Jean-Pierre Danthine and John B. Donaldson

INTERMEDIATE FINANCIAL THEORY

THIRD EDITION

Intermediate Financial Theory

Intermediate Financial Theory

Third Edition

Jean-Pierre Danthine
Swiss National Bank
Bundesplatz 1
Bern, Switzerland

John B Donaldson
Columbia Business School
New York, NY

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA

First edition 2001
The first edition of this book was published by Pearson Education, Inc.

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Catalog Number
A catalog record for this book is available from the Library of Congress

ISBN-13: 978-0-12-386549-6

Preface

For all the suffering that it has caused, the global financial crisis provides a unique opportunity to review what we know or thought we knew about finance. It will challenge and enliven the teaching of finance for years to come. The third edition of Intermediate Financial Theory is affected accordingly. While our own goals for the text have not changed, many new topics have been added and many examples have been taken from recent experience. The task of reviewing the entire material in light of the financial crisis is, however, a work in progress and one that cannot be adequately undertaken within the context of periodic revisions of a textbook of reasonable length. Accordingly, it will be pursued on an ongoing basis on the text's website.

The third edition of Intermediate Financial Theory features 2 entirely new chapters and very substantial revisions to 11 more. With respect to the latter changes, there is greater emphasis on "behavioral finance" and many of the latest developments in portfolio theory are fully featured. The chapter on the consumption capital asset pricing model has been similarly expanded and brought to the theoretical frontier. Integral to the print version of the text are four Web Chapters. The intent of these chapters is to expand on the basic ideas presented in the text in ways that link them more directly to applied practice. Our review and assessment of the recent "financial crisis" is a particular case in point. Lastly, the third edition attempts to strengthen the link between theory and "the data"; that is, of evaluating a particular theory as regards its ability to replicate the associated data patterns, the so-called financial stylized facts.

While the market for financial textbooks is crowded at both the introductory and doctoral levels, it remains much thinner at the intermediate level. Teaching opportunities at this level, however, have greatly increased with the advent of masters of science programs in finance (master's degree programs in computational finance, in mathematical finance, and the like) and the continuing demand for higher-level courses in MBA programs.

The Master in Banking and Finance Program at the University of Lausanne, which admitted its first class in the fall of 1993 is a program of the aforementioned type. One of the first such programs of its kind in Europe, its objective is to provide advanced training to finance specialists in the context of a 1-year theory-based degree program. In designing the curriculum, it was felt that students should be exposed to an integrated course that would

introduce the full range of topics typically covered in financial economics courses at the doctoral level. Such exposure could, however, ignore the detailed proofs and arguments and concentrate on the larger set of issues and concepts to which any advanced practitioner should be exposed. This latest edition of our text retains this philosophy.

Accordingly, our ambition for this third edition is unchanged from earlier ones: first to review rigorously and concisely the main themes of financial economics (those that students should have encountered in prior courses) and, second, to introduce a number of *frontier* ideas of importance for the evolution of the discipline and of relevance from a practitioner's perspective. We want our readers not only to be at ease with the main concepts of standard finance (MPT, CAPM, etc.) but also to be aware of the principal new ideas that have marked the recent evolution of the discipline. Contrary to introductory texts, we aim at depth and rigor; contrary to higher-level texts, we do not emphasize generality. Whenever an idea can be conveyed through an example, this is the approach we choose. We continue to ignore proofs and detailed technical matters unless a reasonable understanding of the related concept mandates their inclusion.

Intermediate Financial Theory is intended primarily for master level students with a professional orientation, a good quantitative background, and who have completed at least one introductory finance course (or have read the corresponding text). As such, the book is targeted especially for masters students in finance, while remaining accessible and appropriate for an advanced MBA class in financial economics, one with the objective of introducing students to the precise modeling of many of the concepts discussed in their capital markets and corporate finance classes. In addition, we believe the book can be a useful reference for doctoral candidates in finance, particularly those whose lack of prior background might prevent them from drawing the full benefits of the very abstract material typically covered at that level. Finally, we hope it will be a useful refresher for well-trained practitioners. Although the mathematical requirements of the book are not great, some confidence in the use of calculus as well as matrix algebra is helpful.

In preparing this third edition, we maintain our earlier emphasis on the valuation of risky cash flows. This subject—asset pricing—constitutes the main focus of modern finance, and its shortcomings have come powerfully to the fore in the recent financial crisis. We also emphasize the distinction between valuation procedures that rely on general equilibrium principles and those based on arbitrage considerations.

At present there are four Web Chapters that are available to readers. These represent substantial extensions of ideas introduced in the print version of the book. Web Chapter A translates the Consumption CAPM model into a fully dynamic production setting so that the mutual influence of the financial markets on and by the macroeconomy can be made more explicit. Web Chapter B goes beyond using the martingale measure to price options to an exploration of the use of options concepts in the evaluation of complex securities, real

investment projects, and strategies of portfolio management, while Web Chapter C returns to a more general treatment of differential information. Lastly, Web Chapter D explores the origins and evolution of the recent financial crisis and does so in a way that is intended to assess the strengths and weaknesses of the current state of financial theory. By placing these chapters on the Web, we can more easily update and add to the material presented therein. More chapters will be added in future years.

Over the years, we have benefited from numerous discussions with colleagues over issues related to the material included in this book. We are especially grateful to Rajnish Mehra, Arizona State University; Elmar Mertens, IMF; Paolo Siconolfi, Lars Lochstoer, Kent Daniel, Andrew Ang and Tano Santos, all of Columbia Business School; and Erwan Morellec, University of Lausanne, the latter for his contribution to the corporate finance review of Chapter 2. We are also indebted to several generations of teaching assistants including François Christen, Philippe Gilliard, Tomas Hricko, Aydin Akgun, Paul Ehling, Oleksandra Hubal, and Lukas Schmid—and of MBF students at the University of Lausanne—who have participated in the shaping up of this material. Teaching and research assistants from the "other side of the Atlantic" have been hugely helpful as well, most especially, J.K. Auh, Mattia Landoni, and Zhongjin Lu. Their questions, corrections, and comments have led to a continuous questioning of the approach we have adopted and have dramatically increased the usefulness of this text. Finally, we reiterate our thanks to the Fondation du 450ème of the University of Lausanne for providing "seed financing" for this project.

Jean-Pierre Danthine
Bern, Switzerland

John B. Donaldson
New York, NY

N'estime l'argent ni plus ni moins qu'il ne vaut:
c'est un bon serviteur et un mauvais maître
(Value money neither more nor less than it is worth:
It is a good servant and a bad master)

Alexandre Dumas, fils, *La Dame aux Camélias* (Préface)

Cover
by Renée-Paule Danthine
« Petite fugue »; 2011 Oil and pastel on Japanese paper

Dedication

J.B. Donaldson wishes to dedicate his involvement in this book enterprise to his parents, Brown and Rachel Donaldson, his wife Charissa Asbury, Mario Gabelli who offered financial support through the provision of his academic chair and, most especially, his physicians, Kareem M. Abu-Elmagd, MD, and Guilherme Costa, MD, of the University of Pittsburgh Medical Center. In various ways, all these persons were vital to the completion of the book.

CHAPTER 1

On the Role of Financial Markets and Institutions

Chapter Outline

1.1 Finance: The Time Dimension

Why do we need financial markets and institutions? We choose to address this question as our introduction to this text on financial theory. In doing so, we touch on some of the most difficult issues in finance and introduce concepts that will eventually require extensive development. Our purpose here is to phrase this question as an appropriate background for the study of the more technical issues that will occupy us at length. We also want to introduce some important elements of the necessary terminology. We ask the reader's patience as most of the sometimes difficult material introduced here will be taken up in more detail in the following chapters.

Fundamentally, a financial system is a set of institutions and markets permitting the exchange of contracts and the provision of services for the purpose of allowing the income and consumption streams of economic agents to be desynchronized—i.e., made less similar. It can, in fact, be argued that indeed the *primary* function of the financial system is to

Intermediate Financial Theory.

permit such desynchronization. There are two dimensions to this function: the time dimension and the risk dimension. Let us start with time. Why is it useful to disassociate consumption and income across time? Two reasons come immediately to mind. First, and somewhat trivially, income is typically received at discrete dates, say monthly, while it is customary to wish to consume continuously (i.e., every day).

Second, and more importantly, consumption spending defines a *standard of living*, and most individuals find it difficult to alter their standard of living from month to month or even from year to year. There is a general, if not universal, desire for a *smooth* consumption stream. Because it deeply affects everyone, the most important manifestation of this desire is the need to save (consumption smaller than income) for retirement so as to permit a consumption stream in excess of income (dissaving) after retirement begins. The *life-cycle* patterns of income generation and consumption spending are not identical, and the latter must be created from the former. The same considerations apply to shorter horizons. Seasonal patterns of consumption and income, for example, need not be identical. Certain individuals (car salespersons, department store salespersons, construction workers) may experience variations in income arising from seasonal events (e.g., most new cars are purchased in the spring and summer; construction activity is much reduced in winter), which they do not like to see transmitted to their ability to consume. There is also the problem created by temporary layoffs due to variation in aggregate economic activity that we refer to as business cycle fluctuations. While they are temporarily laid off and without substantial income, workers do not want their family's consumption to be severely reduced (Box 1.1).

Furthermore, and this is quite crucial for the growth process, some people—entrepreneurs, in particular—are willing to accept a relatively small income (but not necessarily

BOX 1.1 Representing Preference for Smoothness

The preference for a smooth consumption stream has a natural counterpart in the form of the utility function, $U(\ )$, which is typically used to represent the relative benefit a consumer receives from a specific consumption bundle. Suppose the representative individual consumes a single consumption good (or a basket of goods) in each of two periods, now and tomorrow. Let c_1 denote today's consumption level and c_2 tomorrow's, and let $U(c_1) + U(c_2)$ represent the level of utility (benefit) obtained from a given consumption stream (c_1, c_2).

Preference for consumption smoothness must mean, for instance, that the consumption stream $(c_1,c_2) = (4,4)$ is preferred to the alternative $(c_1, c_2) = (3, 5)$, or

$$U(4) + U(4) > U(3) + U(5)$$

Dividing both sides of the inequality by 2, this implies

$$U(4) > \frac{1}{2}U(3) + \frac{1}{2}U(5)$$

(*Continued*)

BOX 1.1 Representing Preference for Smoothness (Continued)

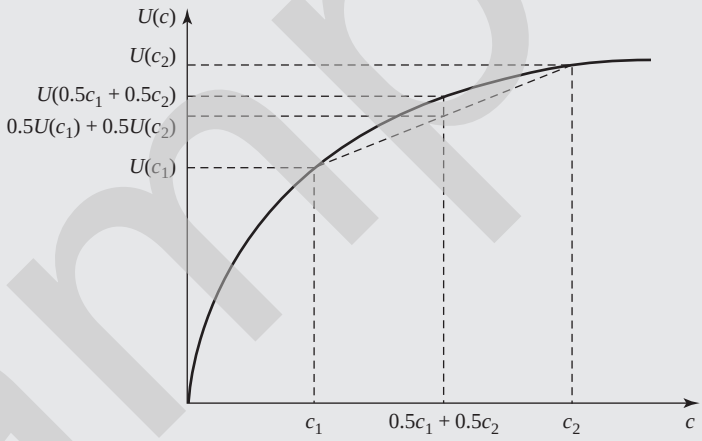


Figure 1.1
A strictly concave utility representation.

As shown in Figure 1.1, when generalized to all possible alternative consumption pairs, this property implies that the function $U(\cdot)$ has the rounded shape that we associate with the term *strict concavity*.

consumption!) for an initial period of time in exchange for the prospect of high returns (and presumably high income) in the future. They are operating a sort of arbitrage over time. This does not disprove their desire for smooth consumption; rather, they see opportunities that lead them to accept what is formally a low-income level initially against the prospect of a much higher income level later (followed by a zero income level when they retire). They are investors who, typically, do not have enough liquid assets to finance their projects and, as a result, need to raise capital by borrowing or by selling shares.

Indeed, the first key element in finance is **time.** In a timeless world, there would be no assets, no financial transactions (although money would be used, it would have only a transaction function), and no financial markets or institutions. The very notion of a security (a financial contract) implies a time dimension.

Asset holding permits the desynchronization of consumption and income streams. The peasant putting aside seeds, the miser burying his gold, or the grandmother putting a few hundred dollar bills under her mattress are all desynchronizing their consumption and income, and in doing so, presumably seeking a higher level of well-being for themselves. A fully developed financial system should also have the property of fulfilling this same function *efficiently*. By that we mean that the financial system should provide versatile and

diverse instruments to accommodate the widely differing needs of savers and borrowers insofar as size (many small lenders, a few big borrowers), timing, and maturity of loans (how to finance long-term projects with short-term money), and the liquidity characteristics of instruments (precautionary saving cannot be tied up permanently). In other words, the elements composing the financial system should aim at *matching* the diverse financing needs of different economic agents as perfectly as possible.

1.2 Desynchronization: The Risk Dimension

We have argued that time is of the essence in finance. When we talk of the importance of time in economic decisions, we think in particular of the relevance of choices involving the present versus the future. But the future is, by its very nature, uncertain: financial decisions with implications (payouts) in the future are necessarily risky. Time and risk are inseparable. This is why **risk** is the second key word in finance.

For the moment, let us compress the time dimension into the setting of a "Now and Then" (present versus future) economy. The typical individual is motivated by the desire to smooth consumption between "Now" and "Then." This implies a desire to identify consumption opportunities that are as similar as possible among the different possibilities that may arise "Then." In other words, *ceteris paribus*—most individuals would like to guarantee their family the same standard of living whatever events transpire tomorrow: whether they are sick or healthy, unemployed or working, confronted with bright or poor investment opportunities, fortunate or hit by unfavorable accidental events.[1] This characteristic of preferences is generally described as "aversion to risk."

A productive way to start thinking about this issue is to introduce the notion of *states of nature* or *states of the world*. A state of nature is a complete description of a possible scenario for the future across all the dimensions relevant for the problem at hand. In a "Now and Then" economy, all possible future events can be represented by an exhaustive list of states of nature. We can thus extend our former argument for smoothing consumption across time by noting that the typical "risk averse" individual would also like to experience similar consumption levels across all future states of nature, whether good or bad.

An efficient financial system offers ways for savers to reduce or eliminate, at a fair price, the risks they are not willing to bear (risk shifting). Fire insurance contracts eliminate the financial risk of fire, while put options contracts can prevent the loss in wealth associated with a stock's price declining below a predetermined level, to mention but two examples. The financial system also makes it possible to obtain relatively safe aggregate returns from a large number of small, relatively risky investments. This is the process of diversification. By permitting economic agents to *diversify*, to *insure*, and to *hedge* their risks, an efficient

[1] *Ceteris paribus* is the Latin phrase for "everything else maintained equal." It is an expression commonplace in the language of economics.

financial system fulfills the function of redistributing purchasing power not only over time, but also across states of nature.[2]

1.3 The Screening and Monitoring Functions of the Financial System

The business of desynchronizing consumption from income streams across time and states of nature is often more complex than our initial description may suggest. If time implies uncertainty, uncertainty may imply not only risk, but often *asymmetric information* as well. By this term, we mean situations where the agents involved have different information, with some being potentially better informed than others. How can a saver be assured that he will be able to find a borrower with a good ability to repay—the borrower himself knows more about this, but he may not wish to reveal all he knows—or an investor (an entrepreneur or a firm) with a good project, yielding the most attractive return for him and hopefully for society as well? Again, the investor is likely to have a better understanding of the project's prospects and of his own motivation to carry it through. What do "good" and "most attractive" mean in these circumstances? Do these terms refer to the highest potential return? What about risk? What if the anticipated return is itself affected by the actions of the investors themselves (a phenomenon labeled "moral hazard")? How does one share the risks of a project in such a way that both investors and savers are willing to proceed, taking actions acceptable to both? It is the task of financial intermediaries—banks, venture capital firms, and private equity firms—to answer these questions and to do so in such a way that brings the socially beneficial projects to fruition. An efficient financial system, and the financial institutions that define it, not only assists in these information and monitoring tasks, but also provides a range of instruments (contractual arrangements) suitable for the largest number of savers and borrowers, thereby contributing to the channeling of savings toward the most efficient projects.[3]

In the words of the preeminent economist, Joseph Schumpeter (1934), "Bankers are the gatekeepers of capitalist economic development. Their strategic function is to screen potential innovators and advance the necessary purchasing power to the most promising."

[2] Both insurance and hedging are risk-reduction strategies but with one critical difference. In the case of insurance, the investor pays money—the insurance premium—to guarantee against a loss in value of some asset that he owns (a house, shares of stock). In the case of hedging, an investor adds to his portfolio, usually at very little cost, another asset (the "hedging asset") with a price pattern that is opposite to that of his original portfolio: if the original portfolio declines in value, the newly added asset increases in value by an equal and offsetting amount (this is the case of a "perfect hedge"). The opposite is true, however, if the investor's original portfolio increases in value: the hedging asset loses an equal and offsetting amount. The investor thus sacrifices potential gains to his portfolio's value in exchange for protection against losses. In the case of insurance, upward potential was not sacrificed, but the investor had to pay the premium.

[3] If the extent of the information asymmetry between buyers and sellers in a market becomes too great, the market may shut down: no trades occur. This exact event occurred at the start of the financial crisis when the investment banks that had been packaging pools of US home mortgages into mortgage-backed securities (MBS) discovered that they could find no buyers. The natural buyers of these securities had become suspicious that their quality was not as advertised. The forced sale of a nearly insolvent Bear Stearns to JPMorgan Chase and the Lehman Brothers bankruptcy ensued.

For highly risky projects, such as the creation of a new firm exploiting a new technology, venture capitalists largely provide this function today.

1.4 The Financial System and Economic Growth

The performance of the financial system matters at several levels. We shall argue that it matters for growth, that it impacts the characteristics of the business cycle, and, most importantly, that it is a significant determinant of economic welfare. We tackle growth first. Channeling funds from savers to investors efficiently is obviously important. Whenever more efficient ways are found to perform this task, society can achieve a greater increase in tomorrow's consumption for a given sacrifice in consumption today. As a result, savings becomes a more attractive alternative to current consumption, and households save more. Intuitively, more savings should lead to greater investment and thus greater future wealth. Figure 1.2 indeed suggests that, for 90 developing countries over the period 1971–1992, there was a strong positive association between saving rates and growth rates. When looked at more carefully, however, the evidence is usually not as strong.[4]

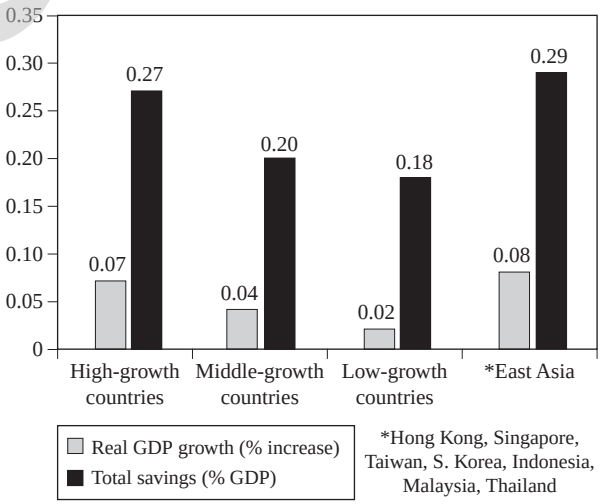


Figure 1.2
Savings and growth in 90 developing countries.

[4] In a straightforward regression in which the dependent variable is the growth rate in real per capita gross national product (GNP), the coefficient on the average fraction of real GNP represented by investment (I/Y) over the prior 5 years is positive but insignificant. Together with other results, this is interpreted as suggesting a possible reverse causation from real per capita GNP growth to investment spending. See Barro and Sala-i-Martin (1995), Chapter 12, for a full discussion. There is also a theoretically important distinction between the effects of increasing investment (savings) (as a proportion of national income) on an economy's *level* of wealth and its *growth rate*. Countries that save more will ceteris paribus be wealthier, but they need not grow more rapidly. The classic growth model of Solow (1956) illustrates this distinction. See Web Chapter A.

One important reason may be that the hypothesized link is, of course, dependent on a ceteris paribus clause: It applies only to the extent savings are invested in appropriate ways. The economic performance of the former USSR reminds us that it is not enough only to save; it is also important to invest judiciously. Historically, the investment/GDP (gross domestic product, the measure of a nation's aggregate economic output) ratio in the USSR was very high in international comparisons, suggesting the potential for very high growth rates.[5] After 1989, however, experts realized that the value of the existing stock of capital was not consistent with the former levels of investment. A great deal of the investment had been effectively wasted—in other words, allocated to poor or even worthless projects. Equal savings rates can thus lead to investments of widely differing degrees of usefulness from the viewpoint of future growth.

A more contemporary version of this same overinvestment phenomenon may be present in China. In 2012, investment in China represented 46% of output which is astonishingly high by international standards (e.g., the United States averages 15%; Switzerland averages around 20%), and a lively debate has arisen as to what fraction of China's investment will ultimately be useful.[6]

Let us go further than these general statements in the analysis of the savings and growth nexus and of the role of the financial system. Following Barro and Sala-i-Martin (1995), one can view the process of transferring funds from savers to investors in the following way.[7] The least efficient system would be one in which all investments are made by the savers themselves. This is certainly inefficient because it requires a sort of "double coincidence" of intentions: good investment ideas occurring in the mind of someone lacking past savings will not be realized. Funds that a nonentrepreneur saves would not be put to productive use. Yet, this unfortunate situation is a clear possibility if the necessary confidence in the financial system is lacking, with the consequence that savers do not entrust the system with their savings. One can thus think of circumstances where savings never enter the financial system, or where only a small fraction does. When it does, it will typically enter via some sort of depository institution. In an international setting, a similar problem arises if national savings are primarily invested abroad, a situation that may reach alarming proportions in the case of underdeveloped

[5] More precisely, GDP refers to the value, at market prices, of all final goods and services (those sold to end users) produced within a nation's geographical boundaries during a specific time period (usually a year).

[6] Excessive investment comes at the price of lower household consumption. Lee et al. (2012) estimate this loss in consumption to have averaged 4% of total Chinese output.

[7] For a broader perspective and a more systematic connection with the relevant literature on this topic, see Levine (1997).

countries.[8] Let FS/S represent, then, the fraction of aggregate savings (S) being entrusted to the financial system (FS).

At a second level, the functioning of the financial system may be more or less costly. While funds transferred from a saver to a borrower via a direct loan are immediately and fully made available to the end user, the different functions of the financial system discussed above are often best fulfilled, or sometimes can only be fulfilled, through some form of intermediation, which typically involves some cost. Let us think of these costs as administrative costs, on the one hand, and costs linked to the reserve requirements of banks, on the other. Different systems will have different operating costs in this large sense, and, as a consequence, the amount of resources transferred to investors will also vary. Let us think of BOR/FS as the ratio of funds transferred from the financial system to borrowers and entrepreneurs.

Borrowers themselves may make diverse use of the funds borrowed. Some, for example, may have pure liquidity needs (analogous to the reserve needs of depository institutions), and if the borrower is the government, it may well be borrowing for consumption! For the savings and growth nexus, the issue is how much of the borrowed funds actually result in productive investments. Let I/BOR represent the fraction of borrowed funds actually invested. Note that BOR stands for borrowed funds whether private or public. In the latter case, a key issue is what fraction of the borrowed funds are used to finance public investment as opposed to public consumption.

Finally, let EFF denote the efficiency of the investment projects undertaken in society at a given time, with EFF normalized at unity; in other words, the average investment project has EFF = 1, the below-average project has EFF $<$ 1, and conversely for the above average project (a project consisting of building a bridge leading nowhere would have an EFF = 0); K is the aggregate capital stock and Ω the depreciation rate. We may then write

$$\dot{K} = \text{EFF} \cdot I - \Omega K \tag{1.1}$$

or, multiplying and dividing I with each of the newly defined variables

$$\dot{K} = \text{EFF} \cdot (I/\text{BOR}) \cdot (\text{BOR}/\text{FS}) \cdot (\text{FS}/S) \cdot (S/Y) \cdot Y - \Omega K \tag{1.2}$$

[8] The problem is slightly different here, however. Although capital flight is a problem from the viewpoint of building up a country's domestic capital stock, the acquisition of foreign assets may be a perfectly efficient way of building a national capital stock. The effect on growth may be negative when measured in terms of GDP, but not necessarily so in terms of national income or GNP. Switzerland is an example of a rich country investing heavily abroad and deriving a substantial income flow from it. It can be argued that the growth rate of the Swiss GNP (but probably not GDP) has been enhanced rather than diminished by this fact.

where our notation is meant to emphasize that the growth of the capital stock at a given savings rate is likely to be influenced by the levels of the various ratios introduced above.[9] Let us now review how this might be the case.

One can see that a financial system performing its matching function efficiently will positively affect the savings rate (S/Y) and the fraction of savings entrusted to financial institutions (FS/S). This reflects the fact that savers can find the right savings instruments for their needs. In terms of overall services net of inconvenience, this acts like an increase in the return to the fraction of savings finding its way into the financial system. The matching function is also relevant for the I/BOR ratio. With the appropriate instruments (like flexible overnight loan facilities), a firm's cash needs are reduced and a larger fraction of borrowed money can actually be used for investment.

By offering a large and diverse set of possibilities for spreading risks (insurance and hedging), an efficient financial system will also positively influence the savings ratio (S/Y) and the FS/S ratio. Essentially this works through improved return/risk opportunities, corresponding to an improved trade-off between future and present consumption (for savings intermediated through the financial system). Furthermore, in permitting entrepreneurs with risky projects to eliminate unnecessary risks by using appropriate instruments, an efficient financial system provides, somewhat paradoxically, a better platform for undertaking riskier projects. If, on average, riskier projects are also the ones with the highest returns, as most of financial theory reviewed later in this book leads us to believe, one would expect that the more efficiently this function is performed, the higher (ceteris paribus) the value of EFF; in other words, the higher, on average, the efficiency of the investment undertaken with the funds made available by savers.

Finally, a more efficient system may be expected to screen alternative investment projects more effectively and to monitor more thoroughly and more cost efficiently the conduct of the investments (efforts of investors). The direct impact is to increase EFF. Indirectly this also means that, on average, the return/risk characteristics of the various instruments offered savers will be improved and one may expect, as a result, an increase in both S/Y and FS/S ratios.

The previous discussion thus tends to support the idea that the financial system plays an important role in permitting and promoting the growth of economies.[10] Yet growth is not an objective in itself. There is such a thing as excessive capital accumulation typically

[9] $\dot{K} = \mathrm{d}K/dt$, i.e., the change in K as a function of time.

[10] There is statistical support asserting the beneficial consequences of financial development for economic growth at the country (King and Levine, 1993a, b), industry (Rajan and Zingales, 1998), and firm levels (Aghion et al., 2007).

funded in some way by financial repression directed at households. In the case of Italy, Jappelli and Pagano (1994) suggest that household borrowing constraints,[11] in general a source of inefficiency and the mark of a less than perfect financial system, may have led to more savings than desired in the 1980s.

The excessive investment rates evident in China (noted earlier) are partially funded by high household savings (the household savings rate in China is presently approximately 25%) in part driven by financial repression in the form of government-mandated low interest rates that banks are permitted to offer depositors and prohibitions on the ownership of certain types of securities (e.g., stocks or bonds issued by foreign-based firms or foreign governments). In the absence of a significant social safety net (no government sponsored pensions or health care), households are thus induced to save a lot, much of which gets channeled into the investments of state-owned enterprises or local governments. Financial "repression" takes a different form in India: there are few banks in rural areas. As a result, many rural households do not commit their savings to the financial system at all, but prefer to buy "gold." "Investment" in the form of gold purchases contributes nothing to India's growth prospects.

While these examples are purely illustrative, they underscore the necessity of adopting a broader and more satisfactory viewpoint and of more generally studying the impact of the financial system on social welfare. This is best done in the context of the theory of general equilibrium, a subject to which we next turn.

1.5 Financial Markets and Social Welfare

Let us next consider the role of financial markets in the allocation of resources and, consequently, their effects on social welfare. The perspective provided here places the process of financial innovation in the context of the theory of general economic equilibrium whose central concepts are closely associated with the *Ecole de Lausanne* and the names of Léon Walras and Vilfredo Pareto.

Our starting point is the first theorem of welfare economics, which defines the conditions under which the allocation of resources implied by the general equilibrium of a decentralized competitive economy is efficient or optimal in the Pareto sense.

First, let us define the terms involved. Assume a timeless economy where a large number of economic agents interact. There is an arbitrary number of goods and services, *n*. Consumers possess a certain quantity (possibly zero) of each of these *n* goods (in particular, they have

[11] By "borrowing constraints," we mean the limitations that the average individual or firm may experience in his or her ability to borrow, at current market rates, from financial institutions.

the ability to work a certain number of hours per period). They can sell some of these goods and buy others at prices quoted in markets. There are a large number of firms, each represented by a production function—i.e., a given ability (constrained by what is technologically feasible) to transform some of the available goods or services (inputs) into others (outputs)—for instance, combining labor and capital to produce consumption goods. Agents in this economy act selfishly: Individuals maximize their well-being (utility), and firms maximize their profits.

General equilibrium theory tells us that, thanks to the action of the price system, order will emerge out of this uncoordinated chaos, provided certain conditions are satisfied. In the main, these hypotheses (conditions) are as follows:

H1: *Complete markets.* There exists a market on which a price is established for each of the n goods valued by consumers.
H2: *Perfect competition.* The number of consumers and firms (i.e., demanders and suppliers of each of the n goods in each of the n markets) is large enough so that no agent is in a position to influence (manipulate) market prices; i.e., all agents take prices as given.
H3: Consumers' preferences are convex.
H4: Firms' production sets are convex as well.

H3 and H4 are technical conditions with economic implications. Somewhat paradoxically, the convexity hypothesis for consumers' preferences approximately translates into strictly concave utility functions. In particular, H3 is satisfied (in substance) if consumers display risk aversion, an assumption crucial for understanding financial markets, and one that will be made throughout this text. As already noted (Box 1.2), risk aversion translates into strictly concave utility functions (see Chapter 4 for details). H4 imposes requirements on

BOX 1.2 Representing Risk Aversion

Let us reinterpret the two-date consumption stream (c_1, c_2) of Box 1.1 as the consumption levels attained "Then" or "Tomorrow" in two alternative, equally likely, states of the world. The desire for a smooth consumption stream across the two states, which we associate with risk aversion, is obviously represented by the same inequality

$$U(4) > \frac{1}{2}U(3) + \frac{1}{2}U(5)$$

and it implies the same general shape for the utility function. In other words, assuming plausibly that decision makers are **risk averse**, an assumption in conformity with most of financial theory, implies that the utility functions used to represent agents' preferences are **strictly concave.**

the production technology. It specifically rules out increasing returns to scale in production. Although important, this assumption is nevertheless not at the heart of things in financial economics since for the most part we will abstract from the production side of the economy.

A **general competitive equilibrium** is a price vector p^* and an allocation of resources, resulting from the independent decisions of consumers and producers to buy or sell each of the n goods in each of the n markets, such that, at the equilibrium price vector p^*, supply equals demand in all markets simultaneously and the action of each agent is the most favorable to him or her among all those he can afford (technologically or in terms of his budget computed at equilibrium prices).

A **Pareto optimum** is an allocation of resources, however determined, where it is impossible to redistribute resources (i.e., to go ahead with further exchanges) without reducing the welfare of at least one agent. In a Pareto-efficient (or Pareto optimal—we will use the two terminologies interchangeably) allocation of resources, it is thus not possible to make someone better off without making someone else worse off. Such a situation may not be just or fair, but it is certainly efficient in the sense of avoiding waste.

Omitting some purely technical conditions, the main results of general equilibrium theory can be summarized as follows:

1. *The existence of a competitive equilibrium*: Under H1 through H4, a competitive equilibrium is guaranteed to exist. This means that there indeed exists a price vector and an allocation of resources satisfying the definition of a competitive equilibrium as stated above.
2. *First welfare theorem*: Under H1 and H2, a competitive equilibrium, if it exists, is a Pareto optimum.
3. *Second welfare theorem*: Under H1 through H4, any Pareto-efficient allocation can be decentralized as a competitive equilibrium.

The second welfare theorem asserts that, for any arbitrary Pareto-efficient allocation, there is a price vector and a set of initial endowments such that this allocation can be achieved as a result of the free interaction of maximizing consumers and producers interacting in competitive markets. To achieve a specific Pareto optimal allocation, some redistribution mechanism will be needed to reshuffle initial resources. The availability of such a mechanism, functioning without distortion (and thus waste), is, however, very much in question. Hence the dilemma between equity and efficiency that faces all societies and their governments.

The necessity of H1 and H2 for the optimality of a competitive equilibrium provides a rationale for government intervention when these hypotheses are not naturally satisfied. The case for antitrust and other "pro-competition" policies is implicit in H2; the case for

intervention in the presence of externalities or in the provision of public goods follows from H1, because these two situations are instances of missing markets.[12]

Note that so far there does not seem to be any role for financial markets in promoting an efficient allocation of resources. To restore that role, we must abandon the fiction of a timeless world, underscoring, once again, the fact that time is of the essence in finance! Introducing the time dimension does not diminish the usefulness of the general equilibrium apparatus presented above, provided the definition of a good is properly adjusted to take into account not only its intrinsic characteristics, but also the time period in which it is available. A cup of coffee available at date t is different from a cup of coffee available at date $t + 1$, and, accordingly, it is traded on a different market and it commands a different price. Thus, if there are two dates, the number of goods in the economy goes from n to $2n$.

It is easy to show, however, that not all commodities need be traded for future as well as current delivery. The existence of a spot and forward market *for one good only* (taken as the numeraire) is sufficient to implement all the desirable allocations, and, in particular, restore, under H1 and H2, the optimality of the competitive equilibrium. This result is contained in Arrow (1964). It provides a powerful economic rationale for the existence of credit markets, markets where money is traded for future delivery.

Now let us go one step further and introduce uncertainty, which we will represent conceptually as a partition of all the relevant future scenarios into separate *states of nature*. To review, a state of nature is an exhaustive description of one possible relevant configuration of future events. Using this concept, we can extend the applicability of the welfare theorems in a fashion similar to that used with time above, by defining goods according not only to the date but also to the state of nature at which they are (might be) available. This is the notion of contingent commodities. Under this construct, we imagine the market for ice cream decomposed into a series of markets: for ice cream today, ice cream tomorrow if it rains and the Dow Jones is at 10,000; if it rains and so on. Formally, this is a straightforward extension of the basic context: there are more goods, but this is not in itself restrictive[13] (Arrow, 1964; Debreu, 1959).

[12] Our model of equilibrium presumes that agents affect one another only through prices. If this is not the case, an economic externality is said to be present. These may involve either production or consumption. For example, there have been substantial negative externalities for fishermen associated with the construction of dams in the western United States: the catch of salmon has declined dramatically as these dams have reduced the ability of the fish to return to their spawning habitats. If the externality affects all consumers simultaneously, it is said to be a public good. The classic example is national defense. If any citizen is to consume a given level of national security, all citizens must be equally secure (and thus consume this public good at the same level). Both are instances of missing markets. Neither is there a market for national defense nor for rights to disturb salmon habitats.

[13] In this context n can be as large as one needs without restriction.

The hypothesis that there exists a market for each and every good valued by consumers becomes, however, much more questionable with this extended definition of a typical good, as the example above suggests. On the one hand, the number of states of nature is, in principle, arbitrarily large and, on the other, one simply does not observe markets where commodities contingent on the realization of individual states of nature can routinely be traded. One can thus state that *if* markets are complete in the above sense, a competitive equilibrium is efficient, but the issue of completeness (H1) then takes center stage. Can Pareto optimality be obtained in a less formidable setup than one where there are complete contingent commodity markets? What does it mean to make markets "more complete?"

It was Arrow (1964), again, who took the first step toward answering these questions. Arrow generalized the result alluded to earlier and showed that it would be enough, in order to effect all desirable allocations, to have the opportunity to trade one good only across all states of nature. Such a good would again serve as the numeraire. The primitive security could thus be a claim promising \$1.00 (i.e., one unit of the numeraire) at a future date, contingent on the realization of a particular state, and zero under all other circumstances. We shall have a lot to say about such *Arrow–Debreu securities* (henceforth A–D securities), which are also called *contingent claims*. Arrow asserted that if there is one such contingent claim corresponding to each and every one of the relevant future date/state configurations, hypothesis H1 could be considered satisfied, markets could be considered complete, and the welfare theorems would apply. Arrow's result implies a substantial decrease in the number of required markets.[14] However, for a complete contingent claim structure to be fully equivalent to a setup where agents could trade a complete set of contingent commodities, it must be the case that agents are assumed to know all future spot prices, contingent on the realization of all individual states of the world. Indeed, it is at these prices that they will be able to exchange the proceeds from their A–D securities for consumption goods. This hypothesis is akin to the hypothesis of rational expectations.[15]

A–D securities are a powerful conceptual tool and are studied in depth in Chapters 9 and 11. They are not, however, the instruments we observe being traded in actual markets. Why is this the case, and in what sense is what we do observe an adequate substitute? To answer these questions, we first allude to a result (derived later on) which states that there is no single way to make markets complete. In fact, potentially a large number of alternative financial structures may achieve the same goal, and the complete A–D securities structure is only one of them. For instance, we shall describe, in Chapter 11, a context in which one might think of achieving an essentially complete market structure with options or derivative securities. We shall make use of this fact for pricing alternative instruments using arbitrage

[14] Example: 2 dates, 3 basic goods, 4 states of nature: complete commodity markets require 12 contingent commodity markets plus 3 spot markets versus 4 contingent claims and 2×3 spot markets in the Arrow setup.

[15] For an elaboration on this topic, see Drèze (1971).

techniques. Thus, the failure to observe anything close to A−D securities being traded is not evidence against the possibility that markets are indeed complete.

In an attempt to match this discussion on the role played by financial markets with the type of markets we see in the real world, one can identify the different needs met by trading A−D securities in a complete markets world. In so doing, we shall conclude that, in reality, different needs are met trading alternative specialized financial instruments (which, as we shall later prove, will all appear as portfolios of A−D securities).

As we have already observed, the time dimension is crucial for finance, and, correspondingly, the need to exchange purchasing power across time is essential. It is met in reality through a variety of specific *noncontingent* instruments, which are promised future payments independent of specific states of nature, except those in which the issuer is unable to meet his obligations (bankruptcies). Personal loans, bank loans, money market and capital market instruments, social security, and pension claims are all assets fulfilling this basic need for redistributing purchasing power in the time dimension. In a complete market setup implemented through A−D securities, the needs met by these instruments would be satisfied by a certain configuration of positions in A−D securities. In reality, the specialized instruments mentioned above fulfill the demand for exchanging income through time.

One reason for the formidable nature of the complete markets requirement is that a state of nature, which is a complete description of the relevant future for a particular agent, includes some purely personal aspects of almost unlimited complexity. Certainly the future is different for you, in a relevant way, if you lose your job, or if your house burns, without these contingencies playing a very significant role for the population at large. In a pure A−D world, the description of the states of nature should take account of these *individual contingencies* viewed from the perspective of each and every market participant! In the real world, insurance contracts are the specific instruments that deal with the need for exchanging income across purely individual events or states. The markets for these contracts are part and parcel of the notion of complete financial markets. Although such a specialization makes sense, it is recognized as unlikely that the need to trade across individual contingencies will be fully met through insurance markets because of specific difficulties linked with the hidden quality of these contingencies (i.e., the inherent asymmetry in the information possessed by suppliers and demanders participating in these markets). The presence of these asymmetries strengthens our perception of the impracticality of relying exclusively on pure A−D securities to deal with personal contingencies.

Beyond time issues and personal contingencies, most other financial instruments not only imply the exchange of purchasing power through time, but are also more specifically contingent on the realization of particular events. The relevant events here, however, are

defined on a collective basis rather than being based on individual contingencies; they are contingent on the realization of events affecting groups of individuals and observable by everyone. An example is the situation where a certain level of profits for a firm implies the payment of a certain dividend against the ownership of that firm's equity. Another is the payment of a certain sum of money associated with the ownership of an option or a financial futures. In the later cases, the contingencies (sets of states of nature) are dependent on the value of the underlying asset itself.

1.6 Financial Intermediation and the Business Cycle

Business cycles are the mark of all developed economies. According to much of current research, they are in part the result of external shocks with which these economies are repeatedly confronted. The depth and amplitude of these fluctuations, however, may well be affected by some characteristics of the financial system. This is at least the import of the recent literature on the financial accelerator. The mechanisms at work here are numerous, and we limit ourselves to giving the reader a flavor of the discussion.

The financial accelerator is manifest most straightforwardly in the context of monetary policy implementation. Suppose the monetary authority wishes to reduce the level of economic activity (inflation is feared) by raising real interest rates. The primary effect of such a move will be to increase firms' cost of capital and, as a result, to induce a decrease in investment spending as marginal projects are eliminated from consideration.

According to the financial accelerator theory, however, there may be further, substantial, secondary effects. In particular, the interest rate rise will reduce the value of firms' collateralizable assets. For some firms, this reduction may significantly diminish their access to credit, making them credit constrained. As a result, the fall in investment may exceed the direct impact of the higher cost of capital; tighter financial constraints may also affect input purchases or the financing of an adequate level of finished goods inventories. For all these reasons, the output and investment of credit-constrained firms will be more strongly affected by the action of the monetary authorities, and the economic downturn may be made correspondingly more severe. By this same mechanism, any economywide reduction in asset values may have the effect of reducing economic activity under the financial accelerator.

Which firms are most likely to be credit constrained? We would expect that small firms, those for which lenders have relatively little information about the long-term prospects, would be principally affected. These are the firms from which lenders demand high levels of collateral. Bernanke et al. (1996) provide empirical support for this assertion using US data from small manufacturing firms.

The financial accelerator has the power to make an economic downturn, of whatever origin, more severe. If the screening and monitoring functions of the financial system can be tailored more closely to individual firm needs, lenders will need to rely to a lesser extent on collateralized loan contracts. This would diminish the adverse consequences of the financial accelerator and perhaps the severity of business cycle downturns.

1.7 Financial Crises

A more radical version of the link between the financial markets and the business cycle is present in the experience of a financial crisis. It reflects the notion of a financial crisis as either a catalyst for, or the initial cause of, a severe and prolonged business cycle downturn.[16] The Great Depression of the early 1930s and the recent "Great Recession" of 2007–2009 are only the most dramatic cases in point. Both were associated with a large decline in output, a dramatic decline in investment, and a large increase in the number of persons unemployed.

Financial crises of this magnitude are typically preceded by the ending of a price "bubble" in some important asset type, with ensuing price declines in it and many other asset categories. (The "Great Recession" in the United States, in particular, was preceded by the ending of a residential real estate price bubble chiefly in the states of the Southwest, California, Texas, and Florida. Since these homes had been purchased with large mortgages and very little equity, the prospect of numerous defaults immediately arose.) With bank assets (mortgages or MBSs in the Great Recession case) declining in value, many banks face the very real possibility of insolvency, and some collapse (Lehman Brothers).[17] In either case, banks quickly become reluctant to extend any further risky loans for fear of making their own financial situations even worse if the loans cannot be repaid. As a consequence, a "credit crunch" ensues whereby many firms, and especially small and medium sized ones, find it essentially impossible to obtain loans to fund their investments and continuing operations. See Figure 1.3 for some sense of the drop in US lending activity in the first years of the Great Recession. Aggregate investment spending "dries up." Note that the resulting economic contraction parallels the one associated with the financial accelerator of Section 1.6, except that it is typically more immediate and more intense. In effect, the financial system ceases to perform the functions assigned to it by society as described in Section 1.4. The aforementioned drop in investment spending is often accompanied by a reduction in consumption spending as households react to their reduced wealth resulting from the decline in asset prices. Output contracts further.

[16] For a detailed historical analysis, see Reinhart and Rogoff (2009).

[17] Lehman Brothers was not technically a (commercial) bank because it was not permitted to take deposits. Neither did it have the right to receive loans from the US Federal Reserve, the lender of last resort. Lehman Brothers funded its assets (real estate, MBSs) with very short-term loans that had to be rolled over daily.

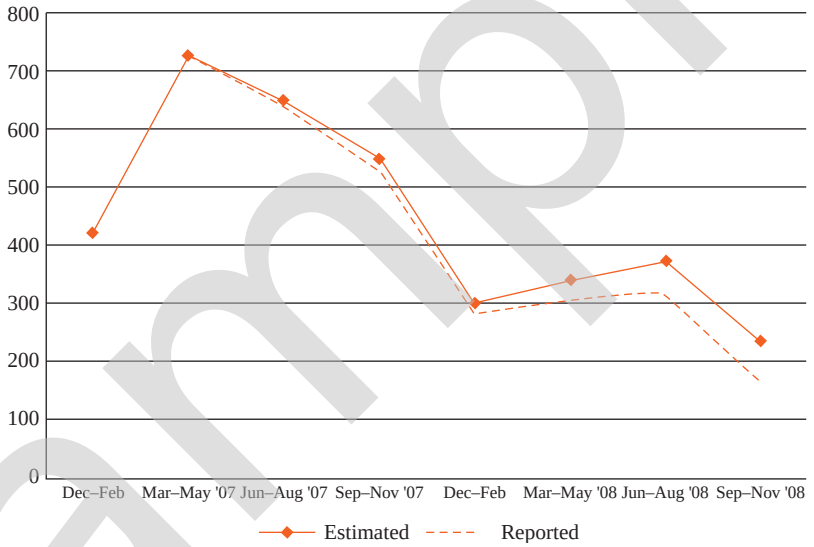


Figure 1.3: Total loan issuance, US Corporate Loans
Compiled from DealScan database of loan originals. Reported corresponds to loans reported in DealScan as of December 1, 2008. *Source: Ivashina and Scharfstein (2008), Figure 1.*

The financial crisis of 2007–2009 is estimated to have cost the US economy 22 trillion dollars, not only in the form of declines in asset values but also in the form of lost output.[18] Worldwide losses have been estimated to be not less than 1 year's world output, and as much as five times larger (Haldane (2010)). The cost of a malfunctioning financial services industry is clearly very great and, as of this writing, the world economy has yet to recover fully from the financial crisis of 2007–2009.[19]

The "Great Recession" was also the culmination of a large and rapid expansion of financial services along a number of dimensions. As a percentage of US GDP, the financial services sector rose from 4.9% in 1980 to 8.3% in 2006. The value of all United States issued private financial claims (stocks, bonds, etc.) rose from five times US GDP in 1980 to 10 times US GDP in 2007 (Greenwood and Scharfstein, 2013), a phenomenon observed in other well developed countries (see Figure 1.4). The provision of household credit similarly rose from an aggregate value equal to 0.48 GDP in 1980 to 0.99 GDP in 2007.

This enormous expansion in financial activity, in conjunction with its subsequent collapse, has led some to question whether the finance industry has grown too large. Pagano (2012), in particular, makes a strong case in this regard. For a large sample of countries,

[18] Source: General Accounting Office Report #GAO-13-180, "Financial Regulatory Reform: Financial Crisis Losses and Potential Impacts of the Dodd-Frank Act," January 16, 2013.
[19] We give a detailed overview of the 2007 financial crisis in Web Chapter D.

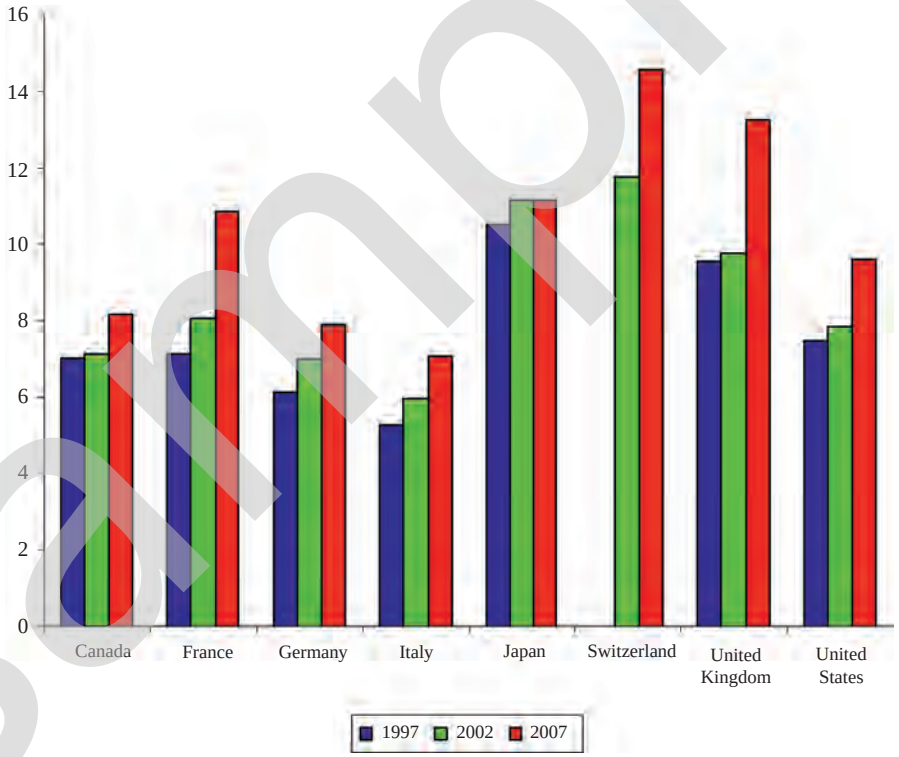


Figure 1.4: Financial deepening, advanced economies[21]
Ratio of total financial assets to GDP. *Source: Financial Accounts Statistics, OECD and Eurostat.*

he measures a country's degree of financial development either by the ratio of outstanding private credit to GDP or the ratio of aggregate stock market value to GDP (in both cases averages for the 1980–1995 period). Using these measures, he demonstrates a strong positive association between the growth in value added for industries that are highly dependent on external finance and either measure of financial development in non-OECD countries. For OECD countries, however, the association is small and not statistically different from zero. These results suggest that for countries in the earlier stages of their economic development, an expansion of the financial industry can enhance economic growth, but for countries with well-developed financial markets, this is no longer the case. Pagano (2012) concludes: "Beyond a certain point, financial development does not appear to contribute significantly to economic activity."[20]

[20] Pagano (2012), page 3.
[21] This figure is taken from Milesi-Ferretti and Tille (2010).

He then goes on to study the relationship of bank credit-worthiness (as captured by a special index, which he constructs) and the ratio of private credit issued by deposit banks and other financial institutions to GDP as the measure of financial development. For developing countries where the private credit to GDP ratio is less than 50%, he finds a positive association (correlation) between these quantities; for developed countries with private credit to GDP ratios above 50%, however, the correlation turns negative. It becomes more negative for those countries where private credit as a fraction of GDP exceeds 100%, which is the case for the United States and the United Kingdom. While not formally conclusive, the Pagano (2012) results do suggest that in some countries the financial industry may have grown so large that it now has "a life of its own," one with self interests that potentially compromise its primary role of matching savers to investors in an efficient way.

1.8 Conclusion

To conclude this introductory chapter, we advance a vision of the financial system progressively evolving toward the complete markets paradigm, starting with the most obviously missing markets and slowly, as technological innovation decreases transaction costs and allows the design of more sophisticated contracts, completing the market structure. Have we arrived at a complete market structure? Have we come significantly closer? There are opposing views on this issue. While a more optimistic perspective is proposed by Merton (1990) and Allen and Gale (1994), we choose to close this chapter on two healthily skeptical notes. Tobin (1984, p. 10), for one, provides an unambiguous answer to the above question:

> *New financial markets and instruments have proliferated over the last decade, and it might be thought that the enlarged menu now spans more states of nature and moves us closer to the Arrow–Debreu ideal. Not much closer, I am afraid. The new options and futures contracts do not stretch very far into the future. They serve mainly to allow greater leverage to short-term speculators and arbitrageurs, and to limit losses in one direction or the other. Collectively they contain considerable redundancy. Every financial market absorbs private resources to operate, and government resources to police. The country cannot afford all the markets the enthusiasts may dream up. In deciding whether to approve proposed contracts for trading, the authorities should consider whether they really fill gaps in the menu and enlarge the opportunities for Arrow–Debreu insurance, not just opportunities for speculation and financial arbitrage.*

Shiller (1993, pp. 2–3) is even more specific with respect to missing markets:

> *It is odd that there appear to have been no practical proposals for establishing a set of markets to hedge the biggest risks to standards of living. Individuals and organizations could hedge or insure themselves against risks to their standards of living if an array of*

risk markets—let us call them macro markets—could be established. These would be large international markets, securities, futures, options, swaps or analogous markets, for claims on major components of incomes (including service flows) shared by many people or organizations. The settlements in these markets could be based on income aggregates, such as national income or components thereof, such as occupational incomes, or prices that value income flows, such as real estate prices, which are prices of claims on real estate service flows.

References

Aghion, P., Fally, T., Scarpetta, S., 2007. Credit constraints as a barrier to entry and post-entry growth of firms. Econ. Policy. 22, 731–779.

Allen, F., Gale, D., 1994. Financial Innovation and Risk Sharing. MIT Press, Cambridge, MA.

Arrow, K.J., 1964. The role of securities in the allocation of risk. Rev. Econ. Stud. 31, 91–96.

Barro, R.J., Sala-i-Martin, X., 1995. Economic Growth. McGraw-Hill, New York, NY.

Bernanke, B., Gertler, M., Gilchrist, S., 1996. The financial accelerator and the flight to quality. Rev. Econ. Stat. 78, 1–15.

Bernstein, P.L., 1992. Capital Ideas. The Improbable Origins of Modern Wall Street. The Free Press, New York, NY.

Debreu, G., 1959. Theory of Value: An Axiomatic Analysis of Economic Equilibrium. John Wiley & Sons, New York, NY.

Drèze, J.H., 1971. Market Allocation Under Uncertainty. Eur. Econ. Rev. 2, 133–165.

General Accounting Office Report #GAO-13-180, Financial Regulatory Reform: Financial Crisis Losses and Potential Impacts of the Dodd-Frank Act, January 16, 2013.

Greenwood, R., Scharfstein, D., 2013. The growth of finance. J. Econ. Perspect. 27, 3–28.

Haldane, A., 2010. The $100 Billion Question. Comments given at the Institute of Regulation and Risk, Hong Kong, March 30. <http://bankofengland.co.uk/publications/documents/speeches/2010/speech433.pdf>.

Ivashina, V., Scharfstein, D., 2008. Bank lending during the financial crisis of 2008. Working paper, Harvard Business School. Published under the same title in Journal of Financial Economics, 2010 (97), 319–338.

Jappelli, T., Pagano, M., 1994. Savings, growth, and liquidity constraints. Q. J. Econ. 109, 83–109.

King, R., Levine, R., 1993a. Finance and growth: Schumpeter may be right. Q. J. Econ. 108, 713–737.

King, R., Levine, R., 1993b. Finance, entrepreneurship and growth. J. Monet. Econ. 32, 513–542.

Lee, I.H., Syed, M., Xueyau, L., 2012. Is China overinvesting and does it matter? IMF Working Paper WP/12/277.

Levine, R., 1997. Financial development and economic growth: views and agenda. J. Econ. Lit. 35, 688–726.

Merton, R.C., 1990. The financial system and economic performance. J. Financ. Serv. 4, 263–300.

Milesi-Ferretti, G.-M., Tille, C., 2010. The great retrenchment: international capital flows during the global financial crisis. Working Paper, CEPR, Economic Policy.

Pagano, M., 2012. Finance: economic lifeblood or toxin? Working Paper, CEPR and University of Naples Federico II.

Rajan, R., Zingales, L., 1998. Financial dependence and growth. Am. Econ. Rev. 88, 559–587.

Reinhart, C., Rogoff, K., 2009. This Time Is Different: Eight Centuries of Financial Folly. Princeton University Press, Princeton.

Schumpeter, J., 1934. The Theory of Economic Development, Duncker & Humblot, Leipzig. Trans. R. Opie (1934),Harvard University Press, Cambridge, MA.

Shiller, R.J., 1993. Macro Markets—Creating Institutions for Managing Society's Largest Economic Risks. Clarendon Press, Oxford.

Solow, R.M., 1956. A contribution to the theory of economic growth. Q. J. Econ. 32, 65–94.

Tobin, J., 1984. On the efficiency of the financial system. Lloyds Bank Rev.1–15.

Complementary Readings

As a complement to this introductory chapter, the reader will be interested in the historical review of financial markets and institutions found in the first chapter of Allen and Gale (1994). Bernstein (1992) provides a lively account of the birth of the major ideas making up modern financial theory, including personal portraits of their authors.

Appendix: Introduction to General Equilibrium Theory

The goal of this appendix is to provide an introduction to the essentials of general equilibrium theory, thereby permitting a complete understanding of Section 1.6 and facilitating the discussion of subsequent chapters (from Chapter 8 onward). To make this presentation as simple as possible, we will take the case of a hypothetical exchange economy (i.e., one with no production) with two goods and two agents. This permits using a very useful pedagogical tool known as the Edgeworth–Bowley box.

Let us analyze the problem of allocating efficiently a given economywide endowment of 10 units of good 1 and 6 units of good 2 among two agents, A and B. In Figure A1.1, we measure good 2 on the vertical axis and good 1 on the horizontal axis. Consider the choice problem from the origin of the axes for Mr. A, and upside down (i.e., placing the origin in the upper right corner), for Ms. B. An allocation is then represented as a point in a rectangle of size 6×10. Point E is an allocation at which Mr. A receives 4 units of

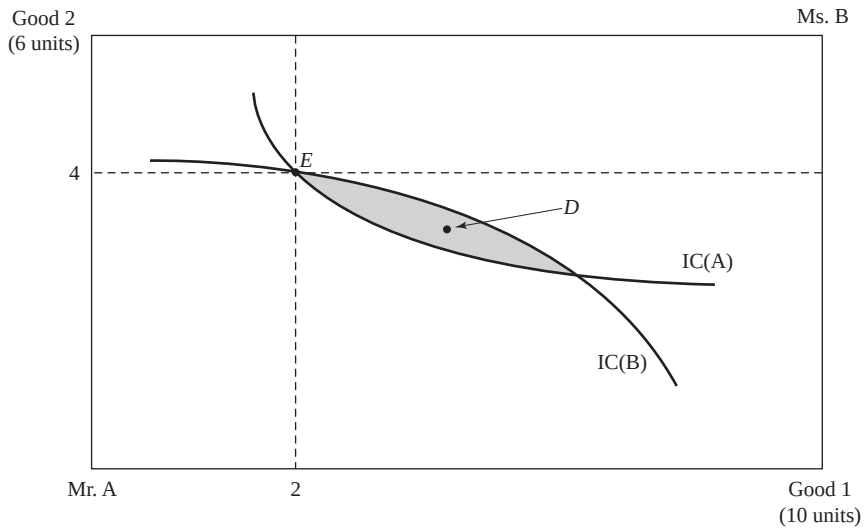


Figure A1.1
The Edgeworth–Bowley box: the set of Pareto superior allocations.

good 2 and 2 units of good 1. Ms. B gets the rest, 2 units of good 2 and 8 units of good 1. All other points in the box represent feasible allocations, i.e., alternative ways of allocating the resources available in this economy.

Pareto Optimal Allocations

In order to discuss the notion of Pareto optimal or efficient allocations, we need to introduce agents' preferences. They are fully summarized, in the graphical context of the Edgeworth–Bowley box, by indifference curves (IC) or utility level curves. Starting from the allocation E represented in Figure A1.1, we can thus record all feasible allocations that provide the same utility to Mr. A. The precise shape of such a level curve is person specific, but we can at least be confident that it slopes downward. If we take away some units of good 1, we have to compensate him with some extra units of good 2 if we are to leave his utility level unchanged. It is easy to see as well that the ICs of a consistent person do not cross, a property associated with the notion of transitivity (and with rationality) in Chapter 3. And we have seen in Boxes 1.1 and 1.2 that the preference for smoothness translates into a strictly concave utility function, or, equivalently, convex-to-the-origin level curves as drawn in Figure A1.1. The same properties apply to the IC of Ms. B, of course viewed upside down with the upper right corner as the origin.

With this simple apparatus we are in a position to discuss further the concept of Pareto optimality. Arbitrarily tracing the level curves of Mr. A and Ms. B as they pass through allocation E (but in conformity with the properties derived in the previous paragraph), only two possibilities may arise: they cross each other at E or they are tangent to one another at point E. The first possibility is illustrated in Figure A1.1, the second in Figure A1.2. In the first case, allocation E cannot be a Pareto optimal allocation. As the picture illustrates clearly, by the very

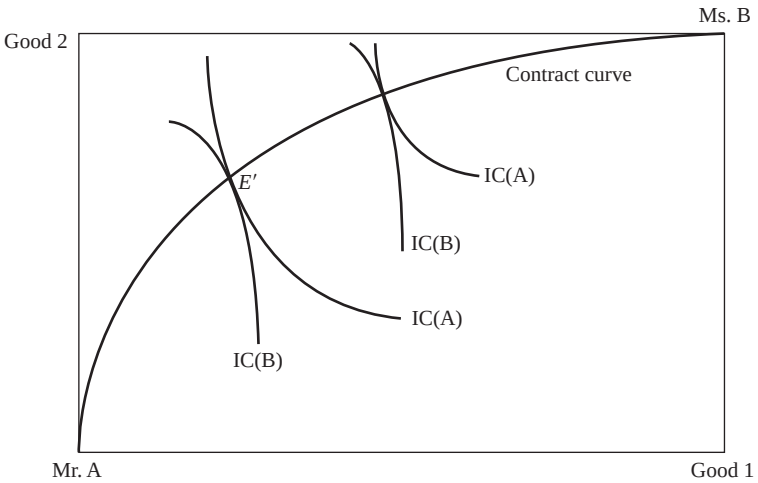


Figure A1.2
The Edgeworth–Bowley box: the contract curve.

definition of level curves, if the ICs of our two agents cross at point E, there is a set of allocations (corresponding to the shaded area in Figure A1.1) that both Mr. A and Ms. B simultaneously prefer to E. These allocations are Pareto superior to E, and, in that situation, it would indeed be socially inefficient or wasteful to distribute the available resources as indicated by E. Allocation D, for instance, is feasible and preferred to E by both individuals.

If the ICs are tangent to one another at point E′ as in Figure A1.2, no redistribution of the given resources exists that would be approved by both agents. Inevitably, moving away from E′ decreases the utility level of one of the two agents if it favors the other. In this case, E′ is a Pareto optimal allocation. Figure A1.2 illustrates that it is not generally unique, however. If we connect all the points where the various ICs of our two agents are tangent to each other, we draw the line, labeled the contract curve, representing the infinity of Pareto optimal allocations in this simple economy.

An indifference curve for Mr. A is defined as the set of allocations that provide the same utility to Mr. A as some specific allocation; e.g., allocation E: $\{(c_1^A, c_2^A): U(c_1^A, c_2^A) = U(E)\}$. This definition implies that the slope of the IC can be derived by taking the total differential of $U(c_1^A, c_2^A)$ and equating it to zero (no change in utility along the IC), which gives:

$$\frac{\partial U(c_1^A, c_2^A)}{\partial c_1^A} \mathrm{d}c_1^A + \frac{\partial U(c_1^A, c_2^A)}{\partial c_2^A} \mathrm{d}c_2^A = 0 \tag{1.3}$$

and thus

$$-\frac{\mathrm{d}c_2^A}{\mathrm{d}c_1^A} = \frac{\dfrac{\partial U(c_1^A, c_2^A)}{\partial c_1^A}}{\dfrac{\partial U(c_1^A, c_2^A)}{\partial c_2^A}} \equiv \mathrm{MRS}_{1,2}^A \tag{1.4}$$

That is, the negative (or the absolute value) of the slope of the IC is the ratio of the marginal utility of good 1 to the marginal utility of good 2 specific to Mr. A and to the allocation (c_1^A, c_2^A) at which the derivatives are taken. It defines Mr. A's marginal rate of substitution (MRS) between the two goods.

Equation (1.4) permits a formal characterization of a Pareto optimal allocation. Our former discussion has equated Pareto optimality with the tangency of the ICs of Mr. A and Ms. B. Tangency, in turn, means that the slopes of the respective ICs are identical. Allocation E, associated with the consumption vector $(c_1^A, c_2^A)^E$ for Mr. A and $(c_1^B, c_2^B)^E$ for Ms. B, is thus Pareto optimal if, and only if,

$$\mathrm{MRS}_{1,2}^A = \frac{\dfrac{\partial U(c_1^A, c_2^A)^E}{\partial c_1^A}}{\dfrac{\partial U(c_1^A, c_2^A)^E}{\partial c_2^A}} = \frac{\dfrac{\partial U(c_1^B, c_2^B)^E}{\partial c_1^B}}{\dfrac{\partial U(c_1^B, c_2^B)^E}{\partial c_2^B}} = \mathrm{MRS}_{1,2}^B \tag{1.5}$$

Equation (1.5) provides a complete characterization of a Pareto optimal allocation in an exchange economy except in the case of a corner allocation, i.e., an allocation at the frontier of the box where one of the agents receives the entire endowment of one good and the other agent receives none. In that situation, it may well be that the equality could not be satisfied except, hypothetically, by moving to the outside of the box, i.e., to allocations that are not feasible since they require giving a negative amount of one good to one of the two agents.

So far we have not touched on the issue of how the discussed allocations may be determined. This is the viewpoint of Pareto optimality, which analysis is exclusively concerned with deriving efficiency properties of given allocations, regardless of how they were achieved. Let us now turn to the concept of competitive equilibrium.

Competitive Equilibrium

Associated with the notion of competitive equilibrium is the notion of markets and prices. One price vector (one price for each of our two goods), or simply a relative price taking good 1 as the numeraire, and setting $p_1 = 1$, is represented in the Edgeworth–Bowley box by a downward-sloping line. From the viewpoint of either agent, such a line has all the properties of the budget line. It also represents the frontier of their opportunity set. Let us assume that the initial allocation, before any trade, is represented by point I in Figure A1.3. Any line sloping downward from I does represent the set of allocations that Mr. A, endowed with I, can obtain by going to the market and exchanging (competitively, taking prices as given)

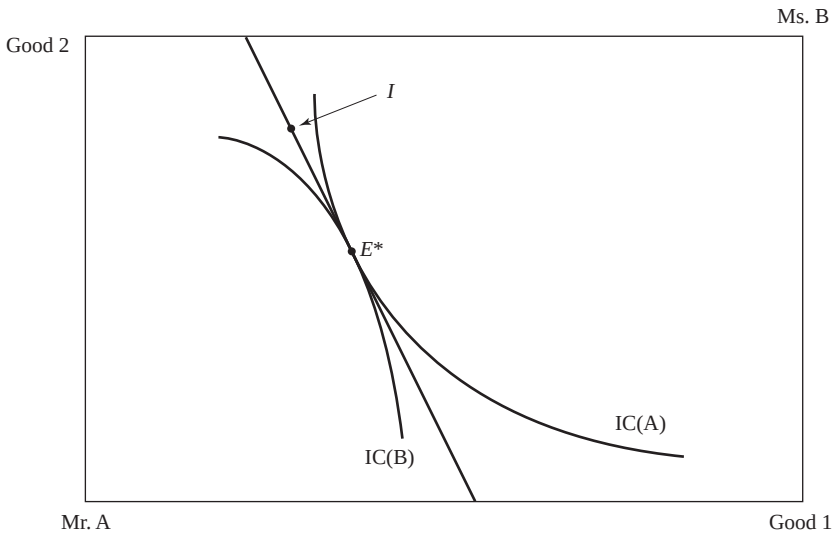


Figure A1.3
The Edgeworth–Bowley box: equilibrium achieved at E*.

good 1 for 2 or vice versa. He will maximize his utility subject to this budget constraint by attempting to climb to the highest IC making contact with his budget set. This will lead him to select the allocation corresponding to the tangency point between one of his ICs and the price line. Because the same prices are valid for both agents, an identical procedure, viewed upside down from the upper right-hand corner of the box, will lead Ms. B to a tangency point between one of her ICs and the price line. At this stage, only two possibilities may arise: Mr. A and Ms. B have converged to the same allocation (the two markets, for good 1 and 2, clear—supply and demand for the two goods are equal and we are at a competitive equilibrium); or the two agents' separate optimizing procedures have led them to select two different allocations. Total demand does not equal total supply, and an equilibrium is not achieved. The two situations are described, respectively, in Figures A1.3 and A1.4.

In the disequilibrium case of Figure A1.4, prices will have to adjust until an equilibrium is found. Specifically, with Mr. A at point A and Ms. B at point B, there is an excess demand of good 2 but insufficient demand for good 1. One would expect the price of 2 to increase relative to the price of good 1 with the likely result that both agents will decrease their net demand for 2 and increase their net demand for 1. Graphically, this is depicted by the price curve tilting with point I as the axis and looking less steep (indicating, for instance, that if both agents wanted to buy good 1 only, they could now afford more of it). With regular ICs, the respective points of tangencies will converge until an equilibrium similar to the one described in Figure A1.3 is reached.

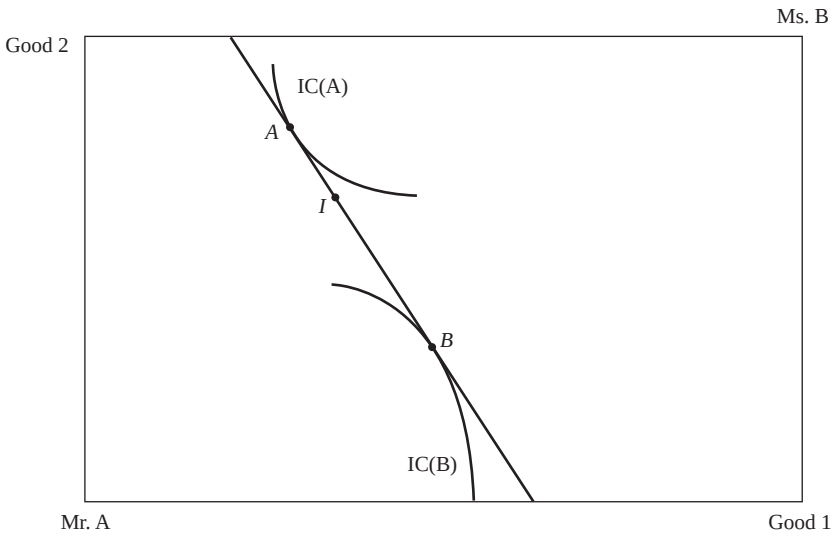


Figure A1.4
The Edgeworth–Bowley box: disequilibrium, excess demand for good 2, excess supply for good 1.

We will not say anything here about the conditions guaranteeing that such a process will converge. Let us rather insist on one crucial necessary precondition: that an equilibrium exists. In the text we have mentioned that assumptions H1 to H4 are needed to guarantee the existence of an equilibrium. Of course, H4 does not apply here. H1 states the necessity of the existence of a price for each good, which is akin to specifying the existence of a price line. H2 defines one of the characteristics of a competitive equilibrium: that prices are taken as given by the various agents and the price line describes their perceived opportunity sets. Our discussion here can enlighten the need for H3. Indeed, in order for an equilibrium to have a chance to exist, the geometry of Figure A1.3 makes clear that the shape of the two agents' ICs is relevant. The price line must be able to separate the "better than" areas of the two agents' ICs passing through a same point—the candidate equilibrium allocation. The better than area is simply the area above a given IC. It represents all the allocations providing higher utility than those on the level curve. This separation by a price line is not generally possible if the ICs are not convex, in which case an equilibrium cannot be guaranteed to exist. The problem is illustrated in Figure A1.5.

Once a competitive equilibrium is observed to exist, which logically could be the case even if the conditions that guarantee existence are not met, the Pareto optimality of the resulting allocation is ensured by H1 and H2 only. In substance this is because once the common price line at which markets clear exists, the very fact that agents optimize taking prices as given leads them to a point of tangency between their highest IC and the common price line. At the resulting allocation, both MRS are equal to the same price line and, consequently, are identical. The conditions for Pareto optimality are thus fulfilled.

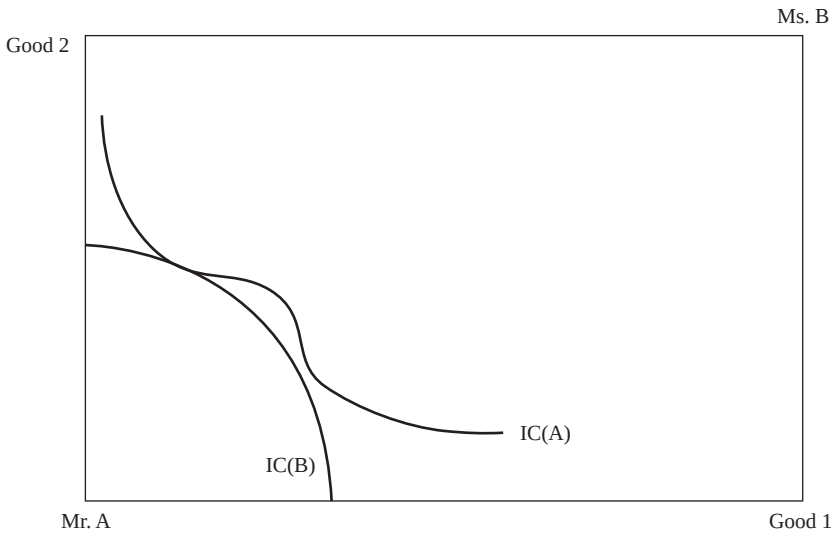


Figure A1.5
The Edgeworth–Bowley box: nonconvex indifference curves.

CHAPTER 2

The Challenges of Asset Pricing: A Road Map

Chapter Outline

2.1 The Main Question of Financial Theory

Valuing risky cash flows or, equivalently, pricing risky assets is at the heart of financial theory.

Our discussion thus far has been conducted from the perspective of society as a whole, and it argues that a progressively more complete set of financial markets will generally enhance societal welfare by making it easier for economic agents to transfer income across future dates and states via the sale or purchase of individually tailored portfolios of securities. The desire of agents to construct such portfolios will depend as much upon the market prices of the relevant securities as on their strict availability, and this leads us to the main topic of the text.

Intermediate Financial Theory.

Indeed, the major practical question in finance is, "How do we value a risky cash flow?" and the main objective of this text is to provide a complete and up-to-date treatment of how it can be answered. For the most part, this book is thus a text on asset pricing. Indeed, an asset is nothing else than the right to future cash flows, whether these future cash flows are the result of interest payments, dividend payments, insurance payments, or the resale value of the asset. Furthermore, when we compute a project's risk-adjusted present value (PV), we are, in effect, asking the question: If this project's cash flow were traded as though it were a security, at what price would it sell given that it should pay the prevailing rate on other securities with the same risk level? We compare its fair market value, estimated in this way, with its cost, P_0. Evaluating a project is thus a special case of evaluating a security.

Viewed in this way and abstracting from risk for the moment, the key object of our attention, be it an asset or an investment project, can be summarized as in Table 2.1.

In Table 2.1, $t = 0,1,2,\ldots\tau,\ldots,T$ represents future dates. The duration of each period, the length of time between $\tau - 1$ and τ, is arbitrary and can be viewed as 1 day, 1 month, 1 quarter, or 1 year. The expression $\tilde{C}F_\tau$ stands for the possibly uncertain cash flows in period τ (whenever useful, we will identify random variables with a tilde), r_τ^f is the risk-free, per-period interest rate prevailing between date 0 and τ, and P_0 denotes the to-be-determined current price or valuation of the future cash flow. If the future cash flows will be available for sure, valuing the flow of future payments is easy. It requires adding the future cash flows after discounting them by the risk-free rate of interest, i.e., adding the cells in the last line of the table. The discounting procedure is indeed at the heart of our problem: it clearly serves to translate future payments into current dollars (those that are to be used to purchase the right to these future cash flows or in terms of which the current value of the future cash flow is to be expressed). In other words, the discounting procedure is what makes it possible to compare future dollars (i.e., dollars that will be available in the future) with current dollars.

If, however, the future cash flows will not be available for certain but are subject to random events—the interest payments depend on the debtor remaining solvent, the dividend payments depend on the financial strength of the equity issuer, the returns to the investment project depend on its commercial success—then the valuation question becomes trickier, so much so that there does not exist a universal way of proceeding that dominates all others.

In the same way that one dollar for sure tomorrow does not generally have the same value as one current dollar, one dollar tomorrow under a set of more or less narrowly defined

Table 2.1: Valuing a risk-free cash flow

$t = 0$	$t = 1$	$t = 2$	$\ldots\tau\ldots$	$t = T$
$\boldsymbol{P_0?}$	$\boldsymbol{\tilde{C}F_1}$	$\boldsymbol{\tilde{C}F_2}$	$\boldsymbol{\tilde{C}F_\tau}$	$\boldsymbol{\tilde{C}F_T}$
	$\frac{CF_1}{(1+r_1^f)}$	$\frac{CF_2}{(1+r_2^f)^2}$	$\frac{CF_\tau}{(1+r_\tau^f)^\tau}$	$\frac{CF_T}{(1+r_T^f)^T}$

circumstances, i.e., in a subset of all possible states of nature, is also not worth even one current dollar discounted at the risk-free rate. Assume the risk-free rate of return is 5% per year, then discounting one dollar available in 1 year at the risk-free rate yields $(1\$/1.05) \cong \$0.95$. This equality is exactly that: it states that \$1 tomorrow will have a market price of \$0.95 today when 1-year risk-free securities earn 5%. It is a market assessment to the extent that the 5% risk-free rate is an equilibrium market rate. Now if \$1 for sure tomorrow is worth \$0.95, it seems likely that \$1 tomorrow "possibly," i.e., in a restricted subset of the states of nature, should certainly be worth less than \$0.95. One can speculate, for instance, that if the probability of \$1 in a year is about $\frac{1}{2}$, then one should not be willing to pay more than $\frac{1}{2} \times \$0.95$ for that future cash flow. But we have to be more precise than this. To that end, several lines of attack will be pursued. Let us outline them.

2.2 Discounting Risky Cash Flows: Various Lines of Attack

First, as in the certainty case, it is plausible to argue (and it can be formally demonstrated) that the valuation process is additive: the value of a sum of future cash flows will take the form of the sum of the values of each of these future cash flows. Second, as already anticipated, we will work with probabilities, so that the random cash flow occurring at a future date τ will be represented by a random variable: $\tilde{CF}_\tau$, for which a natural reference value is its expectation $E\tilde{CF}_\tau$. Another reference value would be this expected future cash flow discounted at the risk-free rate: $E\tilde{CF}_\tau/((1+r_\tau^f)^\tau)$. Now the latter expression cannot generally be the solution to our problem, although it is intuitively understandable that it will be when the risk issue does not matter—i.e., when market participants can be assumed to be risk neutral. In the general case where risk must be taken into account, which typically means that risk-bearing behavior needs to be remunerated, alterations to that reference formula are necessary. These alterations may take any of the following forms:

1. The most common strategy consists of discounting at a rate that is higher than the risk-free rate, i.e., to discount at a rate that is the risk-free rate increased by a certain amount π (a risk premium) as in

$$\frac{E\tilde{CF}_\tau}{(1+r_\tau^f+\pi_\tau)^\tau}$$

   The underlying logic is straightforward: To price an asset equal to the present value of its expected future cash flows discounted at a particular rate is to price the asset in a manner such that, at its present value price, it is expected to earn that discount rate. The appropriate rate, in turn, must be the analyst's estimate of the expected rate of return on other financial assets that represent title to cash flows similar in risk and timing to that

of the asset in question. This strategy has the consequence of pricing the asset to pay the prevailing competitive rate for its risk class. When we follow this approach, the key issue is to compute the appropriate risk premium.[1]

2. Another approach, in the same spirit, consists of correcting the expected cash flow itself in such a way that one can continue discounting at the risk-free rate. The standard way of doing this is to decrease the expected future cash flow by a factor Π that once again will reflect some form of risk or insurance premium as in

$$\frac{E\tilde{C}F_\tau - \Pi_\tau}{(1+r_\tau^f)^\tau}$$

3. The same idea can take the form, it turns out quite fruitfully, of distorting the probability distribution over which the expectations operator is applied so that taking the expected cash flow with this modified probability distribution justifies once again discounting at the risk-free rate:

$$\frac{\hat{E}\tilde{C}F_\tau}{(1+r_\tau^f)^\tau}$$

Here $\hat{E}$ denotes the expectation taken with respect to the modified probability distribution.

4. Finally, one can think of decomposing the future cash flow $\tilde{C}F_\tau$ into its state-by-state elements. Denote $(CF(\theta_\tau))$ the actual payment that will occur in the specific possible state of nature θ_τ. If one is able to find the price today of \$1 in period τ conditional on that particular state θ_τ being realized, say $q(\theta_\tau)$, then surely the appropriate current valuation of $\tilde{C}F_\tau$ is

$$\sum_{\theta_\tau \in \Theta_\tau} q(\theta_\tau)CF(\theta_\tau)$$

where the summation takes place over all the possible future states θ_τ. The quantity $q(\theta_t)$ is often referred to as a "state price," and in important applications it will resemble the more traditional discount factor.

The procedures described above are alternative ways of attacking the difficult valuation problem we have outlined, but they can only be given content in conjunction with theories explaining how to compute the risk premia (cases 1 or 2), to identify the distorted

[1] Let us be sure we understand exactly what this expression says: the r_τ^f denotes the period by period rate of return on a default-free security which pays an amount of money τ periods in the future while π_τ is the return risk premium expected to prevail over this same time horizon. We thus discount each future cash flow "individually." In a typical calculation where there are cash flows in many future periods, we frequently assume $r_\tau^f = r_f$ and $\pi_\tau = \pi$ for all τ; i.e., the risk-free rate and risk premium are constant looking forward. All future cash flows thus end up being discounted at the same rate.

probability distribution (case 3) or to price future dollars state by state (case 4). For strategies 1 and 2, this can be done using the capital asset pricing model (CAPM), the consumption capital asset pricing model (CCAPM), or the arbitrage pricing theory (APT); strategy 3 is characteristic of the Martingale approach; strategy 4 describes the perspective of Arrow−Debreu (AD) pricing.

2.3 Two Main Perspectives: Equilibrium versus Arbitrage

There is another, even more fundamental way of classifying alternative valuation theories. All the known valuation theories cited above employ one of two main methodologies: the equilibrium approach or the arbitrage approach.

The traditional equilibrium approach consists of an analysis of the factors determining the supply and demand for the cash flow (asset) in question. The arbitrage approach attempts to value a cash flow on the basis of observations made on the values of the various elements making up that cash flow.

Let us illustrate this distinction with an analogy. You are interested in pricing a bicycle. There are two ways to approach the question. If you follow the equilibrium approach, you will want to study the determinants of supply and demand. Who are the producers? How many bicycles are they able to produce? What are the substitutes, including probably the existing stock of old bicycles potentially appearing on the second-hand market? After dealing with supply, turn to demand: Who are the buyers? What are the forecasts of the demand for bicycles? And so on. Finally, you will turn to the market structure. Is the market for bicycles competitive? If so, we know how the equilibrium price will emerge as a result of the matching process between demanders and suppliers. The equilibrium perspective is a sophisticated, complex approach, with a long tradition in economics, one that has also been applied in finance, at least since the 1950s. We will follow it in the first part of this book, adopting standard assumptions that simplify, without undue cost, the supply and demand analysis for financial objects: the supply of financial assets at any point in time is assumed to be fixed, and financial markets are viewed as competitive. Our analysis can thus focus on the determinants of the demand for financial assets.

This requires that we first spend some time discussing the preferences and attitudes toward risk of investors, those who demand the assets (Chapters 3 and 4), before modeling the investment process, i.e., how the relative demands for the various financial assets are determined (Chapters 5−7). Armed with these tools, we will review the three main equilibrium theories, the CAPM in Chapter 8, AD pricing in Chapter 9, and the CCAPM in Chapter 10.[2]

[2] In Web Chapter A we generalize the CCAPM by making more explicit the macroeconomic setting that underlies the CCAPM. An understanding of these macrofinancial linkages has become more important given the recent financial crisis and the advent of the "Great Recession."

The arbitrage approach to valuing bicycles starts from observing that a bicycle is not (much) more than the sum of its parts. Accordingly, if you know the price of all the necessary components—frame, handlebar, wheel, tire, saddle, brake, and gearshift—you can determine relatively easily the market value of the bicycle. The knowledge of how to assemble the bicycle and the time required to do so, however, are not in infinite supply. These considerations suggest that the arbitrage approach may hold only as an approximation, one that may be rather imprecise in circumstances where the time and intellectual ability required to "assemble the bicycle" from the necessary spare parts are nontrivial; i.e., when the remuneration of the necessary "engineers" matters.[3]

The arbitrage approach is, in a sense, much more straightforward than the equilibrium approach. It is also more robust: if the arbitrage relationship between the price of the bicycle and the price of its parts does not hold, anyone with a little time could become a bicycle manufacturer and make good money. If too many people exploit that idea, however, the prices of parts and the prices of bicycles will start adjusting and be forced into line. This very idea is especially powerful for the object at hand, financial assets, because if markets are complete in the sense discussed in Section 1.6, then it can easily be shown that *all* the component prices necessary to value *any* arbitrary cash flow are available. Furthermore, little time and few resources (relative to the global scale of product markets) are needed to exploit arbitrage opportunities in financial markets.

There is, however, an obvious limitation to the arbitrage approach. Where do we get the price of the parts if not through an equilibrium approach? That is, the arbitrage approach is much less ambitious and more partial than the equilibrium approach. Even though it may be more practically useful in the domains where the price of the parts is readily available, it does not make up for a general theory of valuation and, in that sense, has to be viewed as a complement to the equilibrium approach. In addition, the equilibrium approach, by forcing us to rationalize investors' demand for financial assets, provides useful lessons for the practice of asset management. The foundations of this inquiry will be put in place in Chapters 3−7—which together make up Part II of the book—while Chapter 16 will extend the treatment of this topic beyond the case of the traditional one-period static portfolio analysis and focus on the specificities of long-run portfolio management.

Finally, the arbitrage and equilibrium approaches can be combined. In particular, one fundamental insight that we will develop in Chapter 11 is that any cash flow can be viewed

[3] In a similar vein, "financial engineers" seek to create new securities by cleverly packaging existing ones, or seek to design arbitrage portfolios which increase in value as the relative prices of their constituent securities (by analogy, the bicycle, and its independently traded constituent parts) come into better alignment. While a pure arbitrage portfolio is strictly risk free, most arbitrage portfolios arising from financial engineering will have positive payoffs provided certain low probability ("Black Swan") events do not occur. At the start of the financial crisis, it was these very low probability events that came to pass with disastrous consequences for the institutions themselves (e.g., AIG).

Table 2.2: The road map

	Equilibrium	Arbitrage
Preliminaries	Utility theory—Chapters 3–4 Investment demand—Chapters 5–7	
Computing risk premia	CAPM—Chapter 8 CCAPM—Chapter 10	APT—Chapter 14
Identifying distorted probabilities		Martingale measure—Chapters 12–13
Pricing future dollars state by state	AD pricing I—Chapter 9	AD pricing II—Chapter 11

as a portfolio of AD securities, i.e., it can be replicated with AD securities. This makes it very useful to start using the arbitrage approach with AD securities as the main building blocks for pricing assets or valuing cash flows. Conversely, the same chapter will show that options can be very useful in completing the markets and thus in obtaining a full set of prices for "the parts that will then be available to price the bicycles." In other words, the AD equilibrium pricing theory is a good platform for arbitrage valuation. The link between the two approaches is indeed so tight that we will use our acquired knowledge of equilibrium models—reviewed in Part III—to understand one of the major arbitrage approaches, the Martingale pricing theory (Chapters 12 and 13).[4] We will then propose an overview of the APT in Chapter 14. Chapters 11 through 15 together make up Part IV of this book. This outline is summarized in Table 2.2.

Part V will focus on three extensions. As already mentioned, Chapter 16 deals with long-run asset management. Chapter 17 focuses on some implications of incomplete markets whose consequences are illustrated from the twin viewpoints of the equilibrium and arbitrage approaches. We will use it as a pretext to review the Modigliani–Miller theorem and, in particular, to understand why it depends on the hypothesis of complete markets. Finally, in Chapter 18, we will open up, just a little, the Pandora's box of heterogeneous beliefs. Our goal is to understand a number of issues that are largely swept under the rug in standard asset management and pricing theories and, in the process, restate the efficient market hypothesis.[5]

2.4 Decomposing Risk Premia

Method 1 in Section 2.2 represents the standard methodology for real investment project evaluation and, if only for this reason, deserves a bit more attention. Generalizing

[4] Web Chapter B presents additional illustrations and applications of Martingale pricing theory.

[5] Web Chapter D concludes with a discussion of the recent "financial crisis." While we describe the causes (in our view) and consequences of the crisis, our principal objective is to relate the event to the concepts introduced in this text.

the present value expression to a many period cash flow timing typical of real projects yields

$$V_{\text{project}} = \sum_{t=1}^{T} \frac{E\tilde{C}F_t}{(1+r_{\text{f}}+\pi_{\text{p}})^t}$$

where we simplify the discussion below by assuming a constant risk-free rate of interest yields ($r_t^f \equiv r_{\text{f}}$ for all t) and a constant project risk premium π_{p}. Recall that an objective, market-based valuation of the project requires that π_{p} represents the risk premium on a stock (or portfolio of stocks) whose cash-flow timing and risk characteristics resemble (in a manner to be made formal in later chapters) those of the project; i.e.

$\pi_{\text{p}} = \pi_i = E\tilde{r}_i - r_{\text{f}}$, where $E\tilde{r}_i$ is the expected return to the "approximating stock or stock portfolio i."

Work in empirical asset pricing suggests is that the return premium $\tilde{r}_i - r_{\text{f}}$ on a stock i may be intertemporarily decomposed as a linear combination of fundamental stochastic factors $\tilde{F}^j$, $j = 1, 2, \ldots, J$:

$$\tilde{r}_{i,t} - r_{\text{f}} = \alpha_i + \beta_i^1 \tilde{F}_t^1 + \beta_i^2 \tilde{F}_t^2 + \cdots + \beta_i^J \tilde{F}_t^J + \tilde{\varepsilon}_{i,t} \tag{2.1}$$

where the β_i^j measures the sensitivity of stock i's return to the specific underlying factor j, and $\tilde{\varepsilon}_{i,t}$ is an i.i.d. mean zero random component. Under representation (2.1), all stock returns are thus determined by the same J factors. These factors affect different stocks to differing degrees as measured by a stock's specific factor sensitivities. The impact of these factors subsumes the entire risk premium except for a "white noise" residual.

What are these factors? Some factors define macroeconomic conditions such as the inflation rate or the state of the business cycle as measured by the GDP growth rate. All firms are affected, to varying degrees, by the business cycle. Chen et al. (1986) consider factors such as industrial production, inflation expectations, and oil prices. Other factors are believed to measure various permanent psychological biases on the part of investors, biases that appear, in some cases, to have a permanent influence on equity return patterns.

Why the focus is on estimating risk premia, when our ultimate objective is asset pricing? Given cash-flow estimates, prices and returns are, of course, "dual" to one another in the sense that knowing a project's price determines the implied risk premium and vice versa via the present value relationship. Another motivation is that most "investors" are not undertaking real investment projects but are buying portfolios of individual securities. Rather than thinking of next period's price of a security relative to its price today, it is more natural for investors to think in terms of what the security is expected to earn above r_{f}—its risk premium—over their chosen time horizon.

2.5 Models and Stylized Facts

Financial economics has as its goal the understanding of financial market behavior. While this statement may seem obvious, it has no real content until we clarify what an "understanding" would mean. For asset pricing phenomena, in particular, it means that the event under study can be explained in a model economy where self-interested economic agents determine their demands for various securities based on certain first principles/axioms of economic behavior.[6] The model may also go on to specify how these demands interact with the supplies of the various securities to determine equilibrium prices and returns. Chapters 3–10 of this text essentially construct the basic model paradigms of finance.

A model is necessarily an abstraction or (dramatic) simplification of reality. Many economic mechanisms are ignored with only the most critical retained. As a result, there will inevitably be some aspects of reality which it will be unable to explain. What characteristics, then, describe a good in contrast to a poor economic model? We propose the three criteria listed below:

i. A good model must be simple enough to enrich our intuition. In other words, the principal economic mechanisms within the model that allow it to explain the phenomenon under study must be readily apparent.
ii. The abstraction which the model represents must be tailored to (or rich enough to address meaningfully) the questions being asked of it. Researchers would not seek to understand observed patterns in the US distributions of income and wealth, for example, in a representative (single) agent model.[7]
iii. The model should be able to give precise answers to questions we pose concerning the behavior of the real economy.

How does one acquire confidence in a model that ostensibly satisfies the above criteria? The answer is straightforward: the more historical phenomena the model is able to explain successfully, the more confidence researchers have in its ability to provide answers to current and future real-world questions. It is here that the financial "stylized facts" enter the scene. These "stylized facts" are simply well-documented price, quantity, or return patterns that have been present in financial market data over long periods of time. In the latter sense, they are said to be "secular."[8] At a minimum, it is imperative that a good and trustworthy asset pricing model be able to replicate (reproduce) them.

---

[6] As such economic models differ from scientific ones (e.g., of the atom). Scientific models describe the laws of nature, which are invariant to human activity. In economic models, everything is the result of human activity.

[7] Requirement (ii) is not intended to suggest that radically different models should be proposed to explain different data regularities. The underlying principals must be the same and the results mutually consistent in areas of overlap.

[8] That is, "secular" as in "existing or continuing through ages and centuries" (Webster's Collegiate Dictionary, Ninth Edition).

Although his focus is on the mechanisms underlying business cycles, we adopt the general modeling perspective of R.E. Lucas, Jr. as expressed in Lucas (1980):

> *One of the functions of theoretical economics is to provide fully articulated, artificial economic systems that can serve as laboratories in which policies that would be prohibitively expensive to experiment with in actual economies can be tested out at much lower cost. To serve this function well, it is essential that the artificial "model" economy be distinguished as sharply as possible in discussion from actual economies. Insofar as there is confusion between statements of opinion as to the way we believe actual economies would react to particular policies and statements of verifiable fact as to how the model will react, the theory is not being effectively used to help us to see which opinions about the behavior of actual economies are accurate and which are not. This is the sense in which insistence on the "realism" of an economic model subverts its potential usefulness in thinking about reality. Any model that is well enough articulated to give clear answers to the questions we put to it will necessarily be artificial, abstract, patently "unreal."*
>
> *At the same time, not all well-articulated models will be equally useful. Though we are interested in models because we believe they may help us to understand matters about which we are currently ignorant, we need to test them as useful imitations of reality by subjecting them to shocks for which we are fairly certain how actual economies, or parts of economies, would react. The more dimensions on which the model mimics the answers actual economies give to simple questions, the more we trust its answers to harder questions. This is the sense in which more "realism" in a model is clearly preferred to less.*
>
> *On this general view of the nature of economic theory then, a "theory" is not a collection of assertions about the behavior of the actual economy but rather an explicit set of instructions for building a parallel or analogue system — a mechanical, imitation economy. A "good" model, from this point of view, will not be exactly more "real" than a poor one, but will provide better imitations. Of course, what one means by a "better imitation" will depend on the particular questions to which one wishes answers.*

Accordingly we next highlight a few quantitative and qualitative properties of financial markets that serve as basic stylized facts against which the financial models we will propose in the remainder of this text should be measured. We focus exclusively on capital market phenomena rather than empirical regularities related to the firm's corporate financing activities.

2.5.1 The Equity Premium

A broadly diversified portfolio of stocks (e.g., the $S\&P_{500}$, the DAX, the CAC) consistently earns average returns substantially in excess of the risk-free rate (normally proxied by the return on short-term debt securities issued by a nation's national treasury authority). Tables 2.3 and 2.4 give some ideas as to the magnitudes involved.

Table 2.3: US returns: 1889–2010[a]

Time period	Real Return on a Market Index[b]	Real Return on a Relatively Riskless Security	% Risk Premium
	Mean	Mean	Mean
1889–2010	7.5%	1.1%	6.4%
1889–1978	7.0%	0.8%	6.2%
1926–2010	8.0%	0.8%	7.2%
1946–2010	7.5%	0.8%	6.7%

[a]Data from Mehra (2012); annualized returns.
[b]The $S\&P_{500}$ and its antecedents.

Table 2.4: The equity premium: the principal capital markets[a]

Country	Time Period	% Risk Premium	Country	Time Period	% Risk Premium
Belgium	1900–2010	5.5%	Sweden	1900–2010	6.6%
Holland	1900–2010	6.5%	UK	1900–2010	6.0%
France	1900–2010	8.7%	Australia	1900–2010	8.3%
Germany	1900–2010	9.8%	Canada	1900–2010	5.6%
Ireland	1900–2010	5.3%	India	1991–2004	11.3%
Italy	1900–2010	9.8%	Japan	1900–2010	9.0%

[a]*Source and details*: Dimson et al. (2010); annualized returns.

Note that in the US statistics (Table 2.3), the premium for long horizons never falls below 6%. With minor exceptions, the same results usually carry over to 10-year horizons (i.e., the pattern is secular). Similar if not stronger results carry over to international data (Table 2.4), despite two wars which led to substantial capital destruction and, in some cases, the cessation of organized competitive stock trading (e.g., Germany, France). The experiences of the United Kingdom and Canada compare most closely with the United States in that there was no interruption to trading and the respective governments did not seek to control this form of financial market activity. India's stock market experience is more recent and Japan's stellar performance is largely a post-World War II phenomenon prior to 1990. Across all markets, a robust equity premium over long horizons is an empirical fact.[9]

Within the current class of "rational economic models," those that we are about to describe in the chapters to follow, it is extremely difficult to replicate these statistics, so much so

[9] A prominent exception is the experience of recent history. For the US stock market, the average annual return for the period 2000–2010 was −0.39%, while US Treasury bonds of time to maturity exceeding 10 years paid on average 7.18% over the same period.

Table 2.5: Average annualized excess returns for 10 portfolios sorted on BE/ME[a]

Lowest		→Increasing (BE/ME)→						Highest	
Port 1	Port 2	Port 3	Port 4	Port 5	Port 6	Port 7	Port 8	Port 9	Port 10
6.76	7.64	7.89	7.65	8.43	8.92	9.02	10.88	11.65	12.75

[a]Based on monthly data for the period 1963.1 through 2011.7. These (value weighted) portfolios are reconstructed (i.e., all the Compustat stocks are reassigned to one of the 10 portfolios) at the end of June of each year based on the end of the previous year's BE and ME values. We thank Tano Santos for making this data available to us.

that they are often described as constituting the "equity premium puzzle."[10] Furthermore, despite the substantial risk premia evident in Tables 2.3 and 2.4, most of the US population owns very little stock: the 2007 Survey of Consumer Finance reports that about one-half of US households own no stock at all; for high-income households 23% own no stock. This fact constitutes a second "puzzle," at least as regards the modern theory of portfolio composition (Chapters 6 and 7).

The equity premium is a secular time series property of stock returns.

2.5.2 The Value Premium

The value premium is a statement about the cross section of stock returns. It is the empirically robust observation that stocks with a higher (book value of equity)/(market value of equity) (BE/ME) ratio have, on average, higher excess returns than stocks with low (BE/ME) values. Consider Table 2.5, which describes the annualized average excess returns on 10 portfolios ($Er_{port\ i} - r_f$, $i = 1, 2, \ldots, 10$) of Compustat stocks sorted on the basis of their (BE/ME) ratio.[11]

Notice that the highest (BE/ME) portfolio has nearly twice the average excess returns of the lowest (BE/ME) portfolio. The return pattern observed in Table 2.5 is known as the "value premium." It is characteristic of international stock markets and all historical time periods. The value premium becomes the "value premium puzzle" because financial theory has no all-encompassing explanation as to why it should be observed. Traditional risk-based theories—the idea that investors dislike risky returns and must therefore be compensated with higher average returns in order to hold, willingly, higher risk assets—cannot explain the pattern of Table 2.5. In particular, the CAPM, which we discuss in

[10] In the models we will consider the supply of equities and risk-free bonds is typically assumed to be constant. This is generally an innocuous assumption: the supply of IBM equity shares outstanding, for example, does not change from year to year. The focus of these models is thus principally to characterize security demands by investors, from which follow (given fixed supplies) equilibrium prices and rates of return. The resulting equity premium usually falls far short of what is manifest in the data.

[11] Compustat is a large publicly accessible database of historical stock returns (and much other information).

Chapter 8, and which is the most widely cited risk-based theory of returns, cannot explain Table 2.5.

The value premium is a secular cross-sectional property of stock returns.

2.5.3 The Term Structure

In what follows we will also explore various features of the bond market and, in particular, the market for default free government securities (e.g., US Treasury securities). In this specific market, the fundamental notion is that of the term structure of nominal interest rates: the family of interest rates on zero-coupon, default-free nominal bonds of progressively greater maturity.[12] More specifically, at any time t, it is the collection of interest rates $\{r_{t,\ 1}, r_{t,\ 2}, \ldots, r_{t,\ J}\}$, where $r_{t,j}$ is the period t interest rate on a default-free security with cash-flow pattern:

T	$t+1$	$t+2\cdots$	$t+j-1\cdots$	$t+j$	$t+j+1\cdots$	$t+J$
$-q^b_{t,j}$	0	0	0	\$1000	0	0

where $q^b_{t,j} = \$1000/((1+\mathrm{r}_{t,j})^j)$ with $q^b_{t,j}$ denoting the security's period t market price. With no uncertainty in the payments, these rates reflect the pure time value of money and, as such, constitute one of the building blocks of any discounting procedure (recall Section 2.2). In this sense, the term structure is a fundamental concept. Accordingly, the basic features of the bond market are usually expressed as properties of the term structure. In particular, we will seek to explain the following:

i. The term structure of interest rates is typically upward sloping: default free discount securities with longer times to maturity typically command higher rates. We want to understand not only why this is true but also what drives the exceptions.
ii. The term structure of interest rates generally moves up or down for all maturities simultaneously. This means that an increase in the short rate is accompanied by an increase in rates for bonds of all maturities. For US Treasury securities, 99% of the variation in returns at any maturity is related to shifts in the entire term structure. This fact stands in stark contrast to its analogous relationship for stocks: roughly 80−90% of the variation in returns to any particular stock is generally *unrelated* to aggregate market movements.

    This observation suggests that there may be one macroeconomic quantity (factor) affecting all default-free rates similarly. Can this factor be identified?

[12] By a "nominal" bond we mean one that pays prefixed dollar (CHF, Euro) amounts, that are not adjusted for inflation.

iii. Lastly, there is the phenomenon of relative volatility: properly assessed, returns on long term US Treasury securities are more volatile than the returns on short-term bonds, even when normalized by their higher expected returns (see topic (i)). Such a result is puzzling in that our immediate intuition would suggest otherwise: long-term bond prices and returns should not be as sensitive to business cycle or other macroeconomic events, as these are generally of shorter duration (than the bond's time to maturity) and thus their consequences for bond returns should tend to "average out" over the long bond's time to maturity. Shiller (1979) refers to this phenomenon as the "bond volatility puzzle."[13]

Taken together, the equity premium (a quantitative assessment) and the three qualitative properties of the term structure detailed above illustrate important stylized facts that good models should be able to replicate.

2.6 Asset Pricing Is Not All of Finance!

2.6.1 Corporate Finance

Intermediate financial theory focuses on the valuation of risky cash flows. Pricing a future (risky) dollar is the dominant ingredient in most financial problems. But it is not all of finance! Our capital markets perspective in particular sidesteps many of the issues surrounding how the firm generates and protects the cash-flow streams to be priced. It is this concern that is at the core of corporate financial theory or simply *corporate finance*.

In a broad sense, corporate finance is concerned with decision making at the firm level whenever it has a financial dimension, has implications for the financial situation of the firm, or is influenced by financial considerations. In particular, it is a field concerned, first and foremost, with the investment decision (what projects should be accepted), the financing decision (what mix of securities should be issued and sold to finance the chosen investment projects), the payout decision (how should investors in the firm, and in particular the equity investors, be compensated), and risk management (how corporate resources should be protected against adverse outcomes). Corporate finance also explores

[13] Strictly speaking, these are statements about default-free coupon bonds. In fact, the US Treasury, for instance, does not issue zero coupon bonds of more than 6 months time to maturity. It is possible, however, to extract the implied default-free zero coupon bond prices from the prices and cash flows associated with default-free coupon bonds, if a sufficient number of distinct types (different cash flows) are issued. The set of IRRs (internal rates of return) on default free coupon bonds of successively greater maturity is referred to as the "yield curve." Its qualitative properties do not differ significantly from those of the term structure (e.g., both move in tandem). Typically there is little quantitative difference as well, and the expressions "term structure" and "yield curve" are often used (incorrectly) interchangeably. See Chapter 11 for a more detailed discussion. The percentages are from Ang (2012).

issues related to the size and the scope of the firm, e.g., mergers and acquisitions and the pricing of conglomerates, the internal organization of the firm, the principles of corporate governance, and the forms of remuneration of the various stakeholders.[14]

All of these decisions individually and collectively influence the firm's free cash-flow stream and, as such, have asset pricing implications. The decision to increase the proportion of debt in the firm's capital structure, for example, increases the riskiness of its equity cash-flow stream and the standard deviation of the equilibrium return on equity.

Of course, when we think of the investment decision itself, the solution to the valuation problem is of the essence. Indeed, many of the issues typically grouped under the heading of capital budgeting are intimately related to the focus of the present text. We will be silent, however, on most of the other issues listed above, which are better viewed as arising in the context of bilateral (rather than market) relations and, as we will see, in situations where asymmetries of information play a dominant role.

The goal of this section is to illustrate the difference in perspectives by reviewing, selectively, the corporate finance literature, particularly as regards the capital structure of the firm and contrasting it with the capital markets perspective that we will be adopting throughout this text. In so doing, we also attempt to give the flavor of an important research area while reminding the reader of the many important topics this text elects not to address.

2.6.2 Capital Structure

We focus on the capital structure issue in Chapter 17 where we explore the assumption underlying the famous Modigliani–Miller irrelevance result: in the absence of taxes, subsidies, and contracting costs, the value of a firm is independent of its capital structure if the firm's investment policy is fixed and financial markets are complete. Our emphasis will concern how this result fundamentally rests on the complete markets assumption.

The corporate finance literature has not ignored the completeness issue but rather has chosen to explore its underlying causes, most specifically information asymmetries between the various agents concerned, managers, shareholders, and so forth.[15] While we touch on

[14] The recent scandals (Hewlett-Packard, AIG) in the United States, Europe (gigantic trading losses in JPM Chase, UBS, and Société Générale, due to rogue or unsupervised trading) and in Japan (Olympus Optical) place in stark light the responsibilities of boards of directors for ultimate firm oversight as well as their frequent failure to provide it. The large question here is what sort of board structure is consistent with superior long-run firm performance?

[15] Tax issues have tended to dominate the corporate finance capital structure debate until recently, and we will review this arena shortly. The relevance of taxes is not a distinguishing feature of the corporate finance perspective alone. Taxes also matter when we think of valuing risky cash flows, although we will have very little to say about it except that all the cash flows we consider are to be thought of as after-tax cash flows.

the issue of heterogeneity of information in a market context, we do so only in Chapter 18, emphasizing there that heterogeneity raises a number of tough modeling difficulties. These difficulties justify the fact that most of capital market theory either is silent on the issue of heterogeneity (in particular, when it adopts the arbitrage approach) or explicitly assumes homogeneous information on the part of capital market participants.

In contrast, the bulk of corporate finance builds on asymmetries of information and explores the various problems they raise. These are typically classified as leading to situations of "moral hazard" or "adverse selection." An instance of the former is when managers are tempted to take advantage of their superior information to implement investment plans that may serve their own interests at the expense of those of shareholders or debtholders. An important branch of the literature concerns the design of contracts, which take moral hazard into account. The choice of capital structure, in particular, will be seen potentially to assist in their management (see, for example, Zwiebel, 1996).

A typical situation of adverse selection occurs when information asymmetries between firms and investors make firms with "good" investment projects indistinguishable to outside investors from firms with poor projects. This suggests a tendency for all firms to receive the same financing terms (a so-called pooling equilibrium where firms with less favorable prospects may receive better than deserved financing arrangements). Firms with good projects must somehow indirectly distinguish themselves in order to receive the more favorable financing terms they merit. For instance, they may want to attach more collateral to their debt securities, an action that firms with poor projects may find too costly to replicate (see, for example, Stein, 1992). Again, the capital structure decision may sometimes help in providing a resolution of the "adverse selection" problem. Below we review the principal capital structure perspectives.

2.6.3 Taxes and Capital Structure

Understanding the determinants of a firm's capital structure (the proportion of debt and equity securities it has outstanding in value terms) is the classical problem in corporate finance. Its intellectual foundations lie in the seminal work of Modigliani and Miller (1958), who argue for capital structure irrelevance in a world without taxes and with complete markets (an hypothesis that excludes information asymmetries).

The corporate finance literature has also emphasized the fact that when one security type receives favored tax treatment (typically, this is debt via the tax deductibility of interest), then the firm's securities become more valuable in the aggregate if more of that security is issued, since to do so is to reduce the firm's overall tax bill and thus enhance the free cash flow to the security holders. Since the bondholders receive the same interest and principal

payments, regardless of the tax status of these payments from the firm's perspective, any tax-based cash-flow enhancement is captured by equity holders. Under a number of further specialized assumptions (including the hypothesis that the firm's debt is risk-free), these considerations lead to the classical relationship

$$V_{\mathrm{L}} = V_{\mathrm{U}} + \tau D$$

The value of a firm's securities under partial debt financing (V_{L}, where L denotes leverage in the capital structure) equals its value under all equity financing (V_{U}, where U denotes unlevered or an all-equity capital structure) plus the present value of the interest tax subsidies. This latter quantity takes the form of the corporate tax rate (τ) times the value of debt outstanding (D) when debt is assumed to be perpetual (unchanging capital structure).

In return terms, this value relationship can be transformed into a relationship between levered and unlevered equity returns:

$$r^e_{\mathrm{L}} = r^e_{\mathrm{U}} + (1-\tau)(D/\mathrm{E})(r^e_{\mathrm{U}} - r_{\mathrm{f}})$$

i.e., the return on levered equity, r^e_{L}, is equal to the return on unlevered equity, r^e_{U}, plus a risk premium due to the inherently riskier equity cash flow that the presence of the fixed payments to debt creates. This premium, as indicated, is related to the tax rate, the firm's debt/equity ratio (D/E), a measure of the degree of leverage, and the difference between the unlevered equity rate and the risk-free rate, r_{f}. Immediately we observe that capital structure considerations influence not only expected equilibrium equity returns via

$$Er^e_{\mathrm{L}} = Er^e_{\mathrm{U}} + (1-\tau)D/E(Er^e_{\mathrm{U}} - r_{\mathrm{f}})$$

where E denotes the expectations operator, but also the variance of returns since

$$\sigma^2_{r^e_{\mathrm{L}}} = (1-(1+\tau)D/E)^2\sigma^2_{r^e_{\mathrm{U}}} > \sigma^2_{r^e_{\mathrm{U}}}$$

under the mild assumption that r_f is constant in the very short run. These relationships illustrate but one instance of corporate financial considerations affecting the patterns of equilibrium returns as observed in the capital markets.

The principal drawback to this tax-based theory of capital structure is the natural implication that if one security type receives favorable tax treatment (usually debt), then if the equity share price is to be maximized the firm's capital structure should be composed exclusively of that security type—i.e., all debt, which is not observed. More recent research in corporate finance has sought to avoid these extreme tax-based conclusions by balancing

the tax benefits of debt with various costs of debt, including bankruptcy and agency costs.[16] Our discussion broadly follows Harris and Raviv (1991).

2.6.4 Capital Structure and Agency Costs

This important segment of the literature seeks to explain financial decisions by examining the conflicts of interests among claimholders within the firm. Although agency conflicts can take a variety of forms, most of the literature has focused on manager's incentives to increase investment risk—the asset substitution problem—or to reject positive Net Present Value (NPV) projects—the underinvestment problem. Both of these conflicts increase the cost of debt and thus reduce the firm's value-maximizing debt ratio.

Another commonly discussed determinant of capital structure arises from manager–stockholder conflicts. Managers and shareholders have different objectives. In particular, managers tend to value investment more than shareholders do. Although there are a number of potentially powerful internal mechanisms to control managers, the control technology normally does not permit the costless resolution of this conflict between managers and investors. Nonetheless, the cash-flow identity implies that constraining financing, hedging, and payout policy places indirect restrictions on investment policy. Hence, even though investment policy is not contractible, by restricting the firm in other dimensions, it is possible to limit the manager's choice of an investment policy. For instance, Jensen (1986) argues that debt financing can increase firm value by reducing the free cash flow. This idea is formalized in more recent papers by Stulz (1990) and Zwiebel (1996). Also, by reducing the likelihood of both high and low cash flows, risk management can control not only shareholders' underinvestment incentives but managers' ability to overinvest as well.

More recently, the corporate finance literature has put some emphasis on the cost that arises from conflicts of interest between controlling and minority shareholders. In most countries, publicly traded companies are not widely held but rather have controlling shareholders. Moreover, these controlling shareholders have the power to pursue private benefits at the expense of minority shareholders, within the limits imposed by investor protection. The recent "law and finance" literature following Shleifer and Vishny (1997) and La Porta et al. (1998) argues that the expropriation of minority shareholders by the controlling shareholder

[16] There are many other proposed capital structure theories. Lee (2014), for example, proposes that firms eschew present and future available tax benefits to debt financing and rather maintain large cash balances in order to finance unexpected future investments and as a precaution against bad times (periods of low cash-flow generation). Such firms are unable to raise cash at times of critical need by selling existing assets because of severe capital adjustment costs.

is at the core of agency conflicts in most countries. While these conflicts have been widely discussed in qualitative terms, the literature has largely been silent on the magnitude of their effects.

2.6.5 The Pecking Order Theory of Investment Financing

The seminal reference here is Myers and Majluf (1984) who again base their work on the assumption that investors are generally less well informed (asymmetric information) than insider-managers vis-à-vis the firm's investment opportunities. As a result, new equity issues to finance new investments may be so underpriced (reflecting average project quality) that NPV positive projects from a societal perspective may have a negative NPV from the perspective of existing shareholders and thus not be financed. Myers and Majluf (1984) argue that this underpricing can be avoided if firms finance projects with securities that have more assured payout patterns and thus are less susceptible to undervaluation: internal funds and, to a slightly lesser extent, debt securities, especially risk-free debt. It is thus in the interests of shareholders to finance projects first with retained earnings, then with debt, and lastly with equity. An implication of this qualitative theory is that the announcement of a new equity issuance is likely to be accompanied by a fall in the issuing firm's stock price since it indicates that the firm's prospects are too poor for the preferred financing alternatives to be accessible.

The pecking order theory has led to a large literature on the importance of security design. For example, Stein (1992) argues that companies may use convertible bonds to get equity into their capital structures "through the backdoor" in situations where informational asymmetries make conventional equity issues unattractive. In other words, convertible bonds represent an indirect mechanism for implementing equity financing that mitigates the adverse selection costs associated with direct equity sales. This explanation for the use of convertibles emphasizes the role of the call feature—that will allow good firms to convert the bond into common equity—and costs of financial distress—that will prevent bad firms from mimicking good ones. Thus, the announcement of a convertible bond issue should be greeted with a less negative—and perhaps even positive—stock price response than an equity issue of the same size by the same company.

2.7 Banks

Banks merit our attention at the present juncture because the concepts we have been discussing (e.g., firm leverage) apply as well to banks but to a "unique" degree. To organize our banking discussion, let us first compare the (simplified) balance sheets of a typical industrial firm and a commercial bank.

Balance sheets (all entries measured in value terms)

Industrial Firm		Bank	
Assets	**Liabilities**	**Assets**	**Liabilities**
• Tangible assets (factories, machinery, inventories) • Intangible assets (patents, process technology, trademarks, etc.)	• Debt issued by the firm • Equity of shareholders	• Loans • Cash • Default-free government securities • Other securities (mortgage backed securities, for example)	• Deposits taken by the bank • Debt securities issued by the bank • Equity of shareholders

Note that the deposits of the banks constitute a (large) portion of its liabilities. Pure investment banks, as distinct from commercial banks, are financial companies that extend loans and make equity investments, but they cannot take deposits as a source of funding. The great general-purpose international banks (e.g., Deutsche, UBS, Société Générale, Citi) are both; i.e., they are holding companies with investment banking and commercial banking divisions. In periods of financial distress, either division effectively becomes responsible for the debts of the other since the overall bank is a single legal entity.

For nonfinancial Eurozone corporations, the average (D/E) ratio prior to the financial crisis (2008) was about 0.8; for banks in the Eurozone, the average (D/E) ratio was about 30. Banks were, and continue to be, much more highly leveraged than other firms. In the past, pure investment banks were even more highly leveraged than bank-holding companies in general. Lehman Brothers, for example, was leveraged to a (D/E) ratio exceeding 50. A high leverage ratio is the first unique feature of bank corporations.

A second distinction, already alluded to, concerns the funding banks receive in the form of deposits. Nonbank corporations that take out bank loans or issue long-term bonds know exactly the timing and magnitude of the future interest and principal payments they must make. To the extent they borrow in the form of taking deposits, banks, however, may be required, suddenly and unexpectedly, to repay the "loans" if depositors collectively initiate large aggregate withdrawals. Such an event may occur if depositors suspect that the bank's investments (its loan and securities portfolio) have not maintained their value, leading to the possibility of bankruptcy, and a delayed return of deposits or potential deposit losses. Economists call these circumstances "bank runs."[17] The phenomenon by which some of a bank's funding may suddenly disappear is the second special feature of banks.[18]

[17] To forestall bank runs, governments now provide deposit insurance against deposit losses at least up to a maximum. In the United States, the relevant entity is the Federal Deposit Insurance Corporation (FDIC), which is funded by a small tax on banks.

[18] Anticipating our discussions of portfolio theory, banks can be viewed as very highly leveraged equity portfolios: a large long position in long-term assets financed by a short position in short-term assets.

Lastly, banks are the principle providers of credit to smaller businesses which find it costly to issue their own bonds directly in the capital markets. This notion of credit is very broad: banks provide loans for investment projects, loans for inventories and wage payments, temporary trade financing, etc. When these services become deficient because certain prominent banks are in financial difficulty (a so-called credit crunch), the macroeconomic effects are both negatively severe and long lasting. The ongoing Great Recession is only the most recent case in point. While the bankruptcy of a nonfinancial firm can have devastating effects on the local economy in which it operates, the consequences of bankruptcies or near-bankruptcies among the large players in the international banking system can have adverse national and international consequences. This feature of the banking system constitutes its third unique characteristic. It is an aspect of the "too big to fail" debate.

We return to the unique features of the banking system in web chapter D.

2.8 Conclusions

We have presented four general approaches and two main perspectives on the valuation of risky cash flows. This discussion was meant to provide an organizing principle and a road map for the extended treatment of a large variety of topics on which we are now embarking. We then went on to present some stylized facts of the financial markets and used these to define what a good financial (valuation) model should be. Our brief excursion into corporate finance was intended to suggest some of the agency issues that are part and parcel of a firm's cash-flow determination. That we have elected to focus on pricing issues surrounding those cash-flow streams does not diminish the importance of the many issues surrounding their creation.

References

Ang, A., 2012. Fixed Income. Columbia Business School, Mimeo.

Chen, N.F., Roll, R., Ross, S.A., 1986. Economic forces and the stock market. J. Bus. 59, 383–404.

Dimson, E., Marsh, P., Staunton, M., 2010. Triumph of the Optimists: 101 Years of Global Investment Returns. Princeton University Press, Princeton, NJ.

Harris, M., Raviv, A., 1991. The theory of capital structure. J. Finan. 46, 297–355.

Jensen, M., 1986. Agency costs of free cash flow, corporate finance, and takeovers. Am. Econ. Rev. 76, 323–329.

Lee, J.H., 2014. Debt Servicing Costs and Capital Structure, Working Paper, Columbia University, Department of Economics.

Lucas Jr., R.E., 1980. Methods and problems in business cycle theory. J. Money Credit Bank. 12, 696–715.

Mehra, R., 2012. Consumption-based asset pricing models. Annu. Rev. Finan. Econ. 4, 13.1–13.25.

Modigliani, F., Miller, M., 1958. The cost of capital, corporate finance, and the theory of investment. Am. Econ. Rev. 48, 261–297.

Myers, S., Majluf, N., 1984. Corporate financing and investment decisions when firms have information that investors do not have. J. Finan. Econ. 13, 187–221.

La Porta, R., Lopes de Silanes, F., Shleifer, A., Vishny, R., 1998. Law and finance. J. Polit. Econ. 106, 1113–1155.
Shiller, R., 1979. The volatility of long-term interest rates and expectations models of the term structure. J. Polit. Econ. 87, 1190–1219.
Shleifer, A., Vishny, R., 1997. A survey of corporate governance. J. Finan. 52, 737–783.
Stein, J., 1992. Convertible bonds as backdoor equity financing. J. Finan. Econ. 32, 3–23.
Stulz, R., 1990. Managerial discretion and optimal financial policies. J. Finan. Econ. 26, 3–27.
Zwiebel, J., 1996. Dynamic capital structure under managerial entrenchment. Am. Econ. Rev. 86, 1197–1215.

CHAPTER 3

Making Choices in Risky Situations

Chapter Outline

3.1 Introduction

The first stage of the equilibrium perspective on asset pricing consists of developing an understanding of the determinants of the **demand** for securities of various risk classes. Individuals demand securities (in exchange for current purchasing power) in their attempt to redistribute income across time and states of nature. This is a reflection of the consumption-smoothing and risk-reallocation function central to financial markets. Our endeavor requires an understanding of three building blocks:

1. How financial risk is defined and measured.
2. How an investor's attitude toward or tolerance for risk is to be conceptualized and then measured.
3. How investors' risk attitudes interact with the subjective uncertainties associated with the available assets to determine an investor's desired portfolio holdings (demands).

In this and the next chapter, we give a detailed overview of points 1 and 2; point 3 is treated in succeeding chapters.

3.2 Choosing Among Risky Prospects: Preliminaries

When we think of the "risk" of an investment, we are typically thinking of uncertainty in the future cash-flow stream to which the investment represents title. Depending on the state of nature that may occur in the future, we may receive different payments and, in particular, much lower payments in some states than others. That is, we model an asset's associated cash flow in any future time period as a **random variable**.

Consider, for example, the investments listed in Table 3.1, each of which pays off next period in either of two equally likely states. We index these states by $\theta = 1,2$ with their respective probabilities labeled π_1 and π_2.

First, this comparison serves to introduce the important notion of **dominance**. Investment 3 clearly dominates both investments 1 and 2 in the sense that it pays as much in all states of nature and strictly more in at least one state. The **state-by-state dominance** illustrated here is the strongest possible form of dominance. Without any qualification, we will assume that all rational individuals would prefer investment 3 to either of the other two. Basically, this means that we are assuming the typical individual to be nonsatiated in consumption: she desires more rather than less of the consumption goods these payoffs allow her to buy.

In the case of dominance, the choice problem is trivial and, in some sense, the issue of defining risk is irrelevant. The ranking defined by the concept of dominance is, however, very incomplete. If we compare investments 1 and 2, we see that neither dominates the other. Although it performs better in state 2, investment 2 performs much worse in state 1. There is no ranking possible on the basis of the dominance criterion. The different prospects must be characterized from a different angle. The concept of risk enters necessarily.

On this score, we would probably all agree that investments 2 and 3 are comparatively riskier than investment 1. Of course, for investment 3, the dominance property means that the only risk is an upside risk. Yet, in line with the preference for smooth consumption discussed in Chapter 1, the large variation in date 1 payoffs associated with investment 3 is to be viewed as undesirable in itself. When comparing investments 1 and 2, the qualifier

Table 3.1: Asset payoffs ($)

		Value at $t = 1$	
		$\pi_1 = \pi_2 = 1/2$	
	Cost at $t = 0$	$\theta = 1$	$\theta = 2$
Investment 1	−1000	1050	1200
Investment 2	−1000	500	1600
Investment 3	−1000	1050	1600

"riskier" undoubtedly applies to the latter. In the worst state, the payoff associated with 2 is much lower; in the best state it is substantially higher.

These comparisons can alternatively, and often more conveniently, be represented if we describe investments in terms of their performance on a per dollar basis. We do this by computing the state-contingent rates of return (ROR) that we will typically associate with the symbol r. In the case of the above investments, we obtain the results given in Table 3.2.

One sees clearly that all rational individuals should prefer investment 3 to the other two and that this same dominance cannot be expressed when comparing 1 and 2.

The fact that investment 2 is riskier, however, does not mean that all rational risk-averse individuals would necessarily prefer 1. Risk is not the only consideration, and the ranking between the two projects is, in principle, preference dependent. This is more often the case than not; dominance usually provides a very incomplete way of ranking prospects. This fact suggests we must turn to a description of preferences, the main objective of this chapter.

The most well-known approach at this point consists of summarizing such investment return distributions (i.e., the random variables representing returns) by their mean (Er_i) and variance (σ_i^2), $i = 1,2,3$. The variance (or its square root, the standard deviation) of the rate of return is then naturally used as the measure of "risk" of the project (or the asset). For the three investments just listed, we have:

$$Er_1 = 12.5\%; \quad \sigma_1^2 = \frac{1}{2}(5-12.5)^2 + \frac{1}{2}(20-12.5)^2 = (7.5)^2, \quad \text{or} \quad \sigma_1 = 7.5\%$$

$$Er_2 = 5\%; \quad \sigma_2 = 55\% \text{ (similar calculation)}$$

$$Er_3 = 32.5\%; \quad \sigma_3 = 27.5\%$$

If we decided to summarize these return distributions by their means and variances only, investment 1 would clearly appear more attractive than investment 2: it has both a higher mean return and a lower variance. In terms of the mean−variance criterion, investment 1 dominates investment 2; 1 is said to *mean−variance dominate* 2. Our previous discussion makes it clear that **mean−variance dominance** neither implies nor is implied by state-by-state dominance. Investment 3 mean−variance dominates 2 but not 1, although it dominates

Table 3.2: State-contingent ROR (r)

	$\theta = 1$	$\theta = 2$
Investment 1	5%	20%
Investment 2	−50%	60%
Investment 3	5%	60%

them both on a state-by-state basis! This is surprising and should lead us to be cautious when using any mean−variance return criterion. Later, we will detail circumstances where it is fully reliable. At this point, let us anticipate that it is not generally so and that restrictions will have to be imposed to legitimize its use.

The notion of mean−variance dominance, which plays a prominent role in modern portfolio theory, can be expressed in the form of a criterion for selecting investments of equal magnitude:

1. For investments of the same Er, choose the one with the lowest σ.
2. For investments of the same σ, choose the one with the greatest Er.

In the framework of modern portfolio theory, one could not understand a rational agent choosing investment 2 rather than investment 1.

We cannot limit our inquiry to the concept of dominance, however. Mean−variance dominance provides only an incomplete ranking among uncertain prospects, as Table 3.3 illustrates.

When we compare these two investments, we do not clearly see which is best; there is no dominance in either state-by-state or mean−variance terms. Investment 5 is expected to pay 1.25 times the expected return of investment 4, but, in terms of standard deviation, it is also 3 times riskier. The choice between 4 and 5, when restricted to mean−variance characterizations, would require specifying the terms at which the decision maker is willing to *substitute* expected return for a given risk reduction. In other words, what decrease in expected return is the decision maker willing to accept for a 1% decrease in the standard deviation of returns? Or, conversely, does the 1 percentage point additional expected return associated with investment 5 adequately compensate for the (3 times) larger risk? Responses to such questions are preference dependent (i.e., they vary from individual to individual).

Suppose, for a particular individual, the terms of the trade-off are well represented by the index E/σ. Since $(E/\sigma)_4 = 4$ while $(E/\sigma)_5 = 5/3$, investment 4 is better than investment 5 for that individual. Of course, another investor may be less risk averse; i.e., he may be willing to accept more extra risk for the same expected return. For example, his preferences may be

Table 3.3: State-contingent ROR (r)

	$\theta = 1$	$\theta = 2$
Investment 4	3%	5%
Investment 5	2%	8%
	$\pi_1 = \pi_2 = \frac{1}{2}$ $ER_4 = 4\%$; $\sigma_4 = 1\%$ $ER_5 = 5\%$; $\sigma_5 = 3\%$	

adequately represented by $(E - 1/3\sigma)$ in which case he would rank investment 5 (with an index value of 4) above investment 4 (with a value of $3\frac{2}{3}$).[1]

All these considerations strongly suggest that we have to adopt a more general viewpoint for comparing potential return distributions. This viewpoint is part of utility theory, to which we now turn after describing some of the problems associated with the empirical characterization of return distributions in Box 3.1.

BOX 3.1 Computing Means and Variances in Practice

Useful as it may be conceptually, calculations of distribution moments such as the mean and the standard deviation are difficult to implement in practice: we rarely know what the *future* states of nature are, let alone their probabilities. We also do not know the returns in each state. A frequently used proxy for a future return distribution is its historical distribution. This amounts to selecting a historical time period and a periodicity, say monthly prices for the past 60 months, and computing the historical (net) returns as follows:

a. Discrete compounding

$$r_j^e = (\text{net})\ \text{return to stock ownership in month}\ j = ((q_j^e + d_j)/q_{j-1}^e) - 1$$

where q_j^e is the price of the stock in month j, and d_j its dividend, if any, that month; $1 + r_j^e$ is referred to as the gross return. We then summarize the past distribution of stock returns by the average historical return and the variance of the historical returns. By doing so, we, in effect, assign an equal probability of $\frac{1}{60}$ to each past observation or event.

b. Continuous compounding

To understand how to compute period-by-period returns "under continuous compounding," we must first explain what this convention entails. Conceptually, continuous compounding is the result of discrete compounding when the corresponding time interval becomes infinitesimally small. Suppose an investor's wealth is Y_0, which he invests at a rate r for one period (let us say a month as in (a)). If the rate r is continuously compounded over this single period, the cumulative wealth consequence is as follows:

$$Y_0 \mapsto Y_0 \lim_{n \mapsto \infty} \left(1 + \frac{r}{n}\right)^n = Y_0 e^r$$

(*Continued*)

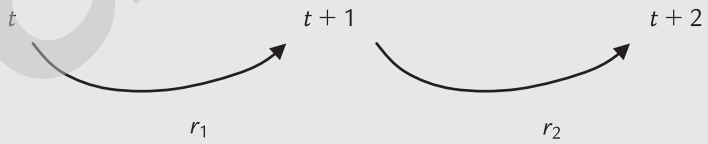
[1] Observe that the proposed index is not immune to the criticism discussed above: investment 3 ($E/\sigma = 1.182$) is inferior to investment 1 ($E/\sigma = 1.667$). Yet, we know that three dominates one because it pays a higher return in every state. This problem is pervasive with the mean−variance investment criterion: whatever the terms of the trade-off between mean and variance or standard deviation, one can produce a paradox such as the one illustrated above. Accordingly, this criterion is not generally applicable without additional restrictions. The index E/σ resembles, but is not identical to, the Sharpe ratio, $(E\tilde{r} - r_f)/\sigma_{\tilde{r}}$, where r_f denotes the risk-free rate.

BOX 3.1 Computing Means and Variances in Practice (Continued)

which exceeds the cumulative effect under discrete compounding ($Y_0 e^r > Y_0(1+r)$ if $r > 0$). It also follows from this identification that if a one period rate r is continuously compounded for a succession of J periods, the cumulative wealth effect will be

$$\begin{array}{ccc} t & & t+J \\ Y_0 & \longmapsto & Y_0 \lim_{n \mapsto \infty} \underbrace{\left\{ \left(1+\frac{r}{n}\right)^n \left(1+\frac{r}{n}\right)^n \cdots \left(1+\frac{r}{n}\right)^n \right\}}_{J \text{ terms}} = Y_0 e^{Jr} \end{array}$$

Lastly, if wealth Y_0 is invested successively at the continuously compounded discrete rates r_1 and r_2, the cumulative effect will be



$$Y_0 \longmapsto Y_0 \lim_{n \mapsto \infty} \left\{ \left(1+\frac{r_1}{n}\right)^n \left(1+\frac{r_2}{n}\right)^n \right\} = Y_0 e^{r_1 + r_2};$$

i.e., under continuous compounding rates of return may simply be added to give their cumulative effect. As we will see in subsequent chapters, this additive feature will allow various calculations to be simplified if the continuous compounding convention is assumed. Note also that our notation assigns the "age interpretation" to time periods: period 10, say, corresponds to the end of the 10th time interval just as a child's 10th birthday is celebrated at the conclusion of his 10th year of life.

Given this setting, how are returns computed from discrete data under the continuous compounding convention? Again, let us assume the following data for some stock is presented to us:

$t+j-1$	$t+j$
q^e_{j-1}	$q^e_j + \text{div}_j$

(*Continued*)

BOX 3.1 Computing Means and Variances in Practice (Continued)

We may now ask the question: what discrete rate of return $r_j^{e,\text{cont.}}$, *when continuously compounded*, was earned by this stock in the course of period j? Equivalently, what rate of return $r_j^{e,\text{cont.}}$ satisfies:

$$q_{j-1}^e \, e^{r_j^{e,\ \text{cont.}}} = q_j^e + \text{div}_j?$$

Thus,

$$r_j^{e,\ \text{cont.}} = \ell n\left(\frac{q_j^e + \text{div}_j}{q_{j-1}^e}\right) = \ln(1 + r_j^e)$$

In what follows in this book, the unspoken assumption is that reported return data is computed under the continuous compounding convention as per the above calculation. Accordingly, we generally do not employ the "cont." superscript.[2]

Using the pattern of historical returns (however measured) to infer properties of the future return distribution makes sense if we think the "mechanism" generating these returns is "stationary": that the future will in some sense closely resemble the past. In practice, this hypothesis is rarely fully verified and, at the minimum, it requires careful checking.[3]

[2] When returns are small, $r_j \approx \ln(1 + r_j)$. This is a standard approximation. Note that continuously compounded returns are the natural logarithm of discrete gross returns and, in this sense, are the continuously compounded cointegral to discrete gross returns.

[3] The accuracy of the mean and variance estimates from historical data as stand-ins to the underlying return distribution's true (future) mean and variance is a topic of great significance and to which we will return (c.f. Chapter 7).

3.3 A Prerequisite: Choice Theory Under Certainty

A good deal of financial economics is concerned with how people make choices. The objective is to understand the systematic part of individual behavior and to be able to predict (at least in a loose way) how an individual will react under specific economic circumstances. Economic theory describes individual behavior as the result of a process of optimization under constraints, the objective to be reached being determined by individual preferences, and the constraints being a function of the person's income or wealth level and of market prices. This approach, which defines the *homo economicus* and the notion of **economic rationality**, is justified by the fact that individual behavior is predictable only to the extent that it is systematic, which must mean that there is an attempt to achieve a well-defined objective. It is not to be taken literally or normatively.[4]

[4] By this we mean that economic science does not *prescribe* that individuals maximize, optimize, or simply behave as if they were doing so. It just finds it productive to summarize the systematic behavior of economic agents with such tools.

To develop this sense of rationality systematically, we begin by summarizing the objectives of investors in the most basic way: we postulate the existence of a preference relation, represented by the symbol $\succeq$, describing investors' ability to compare various bundles of goods and services. For two bundles a and b, the expression

$$a \succcurlyeq b$$

is to be read as follows: For the investor in question, bundle a is either strictly preferred to bundle b, or he is indifferent between them. Pure indifference is denoted by $a \sim b$, strict preference by $a \succ b$.

The notion of economic rationality can then be summarized by the following assumptions:

A.1 Every investor possesses such a preference relation and it is *complete*, meaning that he is able to decide whether he prefers a to b, b to a, or both, in which case he is indifferent with respect to the two bundles. That is, for any two bundles a and b, either $a \succeq b$ or $b \succeq a$, or both. If both hold, we say that the investor is indifferent with respect to the bundles and write $a \sim b$.

A.2 This preference relation satisfies the fundamental property of transitivity: For any bundles a, b, and c, if $a \succeq b$ and $b \succeq c$, then $a \succeq c$.

A further requirement is also necessary for technical reasons:

A.3 The preference relation $\succeq$ is continuous in the following sense: Let $\{x_n\}$ and $\{y_n\}$ be two sequences of consumption bundles such that $x_n \mapsto x$ and $y_n \mapsto y$.[5] If $x_n \succeq y_n$ for all n, then the same relationship is preserved in the limit $x \succeq y$.

A key result can now be expressed in the following proposition.

Theorem 3.1 Assumptions A.1 through A.3 are sufficient to guarantee the existence of a continuous, time-invariant, real-valued utility function[6] u, such that for any two objects of choice (consumption bundles of goods and services),

$$a \succcurlyeq b \quad \text{if and only if} \quad u(a) \geq u(b).$$

Proof See, for example, Mas-Colell et al. (1995), Proposition 3.C.1.

This result asserts that to endow decision makers with a utility function (which they are assumed to maximize) is, in reality, no different than to assume their preferences among objects of choice define a relation possessing the (weak) properties summarized in A.1 through A.3.

[5] We use the standard sense of (normed) convergence in R^N.

[6] In other words, u: $R^N \to R^+$.

Note that Theorem 3.1 implies that if $u(\ )$ is a valid representation of an individual's preferences, any increasing transformation of $u(\ )$ is also valid since such a transformation, by definition, will preserve the ordering induced by $u(\ )$. Note also that the notion of a consumption bundle is, formally, very general. Different elements in a bundle may represent the consumption of the same good or service in different time periods. One element might represent a vacation trip in the Bahamas this year; another may represent exactly the same vacation next year. We can further expand our notion of different goods to include the same good consumed in mutually exclusive states of the world. Our preference for hot soup, for example, may be very different if the day is warm rather than cold. These thoughts suggest that Theorem 3.1 is really quite general and can, formally at least, be extended to accommodate uncertainty. Under uncertainty, however, ranking bundles of goods (or vectors of monetary payoffs, see below) involves more than pure elements of taste or preferences. In the hot soup example, it is natural to suppose that our preferences for hot soup are affected by the probability we attribute to the day being hot or cold. Disentangling pure preferences from probability assessments is the subject to which we now turn.

3.4 Choice Theory Under Uncertainty: An Introduction

Under certainty, the choice is among consumption baskets with known characteristics. Under uncertainty, however, our emphasis changes. The objects of choice are typically no longer consumption bundles but vectors of state-contingent money payoffs (we will reintroduce consumption in Chapter 5). Such vectors are formally what we mean by an investment or an *asset* available for purchase. When we purchase a share of a stock, for example, we know that its sale price in one year will differ depending on what events transpire within the firm and in the world economy. Under financial uncertainty, therefore, the choice is among alternative investments leading to different possible income levels and, hence, ultimately different consumption possibilities. As before, we observe that people do make investment choices, and if we are to make sense of these choices, there must be a stable underlying order of preference defined over different alternative investments. The spirit of Theorem 3.1 will still apply. With appropriate restrictions, these preferences can be represented by a utility index defined on investment possibilities, but obviously something deeper is at work. It is natural to assume that individuals have no intrinsic taste for the assets themselves (IBM stock as opposed to Royal Dutch Petroleum stock (hereafter RDS), for example). Rather, they are interested to know what payoffs these assets will yield and with what likelihood (see Box 3.2, however).

BOX 3.2 Investing Close to Home

Although the assumption that investors only care for the final payoff of their investment without any trace of "romanticism" is standard in financial economics, there is some evidence to the contrary and, in particular, for the assertion that many investors, at the margin at least, prefer to purchase the claims of firms whose products or services are familiar to them. In particular, Huberman (2001) examines the stock ownership records of the seven regional Bell operating companies (RBOCs) (due to a series of mergers, these seven firms have combined into presently two entities, Verizon and AT&T). He discovered that, with the exception of residents of Montana, Americans were more likely to invest in their local RBOC than in any other. When they did, their holdings averaged $14,400. For those who ventured farther from home and hold stocks of the RBOC of a region other than their own, the average holding is only $8246. Considering that every local RBOC cannot be a better investment choice than all of the other six, Huberman interprets his findings as suggesting investors' psychological need to feel comfortable with where they put their money.

One may further hypothesize that investor preferences are indeed very simple after uncertainty is resolved: They prefer a higher monetary payoff to a lower one or, equivalently, to earn a higher return rather than a lower one. Of course they do not know *ex ante* (i.e., before the state of nature is revealed) which asset will yield the higher payoff. They have to choose among prospects, or probability distributions representing these payoffs. And, as we saw in Section 3.2, typically, no one investment prospect will strictly dominate the others. Investors will be able to imagine different possible scenarios, some of which will result in a higher return for one asset, with other scenarios favoring other assets. For instance, let us go back to our favorite situation where there are only two states of nature; in other words, two conceivable scenarios and two assets, as seen in Table 3.4.

There are two key ingredients in the choice between these two alternatives. The first is the probability of the two states. All other things being the same, the more likely is state 1, the more attractive IBM stock will appear to prospective investors. The second is the ex post (once the state of nature is known) level of utility provided by the investment. In Table 3.4, IBM yields $100 in state 1 and is thus preferred to RDS, which yields $90 if this scenario is realized. RDS, however, provides $160 rather than $150 in state 2. Obviously, with

Table 3.4: Forecasted price per share in one period

	State 1	State 2
IBM	$100	$150
RDS	$90	$160
	Current price of both assets is $100.	

unchanged state probabilities, things would look different if the difference in payoffs were increased in one state as in Table 3.5.

Here even if state 1 is slightly more likely, the superiority of RDS in state 2 makes it look more attractive. A more refined perspective is introduced if we go back to our first scenario but now introduce a third contender, Sony, with payoffs of $90 and $150, as seen in Table 3.6.

Sony is dominated by both IBM and RDS. But the choice between the latter two can now be described in terms of an improvement of $10 over the Sony payoff, either in state 1 or in state 2. Which is better? The relevant feature is that IBM adds $10 when the payoff is low ($90), while RDS adds the same amount when the payoff is high ($150). Most people would think IBM more desirable, and with equal state probabilities, would prefer IBM. Once again this is an illustration of the preference for smooth consumption (smoother income allows for smoother consumption).[7] In the present context, one may equivalently speak of risk aversion or of the well-known microeconomic assumption of decreasing marginal utility (the incremental utility steadily declines when adding ever more consumption or income).

The expected utility theorem provides a set of hypotheses under which an investor's preference ranking over investments with uncertain money payoffs may be represented by a utility index combining, in the most elementary way (i.e., linearly), the two ingredients just

Table 3.5: Forecasted price per share in one period

	State 1	State 2
IBM	$100	$150
RDS	$90	$200
	Current price of both assets is $100.	

Table 3.6: Forecasted price per share in one period

	State 1	State 2
IBM	$100	$150
RDS	$90	$160
Sony	$90	$150
	Current price of all assets is $100.	

[7] Of course, for the sake of our reasoning, one must assume that nothing else important is going on simultaneously in the background, and that other things, such as income from other sources, if any, and the prices of the consumption goods to be purchased with the assets' payoffs, are unchanged irrespective of what the payoffs actually are.

discussed—the preference ordering on the ex post money payoffs and the respective probabilities of these payoffs.

We first illustrate this notion in the context of the two assets considered earlier. Let the respective probability distributions on the price per share of IBM and RDS be described, respectively, by $\tilde{p}_{\text{IBM}} = p_{\text{IBM}}(\theta_i)$ and $\tilde{p}_{\text{RDS}} = p_{\text{RDS}}(\theta_i)$ together with the probability π_i that the state of nature θ_i will be realized. In a two-state context, the expected utility theorem provides sufficient conditions on an agent's preferences over uncertain asset payoffs, denoted $\succeq$, such that there exists a function $\mathbb{U}(\ )$, defined over uncertain asset payoffs, and an associated utility-of-money function $U(\ )$ such that

i. $\tilde{p}_{\text{IBM}} \succcurlyeq \tilde{p}_{\text{RDS}}$ if and only if $\mathbb{U}(\tilde{p}_{\text{IBM}}) \geq \mathbb{U}(\tilde{p}_{\text{RDS}})$ where

ii. $$\mathbb{U}(\tilde{p}_{\text{IBM}}) = EU(\tilde{p}_{\text{IBM}}) = \pi_1 U(p_{\text{IBM}}(\theta_1)) + \pi_2 U(p_{\text{IBM}}(\theta_2))$$
$$> \pi_1 U(p_{\text{RDS}}(\theta_1)) + \pi_2 U(p_{\text{RDS}}(\theta_2)) = EU(\tilde{p}_{\text{RDS}}) = \mathbb{U}(\tilde{p}_{\text{RDS}})$$

More generally, for these preferences, the utility of any asset A with payoffs $p_A(\theta_1)$, $p_A(\theta_2)$,. . ., $p_A(\theta_N)$ in the N possible states of nature with probabilities π_1, π_2,. . ., π_N can be represented as

$$\mathbb{U}(\tilde{p}_A) = EU(p_A(\theta_i)) = \sum_{i=1}^{N} \pi_i U(p_A(\theta_i))$$

In other words, by the weighted mean of ex post utilities using the state probabilities as weights. $\mathbb{U}(\tilde{p}_A)$ is a real number. Its precise numerical value, however, has no more meaning than if you are told that the temperature is 40° when you do not know if the scale being used is Celsius or Fahrenheit. It is useful, however, for comparison purposes. By analogy, if it is 40° today, but it will be 45° tomorrow, you at least know it will be warmer tomorrow than it is today. Similarly, the expected utility number is useful because it permits attaching a number to a probability distribution and this number is, under appropriate hypotheses, a good representation of the relative ranking of a particular member of a family of probability distributions (assets under consideration).

3.5 The Expected Utility Theorem

We elect to discuss this theorem in the simple context where objects of choice take the form of simple lotteries. A generic lottery will be denoted (x,y,π); it offers payoff (consequence) x with probability π and payoff (consequence) y with probability $1 - \pi$. This notion of a lottery is actually very general and encompasses a huge variety of possible payoff structures. For example, x and y may represent specific monetary payoffs as in Figure 3.1, or x may be a payment while y is a lottery as in Figure 3.2, or even x and y may both be lotteries as in Figure 3.3. Extending these possibilities, some or all of the x_i's

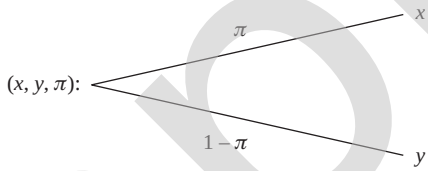


Figure 3.1
A simple lottery (x, y are monetary payoffs).

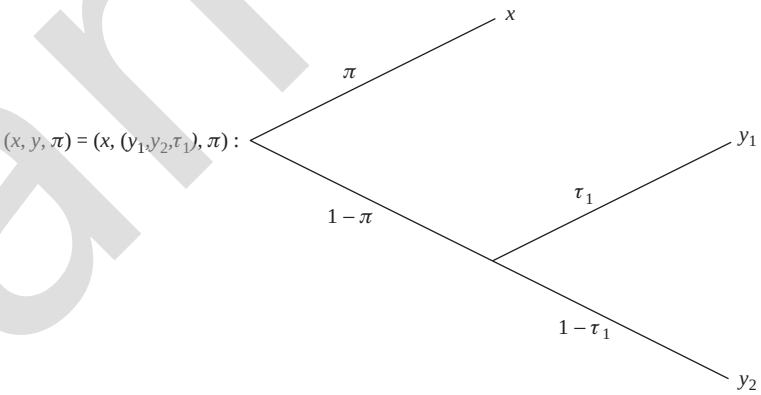


Figure 3.2
A compound lottery (y is itself a lottery).

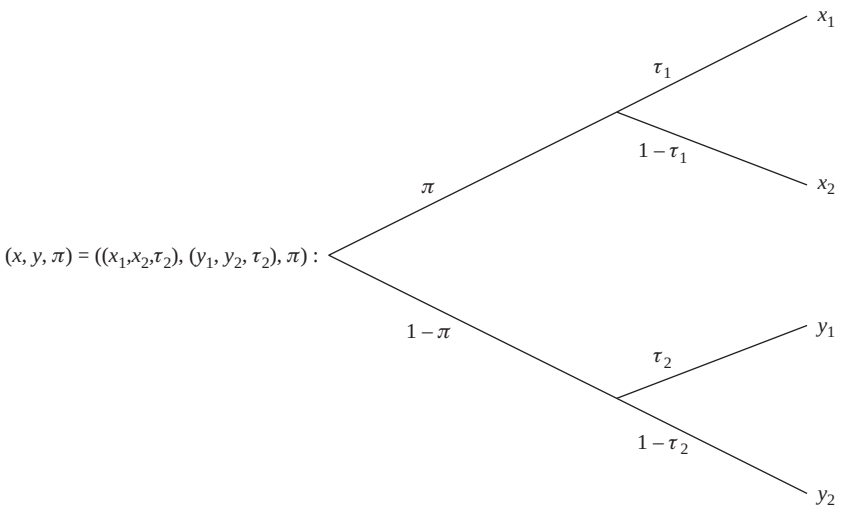


Figure 3.3
A compound lottery (both x and y are themselves lotteries).

and y_i's may themselves be lotteries, and so on. We also extend our choice domain to include individual payments, lotteries where there is one, certain, monetary payoff; for instance,

$$(x, y, \pi) = x \text{ if (and only if) } \pi = 1 \text{ (see axiom C.1)}$$

Moreover, the theorem holds as well for assets paying a continuum of possible payoffs, but our restriction to discrete payoffs makes the necessary assumptions and justifying arguments easily accessible. Our objective is conceptual transparency rather than absolute generality. All the results extend to much more general settings.

Under these representations, we will adopt the following **axioms and conventions**:

C.1. a. $(x,y,1) = x$
b. $(x,y,\pi) = (y,x, 1 - \pi)$
c. $(x,z,\pi) = (x,y,\pi + (1 - \pi)\tau)$ if $z = (x,y,\tau)$

C.1c informs us that agents are concerned with the net cumulative probability of each outcome. Indirectly, it further accommodates lotteries with multiple outcomes; see Figure 3.4, for an example with lotteries (x, y, π'), and $(z, w, \hat{\pi})$, where $\pi_1 = \pi'$, $\pi = \pi_1 + \pi_2$, etc.

C.2. There exists a preference relation $\succeq$, defined on lotteries, which is complete and transitive.

C.3. The preference relation is continuous in the sense of A.3 in Section 3.3.

By C.2 and C.3 alone, we know (Theorem 3.1) that there exists a utility function, which we will denote by $\mathbb{U}(\ )$, defined both on lotteries and on specific payments

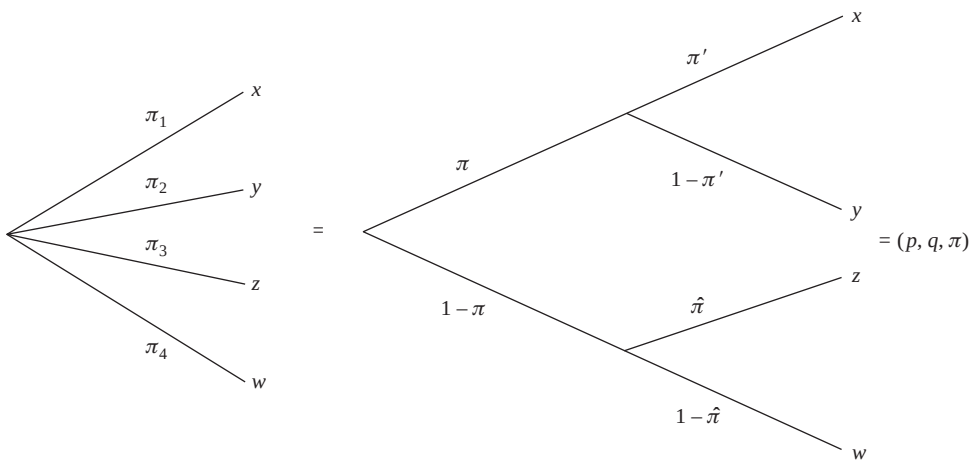


Figure 3.4
A lottery with multiple outcomes reinterpreted as a compound lottery.

since, by assumption C.1a, a payment may be viewed as a (degenerate) lottery. For any payment x, we identify

$$U(x) = \mathbb{U}((x, y, 1)) \tag{3.1}$$

Our remaining assumptions are thus necessary only to guarantee that $\mathbb{U}$ assumes the expected utility form.

C.4. Independence of irrelevant alternatives. Let (x, y, π) and (x, z, π) be any two lotteries; then, $y \succeq z$ if and only if $(x, y, \pi) \succeq (x, z, \pi)$.

C.5. For simplicity, we also assume that there exists a best (i.e., most preferred lottery), b, as well as a worst, least desirable, lottery w.

In our argument to follow (which is constructive, i.e., we explicitly exhibit the expected utility function), it is convenient to use relationships that follow directly from these latter two assumptions. In particular, we will use C.6 and C.7:

C.6. Let x, k, z be consequences or payoffs for which $x > k > z$. Then there exists a π such that $(x, z, \pi) \sim k$.

C.7. Let $x \succ y$. Then $(x, y, \pi_1) \succeq (x, y, \pi_2)$ if and only if $\pi_1 > \pi_2$. This follows directly from C.4.

Theorem 3.2 Consider a preference ordering, defined on the space of lotteries, that satisfies axioms C.1 to C.7. Then there exists a utility function $\mathbb{U}$ defined on the lottery space, with associated utility-of-money function $U(\ )$, such that:

$$\mathbb{U}((x, y, \pi) = \pi U(x) + (1 - \pi)U(y) \tag{3.2}$$

Proof We outline the proof in a number of steps:

By Theorem 3.1 we know that $\mathbb{U}(\ )$ exists with associated $U(\ )$ as its restriction to certain monetary payments, defined as per Eq. (3.1). We must now show that $\mathbb{U}(\ )$ and $U(\ )$ are related by Eq. (3.2).

1. Without loss of generality, we may normalize $\mathbb{U}(\ )$ so that $\mathbb{U}(b) = 1$, $\mathbb{U}(w) = 0$.
2. For all other lotteries z, define $\mathbb{U}(z) = \pi_z$ where π_z satisfies $(b, w, \pi_z) \sim z$
   Constructed in this way $\mathbb{U}(z)$ is well defined since
   a. by C.6, $\mathbb{U}(z) = \pi_z$ exists, and
   b. by C.7, $\mathbb{U}(z)$ is unique. To see this latter implication, assume, to the contrary, that $\mathbb{U}(z) = \pi_z$ and also $\mathbb{U}(z) = \pi'_z$ where $\pi_z > \pi'_z$. By assumption C.4,

$$z \sim (b, w, \pi_z) \succ (b, w, \pi'_z) \sim z, \quad \text{a contradiction}$$

3. It follows also from C.7 that if $m \succ n$, $\mathbb{U}(m) = \pi_m > \pi_n = \mathbb{U}(n)$. Thus, $\mathbb{U}(\ )$ has the property of a utility function.

4. Lastly, we want to show that $\mathbb{U}(\,)$ has the required property. Let x, y be monetary payments, π a probability. By C.1a, $U(x)$, $U(y)$ are well-defined real numbers. By C.6,

$$\begin{aligned}(x, y, \pi) &\sim ((b, w, \pi_x), (b, w, \pi_y)), \pi) \\ &\sim (b, w, \pi\pi_x + (1 - \pi)\pi_y), \text{ by C.1c.}\end{aligned}$$

Thus, by definition of $\mathbb{U}(\,)$,

$$\mathbb{U}((x, y, \pi)) = \pi\pi_x + (1 - \pi)\pi_y = \pi U(x) + (1 - \pi)U(y)$$

Although we have chosen x, y as monetary payments, the same conclusion holds if they are lotteries.

Before going on to refine our understanding of the expected utility theorem, it is important to be absolutely clear on terminology: First, the overall $\mathbb{U}(\,)$ is defined over lotteries. It is referred to as the von Neumann−Morgenstern (VNM) utility function, so named after the originators of the theory, the justly celebrated mathematicians John von Neumann and Oskar Morgenstern. In the construction of a VNM utility function, it is customary first to specify its restriction to certainty monetary payments, the so-called utility-of-money function $U(\,)$ or simply (and hereafter) the *utility function*. Note that the VNM utility function and its associated utility function are not the same. The VNM utility function is defined over uncertain asset payoff structures, while its associated utility function is defined over individual money payments.

The key identifier of the "expected utility" construct is that these two concepts are linearly related: either the utility function is linearly related to the VNM function via the probability weights or the VNM function is linearly related to the state probabilities, with weights being the state-by-state utility-of-money values. The second interpretation leads to the common expression that VNM-expected utility preferences are "linear in the probabilities."

Given the objective specification of probabilities (thus far assumed), it is the utility function that uniquely characterizes an investor. As we will see shortly, different additional assumptions on $U(\,)$ will identify an investor's tolerance for risk. We do, however, impose the maintained requirement that $U(\,)$ be increasing for all candidate utility functions (more money is preferred to less). Note also that the expected utility theorem confirms that investors are concerned only with an asset's final payoffs and the cumulative probabilities of achieving them. For expected utility investors, the structure of uncertainty resolution is thus irrelevant (Axiom C.1c).[8]

Although the introduction to this chapter concentrates on comparing rates of return distributions, our expected utility theorem in fact gives us a tool for comparing different asset

[8] See Section 5.7.1 for a generalization on this score.

payoff distributions. Without further analysis, it does not make sense to think of the utility function as being defined over a rate of return. This is true for a number of reasons. First, returns are expressed on a per unit (per US$, Swiss CHF, etc.) basis and do not identify the magnitude of the initial investment to which these rates are to be applied. We thus have no way to assess the implications of a return distribution for an investor's wealth position. It could, in principle, be anything. Second, the notion of a rate of return implicitly suggests a time interval: the payout is received after the asset is purchased. So far we have only considered the atemporal evaluation of uncertain investment payoffs. In Chapter 4, we generalize the VNM representation to preferences defined over rates of returns.

As in the case of a general order of preferences over bundles of commodities, the VNM-expected utility representation is preserved under linear transformations. If $\mathbb{U}(\,)$ is a von Neuman−Morgenstern utility function, then $\mathbb{V}(\,) = a\mathbb{U}(\,) + b$, where $a > 0$, is also such a function. To verify this assertion, let (x, y, π) be some uncertain payoff, and let $U(\,)$ be the utility-of-money function associated with $\mathbb{U}$.

$$\begin{aligned}\mathbb{V}((x, y, \pi)) &= a\mathbb{U}((x, y, \pi) + b = a[\pi U(x) + (1 - \pi)U(y)] + b \\ &= \pi[a\mathbb{U}(x) + b] + (1 - \pi)[a\mathbb{U}(y) + b] = \pi\mathbb{V}(x) + (1 - \pi)\mathbb{V}(y)\end{aligned}$$

Every linear transformation of an expected utility function is thus also an expected utility function. The utility-of-money function associated with $\mathbb{V}$ is $[aU(\,) + b]$; $\mathbb{V}(\,)$ represents the same preference ordering over uncertain payoffs as $\mathbb{U}(\,)$. On the other hand, a nonlinear transformation does not always respect the preference ordering. It is in that sense that utility is said to be **cardinal**.

Lastly, we need to clarify the direct connection between $U(\,)$ and $u(\,)$ (c.f., Theorem 3.1). Economic science recognizes that money has no value *per se*; its significance lies in the consumption goods that may be purchased with it. Accordingly, consider some financial asset (portfolio) that pays $(Y(\theta_1),\ldots, Y(\theta_N))$, with $Y(\theta_i)$ denoting its money payoff in state θ_i, $i = 1, 2,\ldots, N$. Suppose also that the investor who acquires this asset has available to him J distinct consumption goods in each of the states. The proportions of the consumption goods the investor elects to consume may differ across states reflecting potentially different state-contingent prices as denoted by $(P_1(\theta_i), P_2(\theta_i),\ldots, P_J(\theta_i))$, where $P_j(\theta_i)$ is the price of good j in state θ_i.

Presuming the investor wishes to spend his money as wisely as possible irrespective of what state may be realized, we can define the utility-of-money function $U(Y(\theta_i))$ as the maximum level of consumption utility he may achieve in state i given his income $Y(\theta_i)$ and the consumption goods prices noted above; in effect, we define $U(Y(\theta_i))$ by

$$U(Y(\theta_i)) \equiv_{def} \max_{\{c_1(\theta_i),\ldots,c_J(\theta_i)\}} u(c_1(\theta_i), \ldots, c_J(\theta_i)) \tag{3.3}$$

$$\text{s.t. } c_1(\theta_i)P_1(\theta_i) + \cdots + c_J(\theta_i)P_J(\theta_i) \leq Y(\theta_i)$$

where $c_j(\theta_i)$ is the consumption of good j in state θ_i. The constraint is referred to as the investor's budget constraint. Note that $U(Y(\theta_i))$ subsumes three important quantities: the investor's relative preference for the different goods available in state θ_i (as per $u(\ )$), the relative prices of these goods, and the investor's state θ_i income, $Y(\theta_i)$.

A fuller treatment of this identification would also acknowledge that investors typically save some portion of their income. This consideration requires a multiperiod setting and comes to the fore beginning in Chapter 4.

3.6 How Restrictive Is Expected Utility Theory? The Allais Paradox

Although apparently innocuous, the above set of axioms has been hotly contested as representative of rationality. In particular, it is not difficult to find situations in which investor preferences violate the independence axiom. Consider the following four possible asset payoffs (lotteries):

$$L^1 = (10,000, 0, 0.1) \quad L^2 = (15,000, 0, 0.09)$$
$$L^3 = (10,000, 0, 1) \quad L^4 = (15,000, 0, 0.9)$$

When investors are asked to rank these payoffs, the following ranking is frequently observed:

$$L^2 \succ L^1$$

(presumably because L^2's positive payoff in the favorable state is much greater than L^1's while the likelihood of receiving it is only slightly smaller) and

$$L^3 \succ L^4$$

(Here it appears that the certain prospect of receiving 10,000 is worth more than the potential of an additional 5000 at the risk of receiving nothing.)

By the structure of compound lotteries, however, it is easy to see that:

$$L^1 = (L^3, L^0, 0.1)$$
$$L^2 = (L^4, L^0, 0.1) \quad \text{where } L^0 = (0, 0, 1)$$

By the independence axiom, the ranking between L^1 and L^2, on the one hand, and L^3 and L^4, on the other, should thus be identical!

This is the Allais paradox.[9] There are a number of possible reactions to it.

1. Yes, my choices were inconsistent; let me think again and revise them.
2. No, I'll stick to my choices. The following kinds of things are missing from the theory of choice expressed solely in terms of asset payoffs:
   - the pleasure of gambling, and/or
   - the notion of regret.

The idea of regret is especially relevant to the Allais paradox, and its application in the prior example would go something like this. L^3 is preferred to L^4 because of the regret involved in receiving nothing if L^4 were chosen and the bad state ensued. We would, at that point, regret not having chosen L^3, the certain payment. The expected regret is high because of the nontrivial probability (0.10) of receiving nothing under L^4. On the other hand, the expected regret of choosing L^2 over L^1 is much smaller (the probability of the bad state is only 0.01 greater under L^2, and in either case the probability of success is small), and insufficient to offset the greater expected payoff. Thus L^2 is preferred to L^1.

The Allais paradox is but the first of many phenomena that appear to be inconsistent with standard preference theory. Another prominent example is the general pervasiveness of *preference reversals*, events that may approximately be described as follows. Individuals participating in controlled experiments were asked to choose between two lotteries, (4, 0, 0.9) and (40, 0, 0.1). More than 70% typically chose (4, 0, 0.9). When asked at what price they would be willing to sell the lotteries if they were to own them, however, a similar percentage demanded the higher price for (40, 0, 0.1). At first appearances, these choices would seem to violate transitivity. Let x, y be, respectively, the sale prices of (4, 0, 0.9) and (40, 0, 0.10). Then this phenomenon implies

$$x \sim (4, 0, 0.9) \succ (40, 0, 0.1) \sim y, \quad \text{yet} \quad y > x$$

Alternatively, it may reflect a violation of the assumed principle of procedure invariance, which is the idea that investors' preference for different objects should be indifferent to the manner by which their preference is elicited. Surprisingly, more narrowly focused experiments, which were designed to force a subject with expected utility preferences to behave consistently, gave rise to the same reversals. The preference reversal phenomenon could thus, in principle, be due either to preference intransitivity or to a violation of the independence axiom, or of procedure invariance.

Through a series of carefully constructed experiments, some researchers have attempted to assign responsibility for preference reversals to procedure invariance violations. But this is a particularly alarming conclusion as Thaler (1992) notes. It suggests that "*the context and*

[9] Named after the Nobel Prize-winner Maurice Allais who was the first to uncover the phenomenon. See Allais (1964).

procedures involved in making choices or judgements influence the preferences that are implied by the elicited responses. In practical terms this implies that (economic) behavior is likely to vary across situations which economists (would otherwise) consider identical." This is tantamount to the assertion that the notion of a preference ordering is not well defined. While investors may be able to express a consistent (and thus mathematically representable) preference ordering across television sets with different features (e.g., size of the screen and quality of the sound), this may not be possible with lotteries or consumption baskets containing widely diverse goods.

Grether and Plott (1979) summarize this conflict in the starkest possible terms:

> *Taken at face value, the data demonstrating preference reversals are simply inconsistent with preference theory and have broad implications about research priorities within economics. The inconsistency is deeper than the mere lack of transitivity or even stochastic transitivity. It suggests that no optimization principles of any sort lie behind the simplest of human choices and that the uniformities in human choice behavior which lie behind market behavior result from principles which are of a completely different sort from those generally accepted.*

At this point it is useful to remember, however, that the ultimate goal of financial economics is not to describe individual, but rather market, behavior. There is a real possibility that occurrences of seeming individual irrationality essentially "wash out" when aggregated at the market level. On this score, the proof of the pudding is in the eating and we have little alternative but to see the extent to which the basic theory of choice we are using is able to illuminate financial phenomena of interest. All the while, the discussion above should make us alert to the possibility that unusual phenomena might be the outcome of deviations from the generally accepted preference theory articulated above. While there is, to date, no preference ordering that accommodates preference reversals—and it is not clear there will ever be one—more general constructs than expected utility have been formulated to admit other, seemingly contradictory, phenomena. Further complications arise under collective choice; see Box 3.3.

BOX 3.3 On the Rationality of Collective Decision Making

Although the discussion in the text pertains to the rationality of individual choices, it is a fact that many important decisions are the result of collective decision making. The limitations to such a process are important and, in fact, better understood than those arising at the individual level. It is easy to imagine situations in which transitivity is violated once choices result from some sort of aggregation over more basic preferences.

Consider three portfolio managers who decide which stocks to add to the portfolios they manage by majority voting. The stocks currently under consideration are General Electric

(*Continued*)

BOX 3.3 On the Rationality of Collective Decision Making (Continued)

(GE), Daimler (DAI), and Sony (S). Based on his fundamental research and assumptions, each manager has rational (i.e., transitive) preferences over the three possibilities:

Manager 1: $GE \succeq_1 DAI \succeq_1 S$
Manager 2: $S \succeq_2 GE \succeq_2 DAI$
Manager 3: $DAI \succeq_3 S \succeq_3 GE$

If they were to vote all at once, they know each stock would receive one vote (each stock has its advocate). So they decide to vote on pairwise choices: (GE versus DAI), (DAI versus S), and (S versus GE). The results of this voting (GE dominates DAI, DAI dominates S, and S dominates GE) suggest an intransitivity in the aggregate ordering. Although an intransitivity, it is one that arises from the operation of a collective choice mechanism (voting) rather than being present in the individual orders of preference of the participating agents. There is a large literature on this subject that is closely identified with Arrow's "Impossibility Theorem". See Arrow (1963) for a more exhaustive discussion.

3.7 Behavioral Finance

"The political man of the Greeks, the religious man of the Hebrews and Christians, the enlightened economic man of eighteenth century Europe (the original of that mythical present day character the 'good European') [have] been superseded by a new model for the conduct of life. Psychological man is... more native to American culture than the Puritan sources of that culture would indicate."[10]

The notion of a "rational investor" underlies most of financial theory and, indeed, most of what is presented in this book. A rational investor, very simply, is one with two essential attributes: (i) his preferences over random money payoffs are VNM-expected utility, as just described, and (ii) the probabilities he assigns to these payoffs are objective in that they incorporate all past and present information available to the investor in a manner that respects correct statistical procedure.[11] Unless specified otherwise, VNM-expected utility will represent our default context going forward. But despite the elegant and straightforward nature of the rational investor construct, there remain numerous empirical choice phenomena, which it cannot rationalize. How are they to be understood? One answer to this question lies in the domain of "behavioral finance".

[10] Reiff (2006, page 48). Reiff (2006) is speaking of social trends, but it is not surprising that economic science would be similarly swept along.

[11] In equilibrium contexts, we will go one step further and strengthen the latter requirement to one of "rational expectations." This means that the probabilities objectively computed for the random payoffs in fact coincide with the payoffs' true probability distribution.

Behavioral finance is a theory-in-progress which seeks to fill this gap by departing from the rational investor assumptions in ways that are thought to better reflect various findings in experimental psychology. Most of the assumptions in behavioral finance have not been axiomatized in the context of choices-over-lotteries. Rather, they are supported by circumstantial empirical evidence. We give illustrations of several behavioral notions below. The difficulty in evaluating these concepts in the present context lies precisely in the absence of a formal axiomatic basis underlying them. In that sense we do not yet fully grasp "what they imply."

3.7.1 Framing

"Framing" is simply the notion that individuals' choices may be substantially influenced by the context in which they are presented. As a very simple illustration, would your decision to purchase a steak versus fish for dinner be different if the steak is advertised as:

"90% fat free," or
"10% fat content"?

While *ex post* recognizing that these alternatives convey the same information, it seems apparent that the former description is more likely to elicit a positive "steak" decision for the majority of shoppers.[12]

The same phenomena appear to be present in investment choices. In a classic study, Kahneman and Tversky (1979) explore individual choices across the following lotteries:

i. In the context of first being given \$1000, participants were asked to choose between the following lotteries A and B:
   A: (\$1000, 0, .5)
   B: (\$500, 0, 1)
ii. In a context of first being given \$2000, these same participants were asked to choose between
   C: (−\$1000, 0, .5)
   D: (−\$500, 0, 1).[13]

These lotteries are summarized in Figure 3.5.

[12] Procedure invariance (prior section) and "framing" are not the same notion. Procedure invariance requires that an investor's preference over lotteries be the same irrespective of whether they are directly compared or their certainty equivalents compared; i.e., the ranking is preserved under any methodology as to how the comparison is to be rendered (informally, irrespective of "how the problem is to be solved"). Framing concerns the context of the comparison, once a method of comparison has been chosen.

[13] Kahneman and Tversky (1979) actually conduct their study with payoffs denominated in terms of Israeli currency (lira). At the time, the average monthly income was approximately 3000 lira.

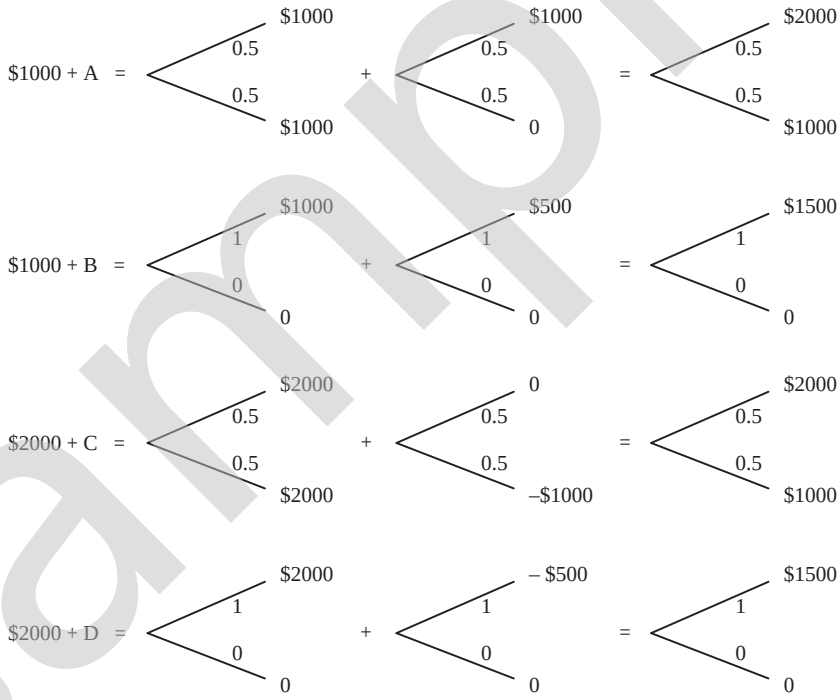


Figure 3.5
Four lotteries with preceding initial payments.

For a majority of those participating in the experiment, B $\succ$ A and C $\succ$ D despite the fact that A and C are equivalent as are B and D when taking full account of the differing initial payments.

How do we interpret these (inconsistent) choices? Apparently, it mattered to most participants that the choice between lotteries A and B was presented as a choice of *gains relative to $1000* while in the second case the choices were presented as *losses relative to $2000.* This distinction is viewed as a manifestation of the phenomenon of framing. Note that under VNM-expected utility, framing, as illustrated above, is irrelevant since only total wealth payoffs matter.

Framing is often cited as one factor potentially contributing to the failure of investors to diversify; i.e., to invest their wealth in portfolios of more than a few (one) assets. It is the idea that when investors consider the acquisition of various assets, they frame the decision on the basis of bilateral comparisons alone, without considering the interaction of multiple asset return patterns and the benefits that may follow. To illustrate this notion in the simple context of three assets and two states of nature θ_1 and θ_2, consider the assets in Figure 3.6.

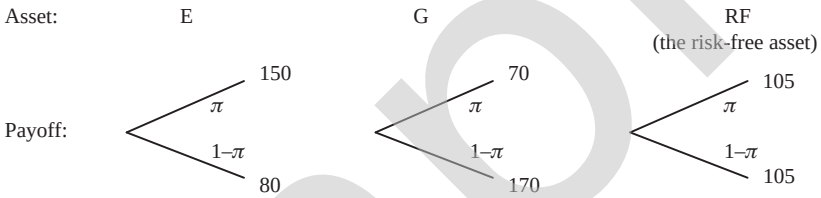


Figure 3.6
Three candidate assets.

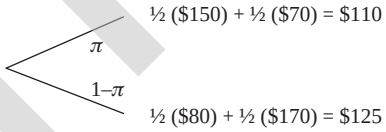


Figure 3.7
Payoff outcomes, portfolio {½E, ½G}.

Assume $\pi = ½$, and let the prices of the assets be $q_E = q_G = q_{RF} = \$100$. Under narrow framing, an investor may individually compare E and G to RF, find each individually less desirable (both E and G have large payoff variances and a substantial probability of significant loss), and end up with a portfolio composed exclusively of asset RF. Nevertheless, we see from Figure 3.7 that RF is clearly state by state *dominated* by the portfolio {½E, ½G}.

It is unclear what utility specification would eliminate all framing phenomena.

3.7.2 Prospect Theory

At present, Kahneman and Tversky's (1979, 1992) Prospect Theory is the most highly developed behavioral theory of choice.[14] It rests on a number of experimental observations:

i Consider the random payoff ($110, −$100, ½). In experimental settings, a majority of participants declined to accept this lottery, irrespective of their level of personal wealth, even if it was offered to them at zero cost.

ii Consider the four basic lotteries from Section 3.7.1

A: ($1000, 0, .5)
B: ($500, 0, 1)
C: (−$1000, 0, .5)
D: (−$500, 0, 1).

[14] Barberis (2013) provides a detailed overview of the theory and applications that have followed from it. This section owes much to him.

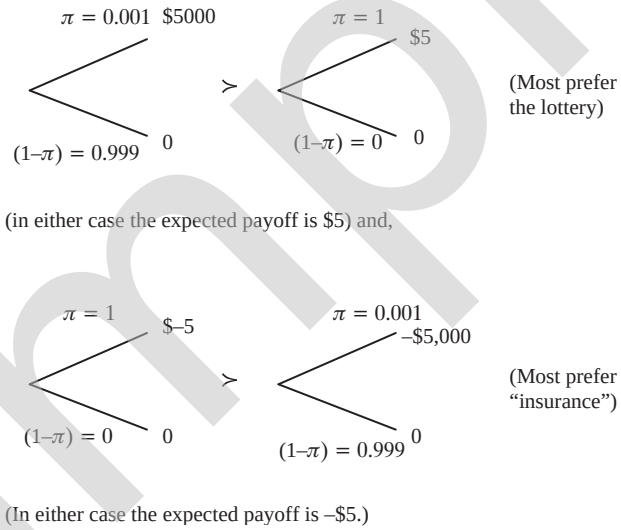


Figure 3.8
Preferences for lotteries and insurance.

When offered for comparison without any prior wealth distinctions (no issues of framing), a majority of participants displayed the following preference:
B ≻ A and C ≻ D.

iii Participants generally displayed a preference for both lotteries and insurance when offered in closely related choice settings; in particular, see Figure 3.8.

Building on these and other observations Kahneman and Tversky (1979) propose a theory of choice under uncertainty (Prospect Theory) with four principal ingredients.

1. Investors ultimately derive utility not from their absolute wealth levels (as in the VNM-expected utility case) but from gains or losses relative to some reference or benchmark value. This (critical) element in their theory is suggested by observation (i): since at very high wealth levels, the acceptance or rejection of ($110, $−100, ½) is really of little consumption consequence, its rejection at all wealth levels suggests that it is the gains or losses themselves, possibly relative to some preconceived benchmark, that really matter to investors. They thus propose a utility-of-money function of the form $U(Y - \overline{Y})$ where $\overline{Y}$ is the benchmark. The benchmark can be thought of as either a minimally acceptable wealth level or, under the proper transformations, a cutoff rate of return. It can be changing through time reflecting prior experience. Unfortunately, Prospect Theory does not offer a general guide as to how the benchmark should be selected in any specific choice setting.
2. Since ($110, $−100, ½) has a positive expected value, its rejection at all wealth levels also suggests that agents feel losses more acutely than gains (of greater comparative

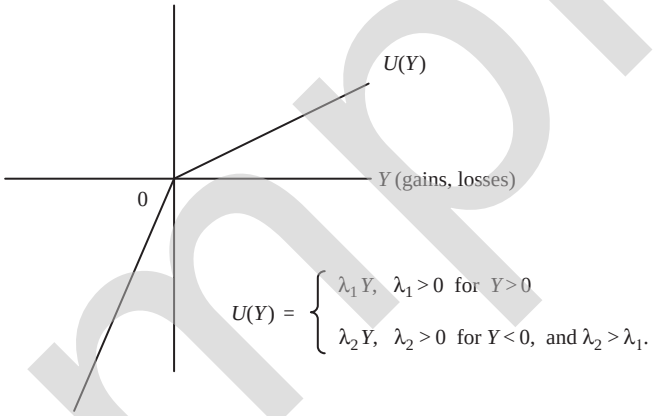


Figure 3.9
Loss-averse utility function.

magnitude). This is the sense of loss-averse preferences or simply "loss aversion." The simplest illustration of a loss-averse utility-of-money function is seen in Figure 3.9, where $\overline{Y} = 0$, a zero benchmark.

3. As a further refinement, consider the choices in observation (ii). The fact that most participants prefer lottery B to lottery A suggests dislike of risk over positive gain lotteries. The preference for lottery C over lottery D, however, suggests "risk loving" behavior over losses. As we will see in the next chapter these features imply that $U(Y - \overline{Y})$ is concave for $Y > \overline{Y}$ but convex for $Y < \overline{Y}$.

   An illustration of a utility representation satisfying (1)−(3) is as follows: Let $\overline{Y}$ denote the benchmark payoff and define the investor's utility-of-money function $U(Y)$ by

$$U(Y) = \begin{cases} \dfrac{(|Y-\overline{Y}|)^{1-\gamma_1}}{1-\gamma_1}, & \text{if } Y \geq \overline{Y} \\ \dfrac{-\lambda(|Y-\overline{Y}|)^{1-\gamma_2}}{1-\gamma_2}, & \text{if } Y \leq \overline{Y} \end{cases}$$

   where $\lambda > 1$ captures the extent of the investor's aversion to "losses" relative to the benchmark, and $\gamma_1 > 0$ and $\gamma_2 > 0$ need not coincide (but $\gamma_1 \neq 1$, $\gamma_2 \neq 1$). In other words, the curvature of the function may differ for deviations above or below the benchmark. See Figure 3.10 for an illustration. Clearly, all three features can have a large impact on the relative ranking of uncertain lottery payoffs.

4. There is a fourth attribute of Prospect Theory that also has its origins in observation. (iii) One interpretation of the choices found in that observation is that investors overweight low probability tail events, both favorable and unfavorable. Following on

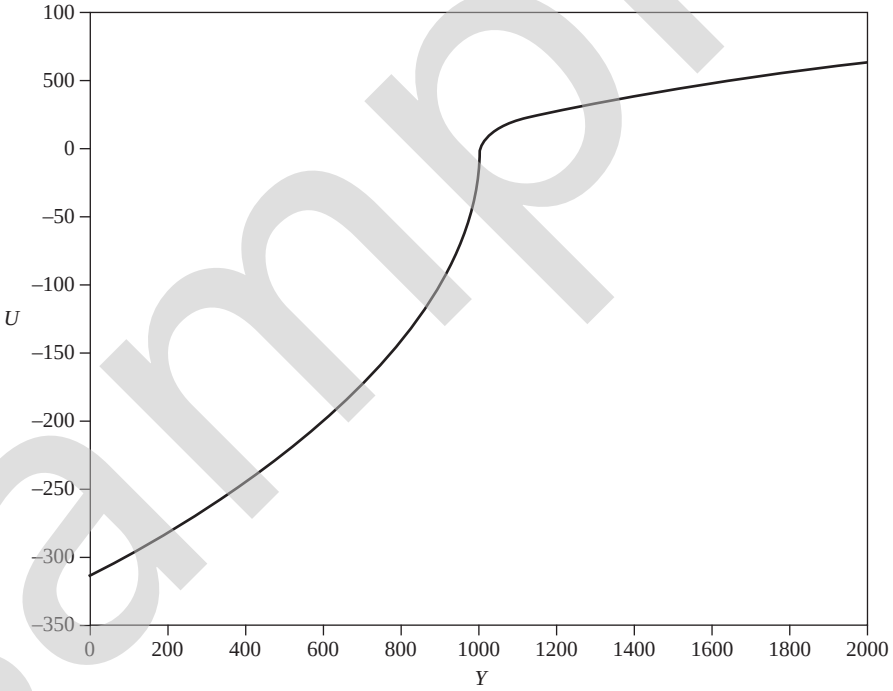


Figure 3.10
Utility function for Prospect Theory.

this possibility, Kahneman and Tversky (1979) elect to weigh the utilities of the various outcomes using a nonlinear function of the true probabilities which is asymmetric in a manner that gives a high weighting to tail events. These weightings are not necessarily to be interpreted as erroneous probability estimates, but as perhaps reflecting relative welfare consequences of the outcomes for the investor that are not observable. See Tversky and Kahneman (1992) for a full discussion. A sample weighting function taken from Tversky and Kahneman (1992) is found in Figure 3.11. Note that this weighting function resembles a probability distribution function where there is a large likelihood of "extreme" events.

As a paradigm for rationalizing laboratory observations (i)−(iv), Prospect Theory has no equal at the moment. But does it have any direct advantage over VNM-expected utility in explaining equilibrium market phenomena? There are at least two relevant works in this regard, Barberis and Huang (2008) and Benartzi and Thaler (1995). In the first paper, the authors show that the skewness in the return distribution of a common stock can have important pricing implications when investors' loss-averse preferences are defined over

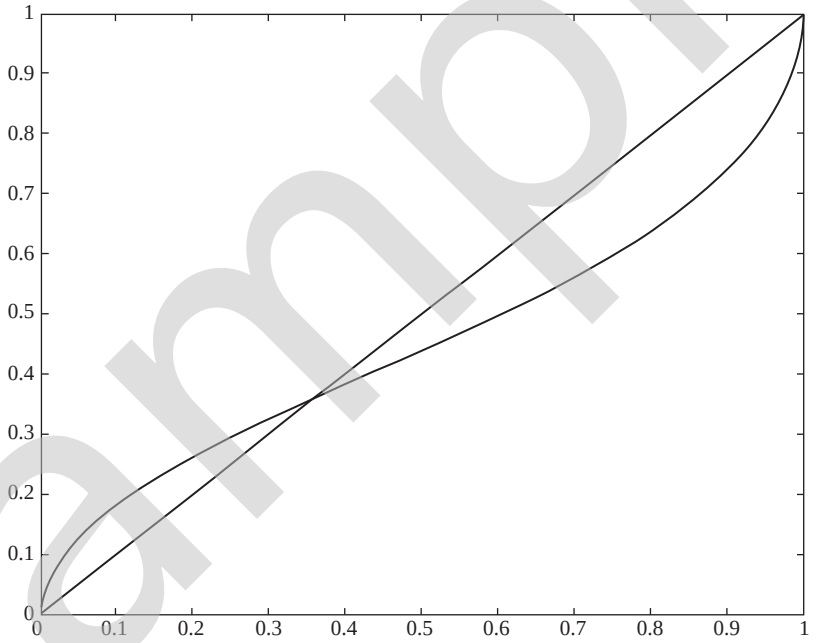


Figure 3.11
The probability weighting function. *Source: This figure is taken from Tversky and Kahneman (1992) as presented in Barberis (2013).*

changes in the value of their portfolios. In particular, securities with positively skewed return distributions are priced higher (and have lower average returns) and negatively skewed securities lower (and have higher average returns) than would be the case in a VNM-expected utility environment, a consequence that follows from the asymmetric weighting function fundamental to Prospect Theory. It is a feature that appears to be present in the cross section of security returns (see, for example, Boyer et al. (2010) and Conrad et al. (2013).

The basic intuition in the Benartzi and Thaler (1995) paper is that loss-averse investors will dislike equity securities because stock market returns are much more highly dispersed relative to bond returns. This fact should lead, they argue, to a higher equity premium in an equilibrium setting where investors have loss-averse preferences than would be possible in the same environment where investors are VNM-expected utility. In a formal dynamic equilibrium quantitative model where investors have loss-averse preferences, Barberis and Huang (2001) go on to confirm the assertions of Benartzi and Thaler (1995), subject to qualifications.

There are many other utility forms that have been proposed over the past 30 years which are linked in some way to Prospect Theory. Some are widely employed in the research literature. We conclude this section by enumerating a few of them below.

3.7.2.1 Preference Orderings with Connections to Prospect Theory

i. Survival benchmark: Investors evaluate lotteries according to $\mathbb{U}(\tilde{Y}) = EU(\tilde{Y} - \overline{Y})$ where $\overline{Y}$ represents a minimum level of income for a decent lifestyle.

ii. Habit formation preferences: Investors become accustomed to a particular income level and measure their utility by the extent to which their present income realization departs from it. For example,

$$\mathbb{U}(\tilde{Y}_t) = EU(\tilde{Y}_t - Y_{t-1})$$

where the habit is identified by the prior period's income (and associated consumption) level, thereby introducing a time dimension. See Constantinides (1990) and Sundaresan (1989).

iii. "Keeping up with the Joneses" or relative income status: Once again investors evaluate lotteries according to

$$\mathbb{U}(\tilde{Y}) = EU(\tilde{Y} - \overline{Y})$$

except that $\overline{Y}$ represents the average income level in the investor's reference community. See Abel (1990).[15]

iv. "Disappointment aversion"; Gul (1991): Here we will need to be a bit more detailed in our representation of the expectations operator E. A disappointment averse investor evaluates lotteries according to

$$\mathbb{U}(\tilde{Y}) = \int_{\underline{Y}}^{B} u(Y)\mathrm{d}F(Y) + AY \int_{B}^{\overline{Y}} u(Y)\mathrm{d}F(Y)$$

where $\underline{Y}$, $\overline{Y}$ denote, respectively, the minimum and maximum payoffs associated with $\tilde{Y}$, A is a number $0 < A < 1$, and B is a certain payment (a certainty equivalent) in exchange for which the investor would be willing to sell the lottery $\tilde{Y}$. With $A < 1$, the investor effectively weighs payments above this benchmark level less heavily than ones below it—a sort of indirect "loss" aversion. He is, essentially, more concerned with low-value outcomes, low in the sense of falling short of the amount for which the investor would have been willing to sell the lottery. Routledge and Zinn (2003) modify the original Gul (1991) representation in a way that endogenizes the construction of B so as to make low payoff realizations even more painful than in Gul (1991).

v. The notion of regret; Loomis and Sugden (1982): The idea here is that if an investor selects one lottery over another, then his utility benefit of a particular state's payment will be diminished had the payoff to the rejected asset in that same state been higher; i.e., given the realized state, the investor experiences regret for not having chosen the

[15] In all these three cases, the benchmark is calculated so that the utility function $U(\ )$ is always defined over a positive quantity.

other asset. More formally, in deciding which of lotteries $\tilde{Y}_1$ and $\tilde{Y}_2$ to choose, the investor computes (for $\tilde{Y}_1$; $\tilde{Y}_2$ is evaluated symmetrically):

$$\begin{aligned}\mathbb{U}(\tilde{Y}_1) &= EU(\tilde{Y}_1) + ER(\tilde{Y}_1 - \tilde{Y}_2) \\ &= \sum_{i=1}^{N} \pi(\theta_1) u(Y_1(\theta_i)) + \delta \sum_{i=1}^{N} \pi(\theta_i) R(Y_1(\theta_i) - Y_2(\theta_i))\end{aligned}$$

where i indexes the states, $\delta > 0$ and, like $U(\ )$, $R(\ )$, the regret function, is a monotone, strictly increasing function which satisfies $R(0) = 0$ and $-R(-\xi) = R(\xi)$ for any $\xi > 0$. $R(\ )$[16] diminishes expected utility in those states where $\tilde{Y}_2$ has the higher outcome. In this sense the alternative asset's payoff serves as the benchmark on a state-by-state basis. Loomes and Sugden (1982) demonstrate that regret preferences can rationalize many of the choice anomalies presented in Kahneman and Tversky (1979). As one might expect, however, regret preferences do not satisfy the VNM transitivity axiom which has the implication that, *per se*, they cannot necessarily be used to isolate the best (highest expected utility) of a collection of eligible lotteries.

Preference representations (i)−(v) are a representative sample of "what's out there". Except for disappointment aversion, all lack a full axiomatic basis grounded in choices-over-lotteries. It is also not clear what the corresponding "reverse engineered" preferences over consumption goods would look like, a problem that is typically sidestepped by assuming one composite consumption good with a normalized price of one so that income and consumption are always numerically equal. In all cases, however, the takeaway is the same: investors evaluate gains and losses relative to a benchmark and do so asymmetrically in a way that tends to "overweight losses" relative to expected utility.

3.7.3 Overconfidence

A variety of studies find evidence that suggests pervasive overconfidence among physicians, nurses, attorneys, engineers, and others.[17] Monitor (2007) finds, in a survey, that more than 70% of professional portfolio managers regard the service they provide as "above average." These various studies measure overconfidence in different ways, consistent with the difficulties inherent in making the notion precise in individual contexts. A more formal study again suggestive of overconfidence is Barber and Odean (2000); see also Odean (1998, 1999). Using data from 78,000 individually managed accounts at a large discount brokerage company, these authors find that these accounts substantially underperform various commonplace benchmarks largely because of the large transaction

[16] Alternatively, $R(\ )$ could assume the form $R(Y_1(\theta_i) - Y_2(\theta_i)) = \min\{0,(Y_1(\theta_i) - Y_2(\theta_i))\}$, etc.

[17] Some references are Baumann et al. (1991), Wagenaar and Kern (1986), Russo and Schoemaker (1992), and De Bont and Thaler (1990) for, respectively, nurses, attorneys, high-level managers, and portfolio managers.