\]

2. For any \(j \neq k\), the marginal probability function \(f(n_j, n_k)\) is multinomial with parameters \(\{p_j, p_k, 1 - p_j - p_k\}\) and \(n\):
\[
f(n_j, n_k) = \frac{n! p_j^{n_j} p_k^{n_k} (1 - p_j - p_k)^{n - n_j - n_k}}{n_j! n_k! (n - n_j - n_k)!},
\]
and hence
\[
E[N_j N_k] = n(n - 1)p_j p_k, \quad \text{Cov}[N_j, N_k] = -np_j p_k.
\]

**Proof.** Hint: These marginal probability functions are obtained by appropriately manipulating the summation over all other \(n_i\)-variates and applying 9.55. \(\blacksquare\)

**Approximate Path-Based Pricing of European Options**

In this section we address the European option pricing result suppressing the distinction between vanilla and path-dependent European options. As noted above in example 9.39, both path-based models fit within a general multinomial framework by properly defining \(R\) and the probabilities \(\{p_j\}_{j=1}^{R}\). Assume then that \(M\) paths have been generated within the given model, with \(\Delta t = T/n\) the time-step used in the definition of period returns \(u(\Delta t)\) and \(d(\Delta t)\) given in 9.1, and an up-return probability \(q(\Delta t)\) given in 9.8, 9.11 or 9.12.

As noted above, \(R = n + 1\) and \(p_j = \binom{n}{j} q^j (1 - q)^{n-j}\) for \(j = 0, 1, ..., n\) for vanilla options. For the path dependent options, \(R = 2^n\) and \(p_j = q^{|k_j|} (1 - q)^{n - |k_j|}\) where \(k_j\) denotes the number of \(u(\Delta t)\) returns in the \(j\)th path sequence, again assuming an arbitrary ordering of these \(2^n\) sequences.

In either case, let \(\{P_k^{(n)}\}_{k=1}^{M}\) denote an \(M\)-sample of \(X\)-price paths, and \(\{M_j\}_{j=1}^{R}\) the number of such paths that fall into each of the \(R\) path buckets. Define the **sample option price** random variable \(O_M^{(n)}(X_0)\):

\[
O_M^{(n)}(X_0) = \frac{e^{-rT}}{M} \sum_{k=1}^{M} O_T\left(P_k^{(n)}\right).
\]  \hspace{1cm} (9.57)

Equivalently:

\[
O_M^{(n)}(X_0) = \frac{e^{-rT}}{M} \sum_{j=1}^{R} M_j O_T\left(P_{k_j}^{(n)}\right),
\]  \hspace{1cm} (9.58)
where \( M_j \) denotes the number of sample paths that equal the \( j \)th path sequence, which we denote \( P^{(n)}_{k_j} \). While \( O_T \left( P^{(n)}_{k_j} \right) \) is defined on each path sequence, it is not necessarily defined differently on different sequences.

To consider \( O^{(n)}_M (X_0) \) as a random variable, we explicitly assume that \( O_T \) is a measurable function. In detail, for vanilla options we assume that \( O_T \) is a measurable function defined on \( \mathbb{R} \), since \( O_T \left( P^{(n)}_k \right) \equiv O_T \left( X^{(n)}_T \right) \), while for the path dependent case, where \( O_T \) is defined as a function of \( n \) asset prices, we assume \( O_T \) is a measurable function defined on \( \mathbb{R}^n \).

The general result follows. For background on convergence of random variables, see sections 5.2 and 8.5 of book 2.

**Proposition 9.41** With \( O^{(n)}_M (X_0) \) defined in 9.58 above:

1. The expected value of \( O^{(n)}_M (X_0) \) equals the theoretical option price:

\[
E[O^{(n)}_M (X_0)] = O^{(n)}_0 (X_0),
\]

where depending on the payoff function specified, \( O^{(n)}_0 (X_0) \) denotes the vanilla price in 9.7 or the path dependent price in 9.52.

2. If \( \text{Var} \left[ O^{(n)}_M \left( P^{(n)}_{k_j} \right) \right] < \infty \), where this variance is defined under \( \{p_j\} \)
defined above, then for any \( \epsilon > 0 \):

\[
\Pr \left[ \left| O^{(n)}_M (X_0) - O^{(n)}_0 (X_0) \right| > \epsilon \right] \to 0 \text{ as } M \to \infty.
\]

In other words, \( O^{(n)}_M (X_0) \) **converges weakly** to \( O^{(n)}_0 (X_0) \) as \( M \to \infty \), denoted \( O^{(n)}_M (X_0) \Rightarrow O^{(n)}_0 (X_0) \).

3. If \( \text{Var} \left[ O^{(n)}_M \left( P^{(n)}_{k_j} \right) \right] < \infty \) defined in 2, then as \( M \to \infty \):

\[
O^{(n)}_M (X_0) \to_{a.s.} O^{(n)}_0 (X_0).
\]

In other words, \( O^{(n)}_M (X_0) \) **converges with probability 1** to \( O^{(n)}_0 (X_0) \) as \( M \to \infty \).

**Proof.** For 1, proposition 9.40 above states that \( \{M_j\} \) are separately binomial, and so:

\[
E[O^{(n)}_M (X_0)] = e^{-rT} \sum_{j=1}^{R} E \left[ \frac{M_j}{M} \right] O_T \left( P^{(n)}_{k_j} \right)
= e^{-rT} \sum_{j=1}^{R} p_j O_T \left( P^{(n)}_{k_j} \right).
\]
Recalling the definition of $R$ and $p_j$ in the respective models, this expectation equals the price in 9.7 or 9.52.

To demonstrate 2, we can simply prove 3 since convergence with probability 1 implies weak convergence by proposition 5.21 of book 2. Additionally, a direct proof using the Chebyshev inequality of proposition 3.32 of book 4 is assigned as an exercise based on a calculation of the variance of $O_M^{(n)}(X_0)$ using proposition 9.40:

$$\text{Var} \left[ O_M^{(n)}(X_0) \right] = \frac{e^{-2rT}}{M} \text{Var} \left[ O_T \left( P_k^{(n)} \right) \right].$$

(9.62)

For 3, define the random variable $Y_k = e^{-rT}O_T \left( P_k^{(n)} \right)$, the option price on the $k$th path. Then $\{Y_k\}_{k=1}^\infty$ are independent random variables by proposition 3.56 of book 2 since $O_T$ is measurable and $\{P_k^{(n)}\}_{k=1}^\infty$ are independent random variables on $\mathbb{R}^n$. Also, $E[Y_k] = O_0^{(n)}(X_0)$ from 1, and $\text{Var}[Y_k] = e^{-2rT}\text{Var} \left[ O_T \left( P_k^{(n)} \right) \right] < \infty$ as noted in 2 using $M = 1$, and so:

$$\sum_{k=1}^\infty \text{Var}[Y_k]/k^2 < \infty.$$

By the strong law of large numbers in proposition 5.45 of book 4:

$$\frac{1}{M} \sum_{k=1}^M Y_k - O_0^{(n)}(X_0) \to_{a.s.} 0,$$

and 9.61 follows by definition. ■

**Remark 9.42** 1. The price of a European derivative security obtained with sampling in the path-based model above will contain **two types of error** compared to the theoretically correct price $O_0(X_0)$ (to the extent this exists in the general case, recalling remark 9.38). Denoting by $O_M^{(n)}(X_0)$ the price obtained with $M$ paths and time-steps of $\Delta t \equiv T/n$:

(a) **Discretization error:** This is identical to that produced by the underlying binomial lattice calculation and depends on $\Delta t$ or equivalently $n$ since $\Delta t = T/n$. This error is defined as:

$$\epsilon^D(\Delta t) = O_0(X_0) - O_0^{(n)}(X_0),$$

and represents the error between the limiting derivative price as $n \to \infty : O_0(X_0) \equiv O_0^{(\infty)}(X_0)$, and the binomial lattice value
or complete path-based value $O_0^{(n)}(X_0)$. This error is caused by discretizing time and asset price movements and the effect of this discretization on the ultimate distribution of derivative payoff values.

(b) **Estimation error:** This is defined as:

$$\varepsilon^E(\Delta t) = O_0^{(n)}(X_0) - O_M^{(n)}(X_0),$$

and represents the error between the path-based option price estimate, $O_M^{(n)}(X_0)$, and the binomial lattice value or complete path-based value $O_0^{(n)}(X_0)$. This error is caused by sampling error, realized here in that for any given sample, $M_j/M \neq p_j$ for every $j$, and hence $O_M^{(n)}(X_0) \neq O_0^{(n)}(X_0)$.

2. As was noted in the above proof, estimation error decreases with $1/M$ by Chebyshev’s inequality:

$$\Pr \left[ \left| O_0^{(n)}(X_0) - O_M^{(n)}(X_0) \right| > \epsilon \right] < \frac{e^{-2\epsilon T}}{M^2 \text{Var}\left[ O_T\left(P_k^{(n)}\right)\right]}.$$

Consequently, we can in theory choose $\epsilon_M \to 0$ in such a way that $M\epsilon_M^2 \to \infty$ and thereby ensure that as $M \to \infty$, estimation error is theoretically eliminated. In practice, however, this will be a slow and painful process since in order for $M\epsilon_M^2 \to \infty$, it will be necessary to have either $\epsilon_M \to 0$ quite slowly, and/or have $M\epsilon_M^2 \to \infty$ quite slowly.

For example, if $\epsilon_M = M^{-a}$ for $0 < a < \frac{1}{2}$, both objectives are achieved, where $a \sim \frac{1}{4}$ provides faster $\epsilon_M \to 0$ and slower $M\epsilon_M^2 \to \infty$, while $a \sim 0$ does the opposite.
References

I have listed below a number of textbook references for the mathematics and finance presented in this series of books. All provide both theoretical and applied materials in their respective areas that are beyond those developed here and are worth pursuing by those interested in gaining a greater depth or breadth of knowledge. This list is by no means complete and is intended only as a guide to further study. In addition, these references include various published research papers if they have been identified in this book’s chapters.

The reader will no doubt observe that the mathematics references are somewhat older than the finance references and upon web searching will find that several of the older texts in each category have been updated to newer editions, sometimes with additional authors. Since I own and use the editions below, I decided to present these editions rather than reference the newer editions which I have not reviewed. As many of these older texts are considered "classics", they are also likely to be found in university and other libraries.

That said, there are undoubtedly many very good new texts by both new and established authors with similar titles that are also worth investigating. One that I will at the risk of immodesty recommend for more introductory materials on mathematics, probability theory and finance is:


**Topology, Measure, Integration, Linear Algebra**


**Probability Theory & Stochastic Processes**


REFERENCES


Finance Applications


Research Papers for Book 6


Index

$\mu$-continuity set, 70

asset model
  real world probability, 255
  risk neutral probability, 257, 261

beta distribution, 41, 142, 194
binomial distribution
  moments, 140, 192
binomial lattice, 257
Black-Scholes-Merton
  European option formulas, 278
  partial differential equation, 285
Black-Scholes-Merton
  option pricing formula, 275
Bochner, Salomon
  Bochner’s theorem, 162
Borel, Émile
  Borel measure, 1
  Borel sigma algebra, 1
boundary
  of a set in 70
  of a set in 74

Cauchy distribution, 142, 194, 197
Cauchy, Augustin Louis
  Cauchy distribution, 38
Cavalieri, Bonaventura
  Cavalieri’s principle, 28
characteristic function
  of a distribution, 133
Cholesky, André-Louis
  Cholesky decomposition, 52
  complex conjugate, 162
compound Poisson distribution, 199
conditional distribution function, 20
continuous from above, 2
convergence in distribution
  of random vectors, 74
convergence in law
  of random vectors, 74
convolution
  of Borel measures, 31
  of distribution functions, 33
correlation, 57
covariance, 48
Cox-Ross-Rubinstein
  binomial lattice model for option
  pricing, 279
Cramér-Wold device, 85, 159, 188
cumulative distribution function (c.d.f.)
  of a random vector, 2
cylinder set
  infinite dimensional product space, 211
degenerate distribution, 197
density function, 95
distribution function (d.f.)
  of a random vector, 2
Distribution functions
  beta, 142, 194
  binomial, 140, 192
  Cauchy, 142, 194, 197
  chi-squared, 194
compound Poisson, 199
continuous uniform, 141, 193
degenerate, 197
exponential, 142, 193, 197
gamma, 142, 193, 197
gerometric, 140, 192, 198
lognormal, 144
multivariate normal, 143, 198
negative binomial, 141, 192, 198
normal, 143, 194, 198
Poisson, 141, 193, 197
rectangular, 140, 191
Student T, 98

Distributions
beta, 41
Cauchy, 38
gamma, 41
multivariate normal, 43, 44, 46, 156
normal, 44
dot product
inner product, 85
exponential distribution, 142, 193, 197

Feller, William
Lindeberg-Feller central limit theorem, 185
Fourier transform
of an integrable function, 139
Fourier-Stieltjes transform
inversion formula, 152
of a finite Borel measure, 138

gamma distribution, 34, 41, 142, 193, 197
Gauss, Carl Friedrich
normal distribution, 44
geometric distribution, 140, 192, 198
Gosset, William Sealy
Student’s T distribution, 98

Hermite, Charles
Hermitian matrix, 164
inner product
dot product, 85
interior
of a set in 74
Isserlis, John
Isserlis’ theorem, 56
Jacobi, Carl Gustav Jacob
Jacobian matrix
determinant, 39
Khintchine, Alexandre
Lévy-Khintchine Representation theorem, 207
Lévy, Paul
Lévy’s continuity theorem, 147, 157
Lévy, Paul
Lévy-Khintchine Representation theorem, 207
Lagrange, Joseph-Louis
Lagrange form of the Taylor remainder, 90
law of one price, 253
Lindeberg, J. W. (Jarl Waldemar)
Lindeberg’s condition, 180
log-X distribution, 209
lognormal distribution, 144, 209
Lyapunov, Aleksandr
Lyapunov’s condition, 186
Mann, Henry
Mann-Wald theorem, 84
marginal distribution function, 15
measure
induced by a transformation, 81
multinomial distribution, 299
INDEX

multinomial theorem, 299

n-increasing, 2

negative binomial distribution, 141, 192, 198

Newton, Isaac
  Newton’s generalized binomial formula, 199

nonnegative definite
  function, 163

normal distribution, 143, 194, 198
  multivariate, 49, 143, 198
  multivariate normal, 43

options
  path dependent European, 293
  vanilla puts and calls, 275

Poisson, Siméon-Denis
  Poisson distribution, 141, 193, 197

positive definite
  matrix, 49

positive semidefinite
  function, 163
  matrix, 46, 156

product space
  infinite dimensional, 211
    cylinder, 211
    product measure, 212

Prokhorov, Yuri Vasilyevich
  Prokhorov’s theorem, 100

random variable, 1
random variable (r.v.)
  independent, 18
random variable sequence
  convergence in distribution, 83
  convergence in probability, 83
  convergence with probability 1, 83
random vector, 1, 2
random vector sequence
  convergence in distribution, 74
rectangular distribution
  moments of discrete, 140, 191
risk-free arbitrage, 254

sample space, 59
  random sample, 59
  statistic, 60

Scheffé, Henry
  Scheffé’s theorem, 94

sign function
  signum function, 153

Slutsky, Evgeny "Eugen"
  Slutsky’s theorem, 87

statistic, 60
  sample mean, 60
  sample variance, 60
  unbiased, 61

Student’s T distribution, 98

Taylor, Brook
  Taylor series, 90, 92

tight
  sequence of probability measures, 93

transformation
  induced measure, 81

triangular array
  of random variables, 179

uniform distribution
  continuous, 141, 193
  uniformly integrable, 110

variance, 48

Wald, Abraham
  Mann-Wald theorem, 84

weak convergence
  of distribution functions, 74
  of probability measures, 69, 74