Lecture Notes for the course Investerings- og Finansieringsteori.

David Lando
Rolf Poulsen

January 2002

Contents

Chapter 1

Preface

These notes are intended for the introductory course 'Investerings- og Finansieringsteori' given in the third year of the joint mathematics-economics program at the University of Copenhagen. At this stage they are still far from complete. The notes (the dominant part of which are written by DL) aim to fill a gap between elementary textbooks such as Copeland and Weston[1] or Brealey and Myers[2], and more advanced books which require knowledge of finance theory and often cover continuous-time modelling, such as Duffie[3] and Campbell, Lo and MacKinlay[4] and Leroy and Werner.[5]

Except for a brief introduction to the Black-Scholes model, the aim is to present important parts of the theory of finance through discrete-time models emphasizing definitions and setups which prepare the students for the study of continuous-time models.

At this stage the notes have no historical accounts and hardly references any original papers or existing standard textbooks. This will be remedied in later versions but at this stage, in addition to the books already mentioned, we would like to acknowledge having included things we learned from the classic Hull [6], the also recommendable Luenberger[7], as well as Jarrow and

[1]T. Copeland and F. Weston: Financial Theory and Corporate Policy

[2]Brealey and Myers: Principles of Corporate Finance.McGraw-Hill 4th ed. 1991.

[3]Duffie, D: Dynamic Asset Pricing Theory.
3rd ed. Princeton 2001.

[4]Campbell, J., A. Lo and A.C. MacKinlay: The Econometrics of Financial Markets. Princeton 1997.

[5]LeRoy, S. L. and J. Werner: Principles of Financial Economics, Cambridge 2001.

[6]Hull, J.: Options, Futures and Other Derivative Securities. Prentice-Hall. 4th ed. 1999

[7]Luenberger, D., "Investment Science", Oxford, 1997.

Turnbull[8], and Jensen. [9]

[8]Jarrow R. and S. Turnbull: Derivative Securities.Cincinnati: South-Western (1996).
[9]Jensen, B.A. Rentesregning. DJØFs forlag. 2001.

Chapter 2

Introduction

A student applying for student loans is investing in his or her human capital. Typically, the income of a student is not large enough to cover living expenses, books etc., but the student is hoping that the education will provide future income which is more than enough to repay the loans. The government subsidizes students because it believes that the future income generated by highly educated people will more than compensate for the costs of subsidy, for example through productivity gains and higher tax revenues.

A first time home buyer is typically not able to pay the price of the new home up front but will have to borrow against future income and using the house as collateral.

A company which sees a profitable investment opportunity may not have sufficient funds to launch the project (buy new machines, hire workers) and will seek to raise capital by issuing stocks and/or borrowing money from a bank.

The student, the home buyer and the company are all in need of money to invest now and are confident that they will earn enough in the future to pay back loans that they might receive.

Conversely, a pension fund receives payments from members and promises to pay a certain pension once members retire.

Insurance companies receive premiums on insurance contracts and delivers a promise of future payments in the events of property damage or other unpleasant events which people wish to insure themselves against.

A new lottery millionaire would typically be interested in investing his or her fortune in some sort of assets (government bonds for example) since this will provide a larger income than merely saving the money in a mattress.

The pension fund, the insurance company and the lottery winner are all looking for profitable ways of placing current income in a way which will provide income in the future.

A key role of financial markets is to find efficient ways of connecting the demand for capital with the supply of capital. The examples above illustrated the need for economic agents to substitute income intertemporally. An equally important role of financial markets is to allow risk averse agents (such as insurance buyers) to share risk.

In understanding the way financial markets allocate capital we must understand the chief mechanism by which it performs this allocation, namely through prices. Prices govern the flow of capital, and in financial markets investors will compare the price of some financial security with its promised future payments. A very important aspect of this comparison is the riskiness of the promised payments. We have an intuitive feeling that it is reasonable for government bonds to give a smaller expected return than stocks in risky companies, simply because the government is less likely to default. But exactly how should the relationship between risk and reward (return on an investment) be in a well functioning market? Trying to answer that question is a central part of this course. The best answers delivered so far are in a set of mathematical models developed over the last 40 years or so. One set of models, CAPM and APT, consider expected return and variance on return as the natural definitions of reward and risk, respectively and tries to answer how these should be related. Another set of models are based on arbitrage pricing, which is a very powerful application of the simple idea, that two securities which deliver the same payments should have the same price. This is typically illustrated through option pricing models and in the modelling of bond markets, but the methodology actually originated partly in work which tried to answer a somewhat different question, which is an essential part of financial theory as well: How should a firm finance its investments? Should it issue stocks and/or bonds or maybe something completely different? How should it (if at all) distribute dividends among shareholders? The so-called Modigliani-Miller theorems provide a very important starting point for studying these issues which currently are by no means resolved.

A historical survey of how finance theory has evolved will probably be more interesting at the end of the course since we will at that point understand versions of the central models of the theory.

But let us start by considering a classical explanation of the significance of financial markets in a microeconomic setting.

2.1 The Role of Financial Markets

Consider the definition of a private ownership economy as in Debreu (1959): Assume for simplicity that there is only one good and one firm with pro-

duction set Y. The ith consumer is characterized by a consumption set X_i, a preference preordering $\preceq_i$, an endowment ω_i and shares in the firm θ_i. Given a price system p, and given a profit maximizing choice of production y, the firm then has a profit of $\pi(p) = p \cdot y$ and this profit is distributed to shareholders such that the wealth of the ith consumer becomes

$$w_i = p \cdot \omega_i \ + \theta_i \pi(p) \tag{2.1}$$

The definition of an equilibrium in such an economy then has three seemingly natural requirements: The firm maximizes profits, consumers maximize utility subject to their budget constraint and markets clear, i.e. consumption equals the sum of initial resources and production. But why should the firm maximize its profits? After all, the firm has no utility function, only consumers do. But note that given a price system p, the shareholders of the firm all agree that it is desirable to maximize profits, for the higher profits the larger the consumers wealth, and hence the larger is the set of feasible consumption plans, and hence the larger is the attainable level of utility. In this way the firm's production choice is separated from the shareholders' choice of consumption. There are many ways in which we could imagine shareholders disagreeing over the firm's choice of production. Some examples could include cases where the choice of production influences on the consumption sets of the consumers, or if we relax the assumption of price taking behavior, where the choice of production plan affects the price system and thereby the initial wealth of the shareholders. Let us, by two examples, illustrate in what sense the price system changes the behavior of agents.

Example 1 Consider a single agent who is both a consumer and a producer. The agent has an initial endowment $e_0 > 0$ of the date 0 good and has to divide this endowment between consumption at date 0 and investment in production of a time 1 good. Assume that only non-negative consumption is allowed. Through investment in production, the agent is able to transform an input of i_0 into $f(i_0)$ units of date 1 consumption. The agent has a utility function $U(c_0, c_1)$ which we assume is strictly increasing. The agent's problem is then to maximize utility of consumption, i.e. to maximize $U(c_0, c_1)$ subject to the constraints $c_0 + i_0 \leq e_0$ and $c_1 = f(i_0)$ and we may rewrite this problem as

$$\begin{aligned} \max \ v(c_0) &\equiv U(c_0, f(e_0 - c_0)) \\ s.t. \ c_0 &\leq e_0 \end{aligned}$$

If we impose regularity conditions on the functions f and U (for example that they are differentiable and strictly concave and that utility of zero consumption in either period is -∞) then we know that at the maximum c_0^* we

will have $0 < c_0^* < e_0$ and $v^{'}(c_0^*) = 0$ i.e.

$$D_1U(c_0^*, f(e_0 - c_0^*)) \cdot 1 - D_2U(c_0^*, f(e_0 - c_0^*))f^{'}(e_0 - c_0^*) = 0$$

where D_1 means differentiation after the first variable. Defining i_0^* as the optimal investment level and $c_1^* = f(e_0 - c_0^*)$, we see that

$$f^{'}(i_0^*) = \frac{D_1U(c_0^*, c_1^*)}{{}_2U(c_0^*, c_1^*)}$$

and this condition merely says that the marginal rate of substitution in production is equal to the marginal rate of substitution of consumption.

The key property to note in this example is that what determines the production plan in the absence of prices is the preferences for consumption of the consumer. If two consumers with no access to trade owned shares in the same firm, but had different preferences and identical initial endowments, they would bitterly disagree on the level of the firm's investment.

Example 2 Now consider the setup of the previous example but assume that a price system (p_0, p_1) (whose components are strictly positive) gives the consumer an additional means of transferring date 0 wealth to date 1 consumption. Note that by selling one unit of date 0 consumption the agent acquires $\frac{p_0}{p_1}$ units of date 1 consumption, and we define $1 + r = \frac{p_0}{p1}$. The initial endowment must now be divided between three parts: consumption at date 0 c_0, input into production i_0 and s_0 which is sold in the market and whose revenue can be used to purchase date 1 consumption in the market.

With this possibility the agent's problem becomes that of maximizing $U(c_0, c_1)$ subject to the constraints

$$\begin{aligned} c_0 + i_0 + s_0 &\leq e_0 \\ c_1 &\leq f(i_0) + (1 + r)s_0 \end{aligned}$$

and with monotonicity constraints the inequalities may be replaced by equalities. Note that the problem then may be reduced to having two decision variables c_0 and i_0 and maximizing

$$v(c_0, i_0) \equiv U(c_0, f(i_0) + (1 + r)(e_0 - c_0 - i_0)).$$

Again we may impose enough regularity conditions on U (strict concavity, twice differentiability, strong aversion to zero consumption) to ensure that it attains its maximum in an interior point of the set of feasible pairs (c_0, i_0) and that at this point the gradient of v is zero, i.e.

$$\begin{aligned} D_1U(c_0^*, c_1^*) \cdot 1 - D_2U(c_0^*, f(i_0^*) + (1 + r)(e_0 - c_0^* - i_0^*))(1 + r) &= 0 \\ D_2U(c_0^*, f(i_0^*) + (1 + r)(e_0 - c_0^* - i_0^*))(f^{'}(i_0^*) - (1 + r)) &= 0 \end{aligned}$$

With the assumption of strictly increasing U, the only way the second equality can hold, is if

$$f^{'}(i_0^*) = (1+r)$$

and the first equality holds if

$$\frac{D_1U(c_0^*, c_1^*)}{D_2U(c_0^*, c_1^*)} = (1+r)$$

We observe two significant features:

First, the production decision is independent of the utility function of the agent. Production is chosen to a point where the marginal benefit of investing in production is equal to the 'interest rate' earned in the market. The consumption decision is separate from the production decision and the marginal condition is provided by the market price. In such an environment we have what is known as Fisher Separation where the firm's decision is independent of the shareholder's utility functions. Such a setup rests critically on the assumptions of the perfect competitive markets where there is price taking behavior and a market for both consumption goods at date 0. Whenever we speak of firms having the objective of maximizing shareholders' wealth we are assuming an economy with a setup similar to that of the private ownership economy of which we may think of the second example as a very special case.

Second, the solution to the maximization problem will typically have a higher level of utility for the agent at the optimal point: Simply note that any feasible solution to the first maximization problem is also a solution to the second. This is an improvement which we take as a 'proof' of the significance of the existence of markets. If we consider a private ownership economy equilibrium, the equilibrium price system will see to that consumers and producers coordinate their activities simply by following the price system and they will obtain higher utility than if each individual would act without a price system as in example 1.

Chapter 3

Payment Streams under Certainty

3.1 Security markets and arbitrage

In this section we consider a very simple setup with no uncertainty. There are three reasons that we do this:

First, the terminology of bond markets is conveniently introduced in this setting, for even if there were uncertainty in our model, bonds would be characterized by having payments whose size at any date are constant and known in advance.

Second, the classical NPV rule of capital budgeting is easily understood in this framework.

And finally, the mathematics introduced in this section will be extremely useful in later chapters as well.

A note on notation: If $v \in \mathbb{R}^N$ is a vector the following conventions are used:

- $v \geq 0$ means that all of v's coordinates are non-negative. This we would also write as $v \in \mathbb{R}^N_+ \cup \{0\}$.
- $v > 0$ means that $v \geq 0$ and that at least one coordinate is strictly positive. This we would also write as $v \in \mathbb{R}^N_+$.
- $v \gg 0$ means that every coordinate is strictly positive. This we would also write as $v \in \mathbb{R}^N_{++}$.

Throughout we use $v^\top$ to denote the transpose of the vector v. Vectors without the transpose sign are always thought of as column vectors.

We now consider a model for a financial market with $T+1$ dates: $0, 1, ..., T$ and no uncertainty.

Definition 1 *A security market consists of a pair (π, C) where $\pi \in \mathbb{R}^N$ and C is an $N \times T-$matrix.*

The interpretation is as follows: By paying the price π_i at date 0 one is entitled to a stream of payments $(c_{i1}, ..., c_{iT})$ at dates $1, \ldots, T$. Negative components are interpreted as amounts that the owner of the security has to pay. There are N different payment streams trading. But by forming portfolios, these payment streams can be bought or sold in any quantity and they may be combined in *portfolios* to form new payment streams:

Definition 2 *A portfolio θ is an element of $\mathbb{R}^N$. The payment stream generated by θ is $C^\top\theta \in \mathbb{R}^T$. The price of the portfolio θ at date 0 is $\pi \cdot \theta$.*

Note that allowing portfolios to have negative coordinates means that we allow securities to be sold. We often refer to a negative position in a security as a *short* position and a positive position as a *long* position. Before we even think of adopting (π, C) as a model of a security market we want to check that the price system is sensible. If we think of the financial market as part of an equilibrium model in which the agents use the market to transfer wealth between periods, we clearly want a payment stream of $(1, ..., 1)$ to have a lower price than $(2, ..., 2)$. We also want payment streams that are non-negative at all times to have a non-negative price. More precisely, we want to rule out arbitrage opportunities in the security market model:

Definition 3 *A portfolio θ is an arbitrage opportunity if it satisfies one of the following conditions:*

1. *$\pi \cdot \theta = 0$ and $C^\top\theta > 0$.*
2. *$\pi \cdot \theta < 0$ and $C^\top\theta \geq 0$.*

The interpretation is that it should not be possible to form a portfolio at zero cost which delivers non-negative payments at all future dates and even gives a strictly positive payment at some date. And it should not be possible to form a portfolio at negative cost (i.e. a portfolio which gives the owner money now) which never has a negative cash flow in the future.

Definition 4 *The security market is arbitrage-free if it contains no arbitrage opportunities.*

To give a simple characterization of arbitrage-free markets we need a lemma which is very similar to Farkas' theorem of alternatives (proved in Matematik 2OK using separating hyperplanes):

Lemma 1 *(Stiemke's lemma) Let A be an $n \times m-$matrix: Then precisely one of the following two statements is true:*

1. There exists $x \in \mathbb{R}^m_{++}$ such that $Ax = 0$.
2. There exists $y \in \mathbb{R}^n$ such that $y^\top A > 0$.

We will not prove this lemma here. But it is the key to our next theorem:

Theorem 2 *The security market (π, C) is arbitrage-free if and only if there exists a strictly positive vector $d \in \mathbb{R}^T_{++}$ such that $\pi = Cd$.*

In the context of our security market the vector d will be referred to as a vector of discount factors. This use of language will be clear shortly.

Proof.

Define the matrix

$$A = \begin{pmatrix} -\pi_1 & c_{11} & c_{12} & \cdots & c_{1T} \\ -\pi_2 & c_{21} & c_{22} & \cdots & c_{2T} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\pi_N & c_{N1} & c_{N2} & \cdots & c_{NT} \end{pmatrix}$$

First, note that the existence of $x \in \mathbb{R}^{T+1}_{++}$ such that $Ax = 0$ is equivalent to the existence of a vector of discount factors since we may define

$$d_i = \frac{x_i}{x_0} \qquad i = 1, \ldots, T.$$

Hence if the first condition of Stiemke's lemma is satisfied, a vector d exists such that $\pi = Cd$. The second condition corresponds to the existence of an arbitrage opportunity: If $y^\top A > 0$ then we have either

$$(y^\top A)_1 > 0 \text{ and } (y^\top A)_i \geq 0 \quad i = 1, \ldots, T+1$$

or

$$(y^\top A)_1 = 0 \ , \ y^\top A \geq 0 \text{ and } (y^\top A)_i > 0 \quad \text{some } i \in \{2, \ldots, T+1\}$$

and this is precisely the condition for the existence of an arbitrage opportunity. Now use Stiemke's lemma.

Definition 5 *The security market is complete if for every* $y \in \mathbb{R}^T$ *there exists a* $\theta \in \mathbb{R}^N$ *such that* $C^\top \theta = y$.

In linear algebra terms this means that the rows of C span $\mathbb{R}^T$, and in our interpretation it means that any desired payment stream can be generated by an appropriate choice of portfolio.

Theorem 3 *Assume that* (π, C) *is arbitrage-free. Then the market is complete if and only if there is a unique vector of discount factors.*

Proof. Since the market is arbitrage-free we know that there exists $d \gg 0$ such that $\pi = Cd$. Now if the model is complete $C^\top$is onto and hence C is one-to-one and therefore d must be unique. For the other direction assume that the model is incomplete and hence C is not one-to-one. Then there exists a vector $d' \neq 0$ such that $0 = Cd'$. Since $d \gg 0$, we may choose $\epsilon > 0$ such that $d + \epsilon d' \gg 0$. Clearly, this produces a vector of discount factors different from d. ■

3.2 Zero-coupon bonds and the term structure of interest rates.

Assume throughout this section that the model (π, C) is complete and arbitrage-free and let $d^\top = (d_1, \ldots, d_T)$ be the unique vector of discount factors. Since there must be at least T securities to have a complete model, C must have at least T rows. On the other hand if C has exactly T linearly independent rows, then adding other securities to C will not add any more possibilities of wealth transfer to the market. Hence we can assume that C is a regular $T \times T$ matrix.

Definition 6 *A zero coupon bond with maturity* t *is given by the* t'*th unit vector* e_t *of* $\mathbb{R}^T$.

Next we see why the words 'discount factors' were chosen:

Proposition 4 *The price of a zero coupon bond with maturity* t *is* d_t.

Proof. Let θ_t be the portfolio such that $C^\top \theta_t = e_t$. Then

$$\pi^\top \theta_t = (Cd)^\top \theta_t = d^\top C^\top \theta_t = d^\top e_t = d_t. \quad \blacksquare$$

Note from the definition of d that we get the value of a stream of payments c by computing $\sum_{t=1}^{T} c_t d_t$. In other words, the value of a stream of payments is

obtained by discounting back the individual components. There is nothing in our definition of d which prevents $d_s > d_t$ even when $s > t$,but in the models we will consider this will not be relevant: It is safe to think of d_t as decreasing in t corresponding to the idea that the longer the maturity of a zero coupon bond, the smaller is its value at time 0.

From the discount factors we may derive various types of interest rates which are essential in the study of bond markets.:

Definition 7 *The spot rate at date* 0 *is given by*

$$r_0 = \frac{1}{d_1} - 1.$$

The (one-period) time $t-$ *forward rate at date* 0, *is equal to*

$$f(0,t) = \frac{d_t}{d_{t+1}} - 1,$$

where $d_0 = 1$ *by convention.*

The interpretation of the spot rate should be straightforward: Buying $\frac{1}{d_1}$ units of a maturity 1 zero coupon bond costs $\frac{1}{d_1}d_1 = 1$ at date 0 and gives a payment at date 1 of $\frac{1}{d_1} = 1 + r_0$.

The forward rate tells us the rate at which we may agree at date 0 to borrow (or lend) between dates t and $t+1$. To see this, consider the following strategy at time 0 :

- Sell 1 zero coupon bond with maturity t.
- Buy $\frac{d_t}{d_{t+1}}$ zero coupon bonds with maturity $t+1$.

Note that the amount raised by selling precisely matches the amount used for buying and hence the cash flow from this strategy at time 0 is 0. Now consider what happens if the positions are held to the maturity date of the bonds:

At date t the cash flow is then -1 and at date $t+1$ the cash flow is $\frac{d_t}{d_{t+1}} = 1 + f(0,t)$.

Definition 8 *The yield (or yield to maturity) at time* 0 *of a zero coupon bond with maturity* t *is given as*

$$y(0,t) = \left(\frac{1}{d_t}\right)^{\frac{1}{t}} - 1.$$

Note that

$$d_t(1+y(0,t))^t = 1.$$

and that one may therefore think of the yield as an 'average interest rate' earned on a zero coupon bond. In fact, the yield is a geometric average of forward rates:

$$1+y(0,t) = ((1+f(0,0))\cdots(1+f(0,t-1)))^{\frac{1}{t}}$$

Definition 9 *The term structure of interest rates (or the yield structure of interest rates) at date 0 is given by* $(y(0,1),\ldots,y(0,T))$.

Note that if we have any one of the vector of yields, the vector of forward rates and the vector of discount factors, we may determine the other two. Therefore we could equally well define a term structure of forward rates and a term structure of discount factors. In these notes unless otherwise stated, we think of *the term structure of interest rates* as the yields of zero coupon bonds as a function of time to maturity. It is important to note that the term structure of interest rate depicts yields of zero coupon bonds. We do however also speak of yields on securities which have no negative payments(and some strictly positive payments):

Definition 10 *The yield (or yield to maturity) of a security* $c^\top = (c_1,\ldots,c_T)$ *with* $c > 0$ *and price* π *is the unique solution* $y > -1$ *of the equation*

$$\pi = \sum_{i=1}^{T} \frac{c_i}{(1+y)^i}.$$

Example 3 (Compounding Periods) In most of the analysis in this chapter the time is "stylized"; it is measured in some unit (which we think of and refer to as "years") and cash-flows occur at dates $\{0,1,2,\ldots,T\}$. But it is often convenient (and not hard) to work with dates that are not integer multiples of the fundamental time-unit. We quote interest rates in units of years^{-1} ("per year'), but to any interest rate there should be a number, m, associated stating how often the interest is compounded. By this we mean the following: If you invest 1 \$ for n years at the m-compounded rate r_m you end up with

$$\left(1+\frac{r_m}{m}\right)^{mn}. \tag{3.1}$$

The standard example: If you borrow 1\$ in the bank, a 12% interest rate means they will add 1% to you debt each month (i.e. $m = 12$) and you will end up paying back 1.1268 \$ after a year, while if you make a deposit,

they will add 12% after a year (i.e. $m = 1$) and you will of course get 1.12\$ back after one year. If we keep r_m and n fixed in (3.1) (and then drop the m-subscript) and and let m tend to infinity, it is well known that we get:

$$\lim_{m\to\infty}\left(1+\frac{r}{m}\right)^{mn} = e^{nr},$$

and in this case we will call r the continuously compounded interest rate. In other words: If you invest 1 \$ and the continuously compounded rate r_c for a period af lenght t, you will get back e^{tr_c}. Note also that a continuously compounded rate r_c can be used to find (uniquely for any m) r_m such that 1 \$ invested at m-compounding corresonds to 1 \$ invested at continuous compounding, i.e.

$$\left(1+\frac{r_m}{m}\right)^{m} = e^{r_c}.$$

This means that in order to avoid confusion – even in discrete models – there is much to be said in favor of quoting interest rates on a continuously compounded basis. But then again, in the highly stylized discrete models it would be pretty artificial, so we will not do it (rather it will always be $m = 1$).

3.3 Annuities, serial loans and bullet bonds.

Typically, zero-coupon bonds of all maturities do not trade in financial markets and one therefore has to deduce prices of zero-coupon bonds from other types of bonds trading in the market. Three of the most common types of bonds which do trade in most bond markets are annuities, serial loans and bullet bonds. We now show how knowing to which of these three types a bond belongs and knowing three characteristics, namely the maturity, the principal and the coupon rate, will enable us to determine the bond's cash flow completely.

Let the principal or face value of the bond be denoted F. Payments on the bond start at date 1 and continue to the time of the bond's maturity, which we denote τ. The payments are denoted c_t. We think of the principal of a bond with coupon rate R and payments $c_1, \ldots, c_\tau$ as satisfying the following difference equation:

$$p_t = (1+R)p_{t-1} - c_t \qquad t = 1, \ldots, \tau, \tag{3.2}$$

with the boundary conditions $p_0 = F$ and $p_\tau = 0$.

Think of p_t as the remaining principal right after a payment at date t has been made. For accounting and tax purposes and also as a helpful tool

in designing particular types of bonds, it is useful to split payments into a part which serves as reduction of principal and one part which is seen as an interest payment. We define the reduction in principal at date t as

$$\delta_t = p_{t-1} - p_t$$

and the interest payment as

$$i_t = Rp_{t-1} = c_t - \delta_t.$$

Definition 11 *An annuity with maturity τ, principal F and coupon rate R is a bond whose payments are constant between date 1 and date τ and whose principal evolves according to (3.2).*

Note that with constant payments we may write the remaining principal at time t as

$$p_t = (1+R)^t F - c\sum_{j=0}^{t-1}(1+R)^j \qquad t = 1, 2, \ldots, \tau.$$

To satisfy the boundary condition $p_\tau = 0$ we must therefore have

$$F - c\sum_{j=0}^{\tau-1}(1+R)^{j-\tau} = 0$$

i.e.

$$\begin{aligned} c &= F\left(\sum_{j=0}^{\tau-1}(1+R)^{j-\tau}\right)^{-1} \\ &= F\frac{R(1+R)^\tau}{(1+R)^\tau - 1}. \end{aligned}$$

It is common to use the shorthand notation

$$\alpha_{n\rceil R} = \text{ (“Alfahage”)} = \frac{(1+R)^n - 1}{R(1+R)^n}.$$

Having found what the size of the payment must be we may derive the interest and the deduction of principal as well:

Let us calculate the size of the payments and see how they split into deduction of principal and interest payments.

First, we derive an expression for the remaining principal:

$$\begin{aligned} p_t &= (1+R)^t F - \frac{F}{\alpha_{\tau\rceil R}} \sum_{j=0}^{t-1} (1+R)^j \\ &= \frac{F}{\alpha_{\tau\rceil R}} \left((1+R)^t \alpha_{\tau\rceil R} - \frac{(1+R)^t - 1}{R} \right) \\ &= \frac{F}{\alpha_{\tau\rceil R}} \left(\frac{(1+R)^\tau - 1}{R(1+R)^{\tau-t}} - \frac{(1+R)^\tau - (1+R)^{\tau-t}}{R(1+R)^{\tau-t}} \right) \\ &= \frac{F}{\alpha_{\tau\rceil R}} \alpha_{\tau-t\rceil R}. \end{aligned}$$

This gives us the interest payment and the deduction immediately for the annuity:

$$\begin{aligned} i_t &= R \frac{F}{\alpha_{\tau\rceil R}} \alpha_{\tau-t+1\rceil R} \\ \delta_t &= \frac{F}{\alpha_{\tau\rceil R}} (1 - R\alpha_{\tau-t+1\rceil R}). \end{aligned}$$

Definition 12 *A bullet bond*[1] *with maturity τ,principal F and coupon rate R is characterized by having $i_t = c_t$ for $t = 1, \ldots, \tau - 1$ and $c_\tau = (1+R)F$.*

The fact that we have no reduction in principal before τ forces us to have $c_t = RF$ for all $t < \tau$.

Definition 13 *A serial bond with maturity τ, principal F and coupon rate R is characterized by having δ_t constant for all $t = 1, \ldots, \tau$.*

Since the deduction in principal is constant every period and we must have $p_\tau = 0$, it is clear that $\delta_t = \frac{F}{\tau}$ for $t = 1, \ldots, \tau$. From this it is straightforward to calculate the interest using $i_t = Rp_{t-1}$.

We summarize the characteristics of the three types of bonds in the table below:

	payment	interest	deduction of principal
Annuity	$F\alpha^{-1}_{\tau\rceil R}$	$R\frac{F}{\alpha_{\tau\rceil R}}\alpha_{\tau-t+1\rceil R}$	$\frac{F}{\alpha_{\tau\rceil R}}(1 - R\alpha_{\tau-t+1\rceil R})$
Bullet	RF for $t < \tau$ $(1+R)F$ for $t = \tau$	RF	0 for $t < \tau$ F for $t = \tau$
Serial	$\frac{F}{\tau} + R\left(F - \frac{t-1}{\tau}F\right)$	$R\left(F - \frac{t-1}{\tau}F\right)$	$\frac{F}{\tau}$

[1] In Danish: Et stående lån

Example 4 (A Simple Bond Market) Consider the following bond market where time is measured in years and where payments are made at dates $\{0, 1, \ldots, 4\}$:

Bond (i)	Coupon rate (R_i)	Price at time 0 ($\pi_i(0)$)
1 yr bullet	5	100.00
2 yr bullet	5	99.10
3 yr annuity	6	100.65
4 yr serial	7	102.38

We are interested in finding the zero-coupon prices/yields in this market. First we have to determine the payment streams of the bonds that are traded (the C-matrix). Since $\alpha_{3\rceil 6} = 2.6730$ we find that

$$C = \begin{bmatrix} 105 & 0 & & 0 \\ 5 & 105 & 0 & 0 \\ 37.41 & 37.41 & 37.41 & 0 \\ 32 & 30.25 & 28.5 & 26.75 \end{bmatrix}$$

Clearly this matrix is invertible so $e_t = C^\top \theta_t$ has a unique solution for all $t \in \{1, \ldots, 4\}$ (namely $\theta_t = (C^\top)^{-1} e_t$). If the resulting t-zero-coupon bond prices, $d_t(0) = \pi(0) \cdot \theta_t$, are strictly positive then there is no arbitrage. Performing the inversion and the matrix multiplications we find that

$$(d_1(0), d_2(0), d_3(0), d_4(0))^\top = (0.952381, 0.898458, 0.839618, 0.7774332),$$

or alternatively the following zero-coupon yields

$$100 * (y(0,1), y(0,2), y(0,3), y(0,4))^\top = (5.00, 5.50, 6.00, 6.50).$$

Now suppose that somebody introduces a 4 yr annuity with a coupon rate of 5 % . Since $\alpha_{4\rceil 5} = 3.5459$ this bond has a unique arbitrage-free price of

$$\pi_5(0) = \frac{100}{3.5459}(0.952381 + 0.898458 + 0.839618 + 0.7774332) = 97.80.$$

Notice that bond prices are always quoted per 100 *units* (e.g. \$ or DKK) of principal. This means that if we assume the yield curve is the same at time 1 the price of the serial bond would be quoted as

$$\pi_4(1) = \frac{d_{1:3}(0) \cdot C_{4,2:4}}{0.75} = \frac{76.87536}{0.75} = 102.50$$

(where $d_{1:3}(0)$ means the first 3 entries of $d(0)$ and $C_{4,2:4}$ means the entries 2 to 4 in row 4 of C).

Example 5 (Reading the Financial Pages) This example gives concrete calculations for a specific Danish Government bond traded at the Copenhagen Stock Exchange(CSX): A bullet bond with a 7 % coupon rate and yearly coupon payments that matures on December 15th 2004. On January 4th, 2000 the following information about the bond was available on the homepage of CSX:

Bond type	Maturity date	Price on Jan. 4rd 2000	Yield
7% government bullet	Dec. 15th 2004	106.33	5.50 %

Let us see how the yield was calculated. First, we have to be aware that bond trades are settled 3 trading days later than the trade is agreed upon. So if we buy this bond on Jan. 4th, the first cash-flow occurs on Jan. 7th, or as we will write it: 2000/01/07. And how large is it? By convention we have to pay the price (106.33) plus compensate the seller of the bond for the accrued interest over the period 1999/12/15-2000/01/07. Since there are more than 30 days (by any counting convention) to the next coupon payment we pay an amount to the seller and receive the next coupon. The amount paid in accrued interest is

$$a = \frac{\#\text{days between 1999/12/15 and 2000/01/07}}{360} \times \text{coupon payment.}$$

By (Danish bond market) convention the distance between dates $Y_1/M_1/D_1$ and $Y_2/M_2/D_2$ is

$$(\tilde{D}_2 - \tilde{D}_1) + 30 * (M_2 - M_1) + 360 * (Y_2 - Y_1)$$

where $\tilde{D}_i = \min(D_i, 30)$. This is (one version of) the day-count convention called 30/360. In this case we find that we have to pay $(22/360) * 7 = 0.42778$ DKK in accrued interest. So now we can write down the cash-flows:

Date	t_k	Cash-flow (c_k)	$d_k = (1+y)^{-t_k}$	PV= $d_k * c_k$
2000/01/07	0.00		-106.7578	
2000/12/15	0.93889	7	0.95097	6.6568
2001/12/15	1.93889	7	0.90140	6.3098
2002/12/15	2.93889	7	0.85440	5.9808
2003/12/15	3.93889	7	0.80986	5.6690
2004/12/15	4.93889	107	0.76764	82.1377
SUM				106.7541

And what can we learn from this example? Besides being able to understand and replicate some the numbers we see in the news, we should know that

bond markets have a variety of conventions that are not very homogeneous (settlement takes place after 3 days in Denmark, but after 7 in Euroland; banks typically use actual days when counting; the convention 30/360 does not mean the same in Europe and the U.S., ...) Of course we are not interested in learning the conventions in this (or any?) course, but we must realize that they can be of great practical importance (especially since bond market transactions can be extremely large).

Example 6 The following example is meant to illustrate the perils of relying too much on yields. Especially if they are used incorrectly! The numbers are taken from Jakobsen and Tanggaard.[2] Consider the following small bond market:

Bond (i)	100*Coupon rate (R_i)	Price at time 0 ($\pi_i(0)$)	100*Yield
1 yr bullet	10	100.00	10.00
2 yr bullet	10	98.4	10.93
3 yr bullet	10	95.5	11.87
4 yr bullet	10	91.8	12.74
5 yr bullet	10	87.6	13.58
5 yr serial	10	95.4	11.98

Now consider a portfolio manager with the following argument: "Let us sell 1 of each of the bullet bonds and use the money to buy the serial bond. The weighted yield on our liabilities (the bonds sold) is

$$\frac{100 * 10 + 98.4 * 10.93 + 95.5 * 11.87 + 91.8 * 12.74 + 87.6 * 13.58}{100 + 98.4 + 95.5 + 91.8 + 87.6} = 11.76\%,$$

while the yield on our assets (the bond we bought) is 11.98%. So we just sit back and take a yield gain of 0.22%." But let us look for a minute at the cash-flows from this arrangement (Note that one serial bond has payments $(30, 28, 26, 24, 22)$ and that we can buy $473.3/95.4 = 4.9612$ serial bonds for

[2]Jakobsen, S. and C. Tanggard: Faldgruber i brugen af effektiv rente og varighed, finans/invest, 2/87.

the money we raise.)

	Time 0	1	2	3	4	5
Liabilities						
1 yr bullet	100	-110	0	0	0	0
2 yr bullet	98.4	-10	-110	0	0	0
3 yr bullet	95.5	-10	-10	-110	0	0
4 yr bullet	91.8	-10	-10	-10	-110	0
5 yr bullet	87.6	-10	-10	-10	-10	-110
Assets						
5 yr serial	-473.3	148.84	138.91	128.99	119.07	109.15
Net position						
	0	-1.26	-1.19	-1.01	-0.93	-0.75

So we see that what have in fact found is a sure-fire way of throwing money away. So what went wrong? The yield on the liability side is not 11.76%. The yield of a portfolio is a non-linear function of all payments of the portfolio, and it is not a simple function (such as a weighted average) of the yields of the individual components of the portfolio. The correct calculation gives that the yield on the liabilities is 12.29%. This suggests that we should perform the exact opposite transactions. And we should, since from the table of cash-flow we see that this is an arbitrage-opportunity ("a free lunch"). But how can we be sure to find such arbitrages? By performing an analysis similar to that in Example 4, i.e. pick out a sufficient number of bonds to construct zero-coupon bonds and check if all other bonds are priced correctly. If not it is easy to see how the arbitrage-opportunities are exploited. If we pick out the 5 bullets and do this, we find that the correct price of the serial is 94.7, which is confirmation that arbitrage-opportunities exists in the market. Note that we do not have to worry if it is the serial that is overpriced or the bullets that are underpriced.

Of course things are not a simple in practice as in this example. Market imperfections (such as bid-ask spreads) and the fact that there are more payments dates the bonds make it a challenging empirical task to estimate the zero-coupon yield curve. Nonetheless the idea of finding the zero-coupon yield curve and using it to find over- and underpriced bonds did work wonders in the Danish bond market in the '80ies (the *19*80'ies, that is).

3.4 IRR, NPV and capital budgeting under certainty.

The definition of *internal rate of return* (IRR) is the same as that of yield, but we use it on arbitrary cash flows, i.e. on securities which may have negative cash flows as well:

Definition 14 *An internal rate of return of a security* $(c_1, \ldots, c_T)$ *with price* $\pi \neq 0$ *is a solution* $y > -1$ *of the equation*

$$\pi = \sum_{i=1}^{T} \frac{c_i}{(1+y)^i}.$$

Hence the definitions of yield and internal rate of return are identical for positive cash flows. It is easy to see that for securities whose future payments are both positive and negative we may have several IRRs. This is one reason that one should be very careful interpreting and using this measure at all when comparing cash flows. We will see below that there are even more serious reasons. When judging whether a certain cash flow is 'attractive' the correct measure to use is Net Present Value:

Definition 15 *The PV and NPV of security* $(c_1, \ldots, c_T)$ *with price* c_0 *given a term structure* $(y(0,1), \ldots, y(0,T))$ *are defined as*

$$\begin{aligned} PV(c) &= \sum_{i=1}^{T} \frac{c_i}{(1+y(0,i))^i} \\ NPV(c) &= \sum_{i=1}^{T} \frac{c_i}{(1+y(0,i))^i} - c_0 \end{aligned}$$

Next, we will see how these concepts are used in deciding how to invest under certainty.

Assume throughout this section that we have a complete security market as defined in the previous section. Hence a unique discount function d is given as well as the associated concepts of interest rates and yields. We let y denote the term structure of interest rates and use the short hand notation y_i for $y(0,i)$.

In capital budgeting we analyze how firms should invest in projects whose payoffs are represented by cash flows. Whereas we assumed in the security market model that a given security could be bought or sold in any quantity

desired, we will use the term *project* more restrictively: We will say that the project is scalable by a factor $\lambda \neq 1$ if it is possible to start a project which produces the cash flow λc by paying λc_0 initially. A project is not scalable unless we state this explicitly and we will not consider any negative scaling.

In a complete financial market an investor who needs to decide on only one project faces a very simple decision: Accept the project if and only if it has positive NPV. We will see why this is shortly. Accepting this fact we will see examples of some other criteria which are generally inconsistent with the NPV criterion. We will also note that when a collection of projects are available capital budgeting becomes a problem of maximizing NPV over the range of available projects. The complexity of the problem arises from the constraints that we impose on the projects. The available projects may be non-scalable or scalable up to a certain point, they may be mutually exclusive (i.e. starting one project excludes starting another), we may impose restrictions on the initial outlay that we will allow the investor to make (representing limited access to borrowing in the financial market), we may assume that a project may be repeated once it is finished and so on. In all cases our objective is simple: Maximize NPV.

First, let us note why looking at NPV is a sensible thing to do:

Proposition 5 *Given a cash flow $c = (c_1, \ldots, c_T)$ and given c_0 such that $NPV(c_0; c) < 0$. Then there exists a portfolio θ of securities whose price is c_0 and whose payoff satisfies*

$$C^\top \theta > \begin{pmatrix} c_1 \\ \vdots \\ c_T \end{pmatrix}.$$

Conversely, if $NPV(c_0; c) > 0$, then every θ with $C^\top \theta = c$ satisfies $\pi^\top \theta > c_0$.

Proof. Since the security market is complete, there exists a portfolio θ^c such that $C^\top \theta^c = c$. Now $\pi^\top \theta^c < c_0$ (why?), hence we may form a new portfolio by investing the amount $c_0 - \pi^\top \theta^c$ in some zero coupon bond (e_1, say) and also invest in θ^c. This generates a stream of payments equal to $C^\top \theta^c + \frac{(c_0 - \pi^\top \theta^c)}{d_1} e_1 > c$ and the cost is c_0 by construction. The second part is left as an exercise! TCIMACRO ■ ■

The interpretation of this lemma is the following: One should never accept a project with negative NPV since a strictly larger cash flow can be obtained at the same initial cost by trading in the capital market. On the other hand, a positive NPV project generates a cash flow at a lower cost than the cost of generating the same cash flow in the capital market. It might seem that

this generates an arbitrage opportunity since we could buy the project and sell the corresponding future cash flow in the capital market generating a profit at time 0. However, we insist on relating the term *arbitrage* to the capital market only. Projects should be thought of as 'endowments': Firms have an available range of projects. By choosing the right projects the firms maximize the value of these 'endowments'.

Some times when performing $NPV-$calculations, we assume that 'the term structure is flat' . What this means is that the discount function has the particularly simple form

$$d_t = \frac{1}{(1+r)^t}$$

for some constant r, which we will usually assume to be non-negative, although our model only guarantees that $r > -1$ in an arbitrage-free market. A flat term structure is very rarely observed in practice - a typical real world term structure will be upward sloping: Yields on long maturity zero coupon bonds will be greater than yields on short bonds. Reasons for this will be discussed once we model the term structure and its evolution over time - a task which requires the introduction of uncertainty to be of any interest. When the term structure is flat then evaluating the NPV of a project having a constant cash flow is easily done by summing the geometric series. The present value of n payments starting at date 1, ending at date n each of size c, is

$$\sum_{i=1}^{n} cd^i = cd\sum_{i=0}^{n-1} d^i = cd\frac{1-d^n}{1-d}, \qquad d \neq 1$$

Another classical formula concerns the present value of a geometrically growing payment stream $(c, c(1+g), \ldots, c(1+g)^{n-1})$ as

$$\begin{aligned} & \sum_{i=1}^{n} c\frac{(1+g)^{i-1}}{(1+r)^i} \\ = \; & \frac{c}{1+r}\sum_{i=0}^{n-1}\frac{(1+g)^i}{(1+r)^i} \\ = \; & \frac{c}{r-g}\left(1-\left(\frac{1+g}{1+r}\right)^n\right). \end{aligned}$$

Although we have not taken into account the possibility of infinite payment streams, we note for future reference, that for $0 \leq g < r$ we have what is known as *Gordon's growth formula:*

$$\sum_{i=1}^{\infty} \frac{c(1+g)^{i-1}}{(1+r)^i} = \frac{c}{r-g}.$$

3.4.1 Some rules which are inconsistent with the NPV rule.

Corresponding to our definition of internal rate of return in Chapter 3, we define an internal rate of return on a project c with initial cost $c_0 > 0$, denoted $IRR(c_0; c)$, as a solution to the equation

$$c_0 = \sum_{i=1}^{T} \frac{c_i}{(1+x)^i}, \qquad x > -1$$

As we have noted earlier such a solution need not be unique unless $c > 0$ and $c_0 > 0$.

Note that an internal rate of return is defined without referring to the underlying term structure. The internal rate of return describes the level of a flat term structure at which the NPV of the project is 0. The idea behind its use in capital budgeting would then be to say that the higher the level of the interest rate, the better the project (and some sort of comparison with the existing term structure would then be appropriate when deciding whether to accept the project at all). But as we will see in the following example, IRR and NPV may disagree on which project is better: Consider the projects shown in the table below (whose last column shows a discount function d):

date	proj 1	proj 2	d
0	-100	-100	1
1	50	50	0.95
2	5	80	0.85
3	90	4	0.75
IRR	0.184	0.197	-
NPV	19.3	18.5	-

Project 2 has a higher IRR than project 1, but 1 has a larger NPV than 2. Using the same argument as in the previous section it is easy to check, that even if a cash flow similar to that of project 2 is desired by an investor, he would be better off investing in project 1 and then reforming the flow of payments using the capital market.

Another problem with trying to use IRR as a decision variable arises when the IRR is not uniquely defined - something which typically happens when the cash flows exhibit sign changes. Which IRR should we then choose?

One might also contemplate using *the payback method* and count the number of years it takes to recover the initial cash outlay - possibly after discounting appropriately the future cash flows. Project 2 in the table has a payback of 2 years whereas project 1 has a payback of three years. The example above therefore also shows that choosing projects with the shortest payback time may be inconsistent with the NPV method.

3.4.2 Several projects.

Consider someone with $c_0 > 0$ available at date 0 who wishes to allocate this capital over the $T+1$ dates, and who considers a project c with initial cost c_0. We have seen that precisely when $NPV(c_0; c) > 0$ this person will be able to obtain better cash flows by adopting c and trading in the capital market than by trading in the capital market alone.

When there are several projects available the situation really does not change much: Think of the $i'th$ project (p_0^i, p) as an element of a set $P_i \subset \mathbb{R}^{T+1}$. Assume that $0 \in P_i$ all i representing the choice of not starting the i'th project. For a non-scalable project this set will consist of one point in addition to 0.

Given a collection of projects represented by $(P_i)_{i \in I}$. Situations where there is a limited amount of money to invest at the beginning (and borrowing is not permitted), where projects are mutually exclusive etc. may then be described abstractly by the requirement that the collection of selected projects $(p_0^i, p^i)_{i \in I}$ are chosen from a feasible subset P of the Cartesian product $\times_{i \in I} P_i$. The NPV of the chosen collection of projects is then just the sum of the NPVs of the individual projects and this in turn may be written as the NPV of the sum of the projects:

$$\sum_{i \in I} NPV(p_0^i; p^i) = NPV\left(\sum_{i \in I} (p_0^i, p^i)\right).$$

Hence we may think of the chosen collection of projects as producing one project and we can use the result of the previous section to note that clearly an investor should choose a project giving the highest NPV. Rather than elaborating on this point, we consider an example.

Example 7 *Consider the following example from Copeland and Weston (1988):*

project	NPV	initial cost
1	30.000	200.000
2	16.250	125.000
3	19.250	175.000
4	12.000	150.000

Assume that all projects are non-scalable, and assume that we can only invest up to an amount of 300.000. This capital constraint forces us to choose, i.e. projects become mutually exclusive to some extent. Clearly, with no constraints all projects would be adopted since the NPVs are positive in all cases. Note that project 1 generates the largest NPV but it also uses a large portion of the budget: If we adopt 1, there is no room for additional projects. The only way to deal with this problem is to stick to the NPV-rule and go through the set of feasible combinations of projects and compute the NPV. It is not hard to see that combining projects 2 and 3 produces the maximal NPV given the capital constraint.

If the projects were assumed scalable, the situation would be different: Then project 1 adopted at a scale of 1.5 would clearly be optimal. This is simply because the amount of NPV generated per dollar invested is larger for project 1 than for the other projects. Exercises will illustrate other examples of NPV-maximization.

The moral of this section is simple: Given a perfect capital market, investors who are offered projects should simply maximize NPV. This is merely an equivalent way of saying that profit maximization with respect to the existing price system (as represented by the term structure) is the appropriate strategy when a perfect capital market exists. The technical difficulties arise from the constraints that we impose on the projects and these constraints easily lead to linear programming problems, integer programming problems or even non-linear optimization problems.

However, real world projects typically do not generate cash flows which are known in advance. Real world projects involve risk and uncertainty and therefore capital budgeting under certainty is really not sophisticated enough for a manager deciding which projects to undertake. A key objective of this course is to try and model uncertainty and to construct models of how risky cash flows are priced. This will give us definitions of NPV which work for uncertain cash flows as well.

3.5 Duration, convexity and immunization.

3.5.1 Duration with a flat term structure.

In this chapter we introduce the notions of duration and convexity which are often used in practical bond risk management and asset/liability management. It is worth stressing that when we introduce dynamic models of the term structure of interest rates in a world with uncertainty, we obtain much more sophisticated methods for measuring and controlling interest rate risk than the ones presented in this section.

Consider a financial market which is arbitrage-free and complete and where the discount function $d = (d_1, \dots d_T)$ satisfies

$$d_i = \frac{1}{(1+r)^i} \text{ for } i = 1, \dots, T.$$

This corresponds to the assumption of a flat term structure. We stress that this assumption is rarely satisfied in practice but we will see how to relax this assumption.

What we are about to investigate are changes in present values as a function of changes in r. This makes perfect sense even in a world of certainty, but sometimes we will speak freely of 'interest changes' occurring even though strictly speaking, we still do not have uncertainty in our model.

With a flat term structure, the present value of a payment stream $c = (c_1, \dots, c_T)$ is given by

$$PV(c; r) = \sum_{t=1}^{T} \frac{c_t}{(1+r)^t}$$

We have now included the dependence on r explicitly in our notation since what we are about to model are essentially derivatives of $PV(c; r)$ with respect to r.

Let c be a non-negative payment stream.

Definition 16 *The duration $D(c; r)$ of c is given by*

$$\begin{aligned} D(c; r) &= \left(-\frac{\partial}{\partial r} PV(c; r) \right) \frac{1+r}{PV(c; r)} \qquad (3.3) \\ &= \frac{1}{PV(c; r)} \sum_{t=1}^{T} t \frac{c_t}{(1+r)^t} \end{aligned}$$

This duration is called the Macaulay duration and is the "classical" one (many more advanced durations have been proposed in the literature). Rather that saying it is based on a flat term structure, we could refer to it as being based on the yield of the bond (or portfolio). Note that defining

$$w_t = \frac{c_t}{(1+r)^t}\frac{1}{PV(c;r)} \tag{3.4}$$

we have $\sum_{t=1}^{T} w_t = 1$, hence

$$D(c;r) = \sum_{t=1}^{T} t\, w_t.$$

Definition 17 *The convexity of c is given by*

$$K(c;r) = \sum_{t=1}^{T} t^2\, w_t. \tag{3.5}$$

where w_t is given by (3.4).

Let us try to interpret D and K by computing the first and second derivatives[3] of $PV(c;r)$ with respect to r.

$$\begin{aligned}
PV'(c;r) &= -\sum_{t=1}^{T} t\, c_t \frac{1}{(1+r)^{t+1}} \\
&= -\frac{1}{1+r}\sum_{t=1}^{T} t\, c_t \frac{1}{(1+r)^{t}} \\
PV''(c;r) &= \sum_{t=1}^{T} t\,(t+1)\,\frac{c_t}{(1+r)^{t+2}} \\
&= \frac{1}{(1+r)^2}\left[\sum_{t=1}^{T} t^2 c_t \frac{1}{(1+r)^t} + \sum_{t=1}^{T} t c_t \frac{1}{(1+r)^t}\right]
\end{aligned}$$

Now consider the relative change in $PV(c;r)$ when r changes to $r+\Delta r$, i.e.

$$\frac{PV(c;r+\Delta r) - PV(c;r)}{PV(c;r)}$$

[3]From now on we write $PV'(c;r)$ and $PV''(c;r)$ instead of $\frac{\partial}{\partial r}PV(c;r)$ resp. $\frac{\partial^2}{\partial r^2}PV(c;r)$

By considering a second order Taylor expansion of the numerator, we obtain

$$\begin{aligned}\frac{PV(c;r+\Delta r)-PV(c;r)}{PV(c;r)} &\approx \frac{PV'(c;r)\Delta r+\frac{1}{2}PV''(c;r)(\Delta r)^2}{PV(c;r)}\\ &= -D\frac{\Delta r}{(1+r)}+\frac{1}{2}(K+D)\left(\frac{\Delta r}{1+r}\right)^2\end{aligned}$$

Hence D and K can be used to approximate the relative change in $PV(c;r)$ as a function of the relative change in r (or more precisely, relative changes in $1+r$, since $\frac{\Delta(1+r)}{1+r}=\frac{\Delta r}{1+r}$).

Sometimes one finds the expression *modified duration* defined by

$$MD(c;r)=\frac{D}{1+r}$$

and using this in a first order approximation, we get the relative change in $PV(c;r)$ expressed by $-MD(c;r)\Delta r$, which is a function of Δr itself. The interpretation of D as a price elasticity gives us no reasonable explanation of the word 'duration', which certainly leads one to think of quantity measured in units of time. If we use the definition of w_t we have the following simple expression for the duration:

$$D(c;r)=\sum_{t=1}^{T} t\,w_t.$$

Notice that w_t expresses the present value of c_t divided by the total present value, i.e. w_t expresses the weight by which c_t is contributing to the total present value. Since $\sum_{t=1}^{T} w_t=1$ we see that $D(c;r)$ may be interpreted as a 'mean waiting time'. The payment which occurs at time t is weighted by w_t.

Example 8 For the government bullet bond in Example 5 the present value of the payment stream is 106.75 and therefore the duration is

$$\frac{\sum_{k=1}^{4} t_k c_k(1+y)^{-t_k}}{PV}=\frac{464.06}{106.75}=4.35$$

while the convexity is

$$\frac{\sum_{k=1}^{4} t_k^2 c_k(1+y)^{-t_k}}{PV}=\frac{2172.753}{106.75}=20.35,$$

and the following table shows the the exact and approximated relative chances in present value when the yield changes:

Yield	△yield	Exact rel. (%) PV-change	First order approximation	Second order approximation
0.04	-0.0150	6.434	6.181	6.430
0.05	-0.0050	2.084	2.060	2.088
0.055	0	0	0	0
0.060	0.0050	-2.036	-2.060	-2.032
0.070	0.0150	-5.941	-6.180	-5.930

Notice that since PV is a decreasing, convex function of y we know that the first order approximation will underestimate the effect of decreasing y (and overestimate the effect of increasing it).

Notice that for a zero coupon bond with time to maturity t the duration is t. For other kinds of bonds with time to maturity t, the duration is less than t. Furthermore, note that investing in a zero coupon bond with yield to maturity r and holding the bond to expiration guarantees the owner an annual return of r between time 0 and time t. This is not true of a bond with maturity t which pays coupons before t. For such a bond the duration has an interpretation as the length of time for which the bond can ensure an annual return of r :

Let $FV(c; r, H)$ denote the (future) value of the payment stream c at time H if the interest rate is fixed at level r. Then

$$\begin{aligned} FV(c; r, H) &= (1+r)^H PV(c; r) \\ &= \sum_{t=1}^{H-1} c_t(1+r)^{H-t} + c_H + \sum_{t=H+1}^{T} c_t \frac{1}{(1+r)^{t-H}} \end{aligned}$$

Consider a change in r which occurs an instant after time 0. How would such a change affect $FV(c; r, H)$? There are two effects with opposite directions which influence the future value: Assume that r decreases. Then the first sum in the expression for $FV(c; r, H)$ will decrease. This decrease can be seen as caused by *reinvestment risk:* The coupons received up to time H will have to be reinvested at a lower level of interest rates. The last sum will increase when r decreases. This is due to *price risk* : As interest rates fall the value of the remaining payments after H will be higher since they have to be discounted by a smaller factor. Only c_H is unchanged.

The natural question to ask then is for which H these two effects cancel each other. At such a time point we must have $\frac{\partial}{\partial r} FV(c; r, H) = 0$ since an infinitesimal change in r should have no effect on the future value. Now,

$$\begin{aligned}\frac{\partial}{\partial r}FV(c;r,H) &= \frac{\partial}{\partial r}\left[(1+r)^H PV(c;r)\right] \\ &= H(1+r)^{H-1}PV(c;r) + (1+r)^H PV'(c;r)\end{aligned}$$

Setting this expression equal to 0 gives us

$$\begin{aligned}H &= \frac{-PV'(c;r)}{PV(c;r)}(1+r) \\ \text{i.e. } H &= D(c;r)\end{aligned}$$

Furthermore, at $H = D(c;r)$, we have $\frac{\partial^2}{\partial r^2}FV(c;r,H) > 0$. This you can check by computing$\frac{\partial^2}{\partial r^2}\left((1+r)^H PV(c;r)\right)$,reexpressing in terms D and K, and using the fact that $K > D^2$. Hence, at $H = D(c;r)$, $FV(c;r,H)$ will have a minimum in r. We say that $FV(c;r,H)$ is *immunized* towards changes in r, but we have to interpret this expression with caution: The only way a bond really can be immunized towards changes in the interest rate r between time 0 and the investment horizon t is by buying zero coupon bonds with maturity t. Whenever we buy a coupon bond at time 0 with duration t, then to a first order approximation, an interest change immediately after time 0, will leave the future value at time t unchanged. However, as date 1 is reached (say) it will *not* be the case that the duration of the coupon bond has decreased to $t-1$. As time passes, it is generally necessary to adjust bond portfolios to maintain a fixed investment horizon, even if r is unchanged. This is true even in the case of certainty.

Later when we introduce dynamic hedging strategies we will see how a portfolio of bonds can be dynamically managed so as to truly immunize the return.

3.5.2 Relaxing the assumption of a flat term structure.

What we have considered above were parallel changes in a flat term structure. Since we rarely observe this in practice, it is natural to try and generalize the analysis to different shapes of the term structure. Consider a family of structures given by a function r of two variables, t and x. Holding x fixed gives a term structure $r(\cdot, x)$.

For example, given a current term structure $(y_1, \ldots, y_T)$ we could have $r(t,x) = y_t + x$ in which case changes in x correspond to additive changes in the current term structure (the one corresponding to $x = 0$). Or we could have $1 + r(t,x) = (1+y_t)x$, in which case changes in x would produce multiplicative changes in the current (obtained by letting $x = 1$) term structure.

Now let us compute changes in present values as x changes:

$$\frac{\partial PV}{\partial x} = -\sum_{t=1}^{T} t c_t \frac{1}{(1+r(t,x))^{t+1}} \frac{\partial r(t,x)}{\partial x}$$

which gives us

$$\frac{\partial PV}{\partial x}\frac{1}{PV} = -\sum_{t=1}^{T} \frac{t w_t}{1+r(t,x)} \frac{\partial r(t,x)}{\partial x}$$

where

$$w_t = \frac{c_t}{(1+r(t,x))^t}\frac{1}{PV}$$

We want to try and generalize the 'investment horizon' interpretation of duration, and hence calculate the future value of the payment stream at time H and differentiate with respect to x. Assume that the current term structure is $r(\cdot, x_0)$.

$$FV(c; r(H,x_0), H) = (1+r(H,x_0))^H PV(c; r(t,x_0))$$

Differentiating

$$\begin{aligned}\frac{\partial}{\partial x} FV(c; r(H,x), H) &= (1+r(H,x))^H \frac{\partial PV}{\partial x} \\ &\quad + H(1+r(H,x))^{H-1} \frac{\partial r(H,x)}{\partial x} PV(c; r(t,x))\end{aligned}$$

Evaluate this derivative at $x = x_0$ and set it equal to 0:

$$\left.\frac{\partial PV}{\partial x}\right|_{x=x_0} \frac{1}{PV} = -H \left.\frac{\partial r(H,x)}{\partial x}\right|_{x=x_0} (1+r(H,x_0))^{-1}$$

and hence we could define the duration corresponding to the given parametrization as the value D for which

$$\left.\frac{\partial PV}{\partial x}\right|_{x=x_0} \frac{1}{PV} = -D \left.\frac{\partial r(D,x)}{\partial x}\right|_{x=x_0} (1+r(D,x_0))^{-1}.$$

The additive case would correspond to

$$\left.\frac{\partial r(D,x)}{\partial x}\right|_{x=0} = 1,$$

and the multiplicative case to

$$\left.\frac{\partial r(D,x)}{\partial x}\right|_{x=1} = 1 + y_D.$$

Note that the multiplicative case gives us the duration measure (called the Fisher-Weil duration)

$$D_{mult} = -\frac{\partial PV}{\partial x}\frac{1}{PV} = \sum_{t=1}^{T} t w_t$$

which is just like the original measure although the weights of course reflect the structure $y(0,t)$.

Example 9 Consider again the small bond market from Example 4. We have already found the zero-coupon yields in the market, and find that the Fisher-Weil duration of the 4 yr serial bond is

$$\frac{1}{102.38}\left(\frac{32}{1.0500} + \frac{2*30.25}{1.0550^2} + \frac{3*28.5}{1.0600^3} + \frac{4*26.75}{1.0650^4}\right) = 2.342,$$

and the following table gives the yields, Macaulay durations based on yields and Fisher-Weil durations for all the coupon bonds:

Bond	Yield ( )	M-duration	FW-duration
1 yr bullet	5	1	1
2 yr bullet	5.49	1.952	1.952
3 yr annuity	5.65	1.963	1.958
4 yr serial	5.93	2.354	2.342

3.5.3 An example

We finish this chapter with an example (with something usually referred to as a barbell strategy) which is intended to cause some concern. Some of the claims are for you to check!

A financial institution issues 100 million $ worth of 10 year bullet bonds with time to maturity 10 years and a coupon rate of 7 percent. Assume that the term structure is flat at $r = 7$ percent. The revenue (of 100 million $) is used to purchase 10-and 20 year annuities also with coupon rates of 7%. The numbers of the 10 and 20 year annuities purchased are chosen in such a way that the duration of the issued bullet bond matches that of the portfolio of annuities. Now there are three facts you need to know at this stage. Letting

T denote time to maturity, r the level of the term structure and γ the coupon rate, we have that the duration of an annuity is given by

$$D_{ann} = \frac{1+r}{r} - \frac{T}{(1+r)^T - 1}.$$

Note that since payments on an annuity are equal in all periods we need not know the size of the payments to calculate the duration.

The duration of a bullet bond is

$$D_{bullet} = \frac{1+r}{r} - \frac{1+r-T(r-R)}{R\left((1+r)^T - 1\right) + r}$$

which of course simplifies when $r = R$.

The third fact you need to check is that if a portfolio consists of two securities whose values are P_1 and P_2 respectively, then the duration of the portfolio $P_1 + P_2$ is given as

$$D(P_1 + P_2\ ) = \frac{P_1}{P_1 + P_2} D(P_1) + \frac{P_2}{P_1 + P_2} D(P_2).$$

Using these three facts you will note that a portfolio consisting of 23.77 million dollars worth of the 10-year annuity and 76.23 million dollars worth of the 20-year annuity will produce a portfolio whose duration exactly matches that of the issued bullet bond. By construction the present value of the two annuities equals that of the bullet bond. The present value of the whole transaction in other words is 0 at an interest level of 7 percent. However, for all other levels of the interest rate, the present value is strictly positive! In other words, any change away from 7 percent will produce a profit to the financial institution. We will have more to say about this phenomenon in the exercises and we will return to it when discussing the term structure of interest rates in models with uncertainty. As you will see then, the reason that we can construct the example above is that we have set up an economy in which there are arbitrage opportunities.

Chapter 4

Arbitrage pricing in a one-period model

One of the biggest success stories of financial economics is the *Black-Scholes model of option pricing*. But even though the formula itself is easy to use, a rigorous presentation of how it comes about requires some fairly sophisticated mathematics. Fortunately, the so-called binomial model of option pricing offers a much simpler framework and gives almost the same level of understanding of the way option pricing works. Furthermore, the binomial model turns out to be very easy to generalize (to so-called multinomial models) and more importantly to use for pricing other derivative securities (i.e. different contract types or different underlying securities) where an extension of the Black-Scholes framework would often turn out to be difficult. The flexibility of binomial models is the main reason why these models are used daily in trading all over the world.

Binomial models are often presented separately for each application. For example, one often sees the "classical" binomial model for pricing options on stocks presented separately from binomial term structure models and pricing of bond options etc.

The aim of this chapter is to present the underlying theory at a level of abstraction which is high enough to understand all binomial/multinomial approaches to the pricing of derivative securities as special cases of one model.

Apart from the obvious savings in allocation of brain RAM that this provides, it is also the goal to provide the reader with a language and framework which will make the transition to continuous-time models in future courses much easier.

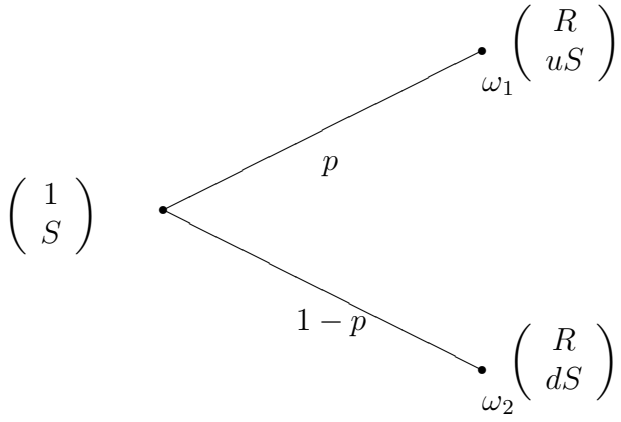
4.1 An appetizer.

Before we introduce our model of a financial market with uncertainty formally, we present a little appetizer. Despite its simplicity it contains most of the insights that we are about to get in this chapter.

Consider a one-period model with two states of nature, ω_1 and ω_2. At time $t = 0$ nothing is known about the time state, at time $t = 1$ the state is revealed. State ω_1 occurs with probability p. Two securities are traded:

- A *stock* which costs S at time 0 and is worth uS at time 1 in one state and dS in the other.
- A *money market account* which costs 1 at time 0 and is worth R at time 1 regardless of the state.

Assume $0 < d < R < u$. (This condition will be explained later.) We summarize the setup with a graph:



Now assume that we introduce into the economy a *European call option on the stock with exercise price K and maturity 1*. At time 1 the value of this call is equal to (where the notation $[y]^+$ (or sometimes $(y)^+$) means $\max(y, 0)$)

$$C_1(\omega) = \begin{cases} [uS - K]^+ & \text{if } \omega = \omega_1 \\ [dS - K]^+ & \text{if } \omega = \omega_2 \end{cases}$$

We will discuss options in more detail later. For now, note that it can be thought of as a contract giving the owner the right but not the obligation to buy the stock at time 1 for K.

To simplify notation, let $C_u = C_1(\omega_1)$ and $C_d = C_1(\omega_2)$. The question is: What should the price of this call option be at time 0? A simple portfolio argument will give the answer: Let us try to form a portfolio at time 0 using only the stock and the money market account which gives the same payoff as the call at time 1 regardless of which state occurs. Let (a, b) denote, respectively, the number of stocks and units of the money market account held at time 0. If the payoff at time 1 has to match that of the call, we must have

$$a(uS) + bR = C_u$$
$$a(dS) + bR = C_d$$

Subtracting the second equation from the first we get

$$a(u-d)S = C_u - C_d$$

i.e.

$$a = \frac{C_u - C_d}{S(u-d)}$$

and algebra gives us

$$b = \frac{1}{R}\frac{uC_d - dC_u}{(u-d)}$$

where we have used our assumption that $u > d$. The cost of forming the portfolio (a, b) at time 0 is

$$\frac{(C_u - C_d)}{S\,(u-d)}S + \frac{1}{R}\frac{uC_d - dC_u}{(u-d)} \cdot 1$$

$$= \frac{R\,(C_u - C_d)}{R\,(u-d)} + \frac{1}{R}\frac{uC_d - dC_u}{(u-d)}$$

$$= \frac{1}{R}\left[\frac{R-d}{u-d}C_u + \frac{u-R}{u-d}C_d\right].$$

We will formulate below exactly how to define the notion of no arbitrage when there is uncertainty, but it should be clear that the argument we have just given shows why the call option must have the price

$$C_0 = \frac{1}{R}\left[\frac{R-d}{u-d}C_u + \frac{u-R}{u-d}C_d\right]$$

Rewriting this expression we get

$$C_0 = \left(\frac{R-d}{u-d}\right)\frac{C_u}{R} + \left(\frac{u-R}{u-d}\right)\frac{C_d}{R}$$

and if we let

$$q = \frac{R-d}{u-d}$$

we get

$$C_0 = q\frac{C_u}{R} + (1-q)\frac{C_d}{R}.$$

If the price were lower, one could buy the call and sell the portfolio (a, b), receive cash now as a consequence and have no future obligations except to exercise the call if necessary.

Some interesting features of this example will be much clearer as we go along:

- The probability p plays no role in the expression for C_0.
- A new set of probabilities

  $$q = \frac{R-d}{u-d} \qquad \text{and} \qquad 1-q = \frac{u-R}{u-d}$$

  emerges (this time we also use that $d < R < u$) and with this set of probabilities we may write the value of the call as

  $$C_0 = E^q\left[\frac{C_1(\omega)}{R}\right]$$

  i.e. an expected value *using* q of the discounted time 1 value of the call.
- If we compute the expected value *using* q of the discounted time 1 stock price we find

  $$E^q\left[\frac{S(\omega)}{R}\right] = \left(\frac{R-d}{u-d}\right)\frac{1}{R}(uS) + \left(\frac{u-R}{u-d}\right)\frac{1}{R}(dS) = S$$

The method of pricing the call really did not use the fact that C_u and C_d were call-values. Any security with a time 1 value depending on ω_1 and ω_2 could have been priced.

4.2 The single period model

The mathematics used when considering a one-period financial market with uncertainty is exactly the same as that used to describe the bond market in a multiperiod model with certainty: Just replace dates by states.

Given two time points $t = 0$ and $t = 1$ and a finite state space

$$\Omega = \{\omega_1, \ldots, \omega_S\}.$$

Whenever we have a probability measure P (or Q) we write p_i (or q_i) instead of $P(\{\omega_i\})$ (or $Q(\{\omega_i\})$).

A *security price system* is a vector $\pi \in \mathbb{R}^N$ and an $N \times S$ matrix D where we interpret the i'th row $(d_{i1}, \ldots, d_{iS})$ of D as the payoff at time 1 of the i'th security in states $1, \ldots, S$, respectively. The price at *time* 0 of the i'th security is π_i. A *portfolio* is a vector $\theta \in \mathbb{R}^N$ whose coordinates represent the number of securities bought at time 0. *The price of the portfolio* θ bought at time 0 is $\pi \cdot \theta$.

Definition 18 *An arbitrage in the security price system* (π, D) *is a portfolio* θ *which satisfies either*

$$\pi \cdot \theta \leq 0 \in \mathbb{R} \text{ and } D^\top \theta > 0 \in \mathbb{R}^S$$

or

$$\pi \cdot \theta < 0 \in \mathbb{R} \text{ and } D^\top \theta \geq 0 \in \mathbb{R}^S$$

A security price system (π, D) *for which no arbitrage exists is called arbitrage-free.*

Remark 1 *Our conventions when using inequalities on a vector in* $\mathbb{R}^k$ *are the same as described in Chapter 3.*

When a market is arbitrage-free we want a vector of prices of 'elementary securities' - just as we had a vector of discount factors in Chapter 3.

Definition 19 $\psi \in \mathbb{R}^S_{++}$ *(i.e.* $\psi \gg 0$*) is said to be a state-price vector for the system* (π, D) *if it satisfies*

$$\pi = D\psi$$

Clearly, we have already proved the following in Chapter 3:

Proposition 6 *A security price system is arbitrage-free if and only if there exists a state-price vector.*

Unlike the model we considered in Chapter 3, the security which pays 1 in every state plays a special role here. If it exists, it allows us to speak of an 'interest rate':

Definition 20 *The system (π, D) contains a riskless asset if there exists a linear combination of the rows of D which gives us $(1, \ldots, 1) \in \mathbb{R}^S$.*

In an arbitrage-free system the price of the riskless asset d_0 is called the *discount factor* and $R_0 \equiv \frac{1}{d_0}$ is the *return on the riskless asset*. Note that when a riskless asset exists, and the price of obtaining it is d_0, we have

$$d_0 = \theta_0^\top \pi = \theta_0^\top D\psi = \psi_1 + \cdots + \psi_S$$

where θ_0 is the portfolio that constructs the riskless asset.

Now define

$$q_i = \frac{\psi_i}{d_0}, i = 1, \ldots, S$$

Clearly, $q_i > 0$ and $\sum_{i=1}^S q_i = 1$, so we may interpret the q_i's as probabilities. We may now rewrite the identity (assuming no arbitrage) $\pi = D\psi$ as follows:

$$\pi = d_0 Dq = \frac{1}{R_0} Dq, \text{ where } q = (q_1, \ldots, q_S)^\top$$

If we read this coordinate by coordinate it says that

$$\pi_i = \frac{1}{R_0} (q_1 d_{i1} + \ldots + q_S d_{iS})$$

which is the discounted expected value using q of the i'th security's payoff at time 1. Note that since R_0 is a constant we may as well say "expected discounted ...".

We assume throughout the rest of this section that a riskless asset exists.

Definition 21 *A security $c = (c_1, \ldots, c_S)$ is redundant given the security price system (π, D) if there exists a portfolio θ_c such that $D^t \theta_c = c$.*

Proposition 7 *Let an arbitrage-free system (π, D) and a redundant security c by given. The augmented system $\left(\widehat{\pi}, \widehat{D}\right)$ obtained by adding π_c to the vector π and $c \in \mathbb{R}^S$ as a row of D is arbitrage-free if and only if*

$$\pi_c = \frac{1}{R_0} (q_1 c_1 + \ldots + q_S c_S) \equiv \psi_1 c_1 + \ldots + \psi_S c_S.$$

Proof. Assume $\pi_c < \psi_1 c_1 + \ldots + \psi_S c_S$. Buy the security c and sell the portfolio θ_c. The price of θ_c is by assumption higher than π_c, so we receive a positive cash-flow now. The cash-flow at time 1 is 0. Hence there is an arbitrage opportunity. If $\pi_c > \psi_1 c_1 + \ldots + \psi_S c_S$ reverse the strategy. ■

Definition 22 *The market is complete if for every* $y \in \mathbb{R}^S$ *there exists a* $\theta \in \mathbb{R}^N$ *such that*

$$D^\top \theta = y$$

i.e. if the rows of D *(the columns of* $D^\top$*) span* $\mathbb{R}^S$*.*

Proposition 8 *If the market is complete and arbitrage-free, there exists precisely one state-price vector* ψ*.*

The proof is exactly as in Chapter 3 and we are ready to do contingent claims pricing! Here is how it is done in a one-period model: Construct a set of securities (the D-matrix,) and a set of prices. Make sure that (π, D) is arbitrage-free. Make sure that either

(a) the model is complete, i.e. there are as many linearly independent securities as there are states

or

(b) the contingent claim we wish to price is redundant given (π, D).

Clearly, (a) implies (b) but not vice versa. (a) is almost always what is done in practice. Given a "contingent claim" $c = (c_1, \ldots, c_S)$. Now compute the price of the contingent claim as

$$\pi(c) = \frac{1}{R_0} E^q(c) \equiv \frac{1}{R_0} \sum_{i=1}^{S} q_i c_i$$

where $q_i = \frac{\psi_i}{d_0} \equiv R_0 \psi_i$. The portfolio generating the claim is the solution to $D^\top \theta_c = c$, and since we can always in a complete model reduce the matrix to an $S \times S$ invertible matrix without changing the model this can be done by matrix inversion.

Let us return to our example in the beginning of this chapter: The security price system is

$$\left\{ \begin{pmatrix} 1 \\ S \end{pmatrix}, \begin{pmatrix} R & R \\ uS & dS \end{pmatrix} \right\}.$$

For this to be arbitrage-free, proposition (6) tells us that there must be a solution (ψ_1, ψ_2) with $\psi_1 > 0$ and $\psi_2 > 0$ to the equation

$$\begin{pmatrix} 1 \\ S \end{pmatrix} = \begin{pmatrix} R & R \\ uS & dS \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}.$$

$u \neq d$ ensures that the matrix $\begin{pmatrix} R & R \\ uS & dS \end{pmatrix}$ has full rank. $u > d$ can be assumed without loss of generality. We find

$$\psi_1 = \frac{R - d}{R(u - d)}$$

$$\psi_2 = \frac{u - R}{R\,(u - d)}$$

and note that the solution is strictly positive precisely when $u > R > d$ (given our assumption that $u > d > 0$).

Clearly the riskfree asset has a return of R, and

$$q_1 = R\psi_1 = \frac{R - d}{u - d}$$

$$q_2 = R\psi_2 = \frac{u - R}{u - d}$$

are the probabilities defining the measure q which can be used for pricing. Note that the market is complete, and this explains why we could use the procedure in the previous example to say what the correct price at time 0 of any claim (c_1, c_2) should be.

4.3 The economic intuition

At first, it may seem surprising that the 'objective' probability p does not enter into the expression for the option price. Even if the the probability is 0.99 making the probability of the option paying out a positive amount very large, it does not alter the option's price at time 0. Looking at this problem from a mathematical viewpoint, one can just say that this is a consequence of the linear algebra of the problem: The cost of forming a replicating strategy does not depend on the probability measure and therefore it does not enter into the contract. But this argument will not (and should not) convince a person who is worried by the economic interpretation of a model. Addressing the problem from a purely mathematical angle leaves some very important economic intuition behind. We will try in this section to get the economic intuition behind this 'invariance' to the choice of p straight. This will provide an opportunity to outline how the financial markets studied in this course fit in with a broader microeconomic analysis.

Before the more formal approach, here is the story in words: If we argue (erroneously) that changing p ought to change the option's price at time 0, the same argument should also lead to a suggested change in S_0. But the experiment involving a change in p is an experiment which holds S_0 *fixed*. The given price of the stock is supposed to represent a 'sensible' model of the market. If we change p without changing S_0 we are implicitly changing our description of the underlying economy. An economy in which the probability of an up jump p is increased to 0.99 while the initial stock price remains fixed must be a description of an economy in which payoff in the upstate has

lost value relative to a payoff in the downstate. These two opposite effects precisely offset each other when pricing the option.

The economic model we have in mind when studying the financial market is one in which utility is a function of wealth in each state and wealth is measured by a scalar (kroner, dollars, ...). Think of the financial market as a way of transferring money between different time periods and different states. A real economy would have a (spot) market for real goods also (food, houses, TV-sets, ...) and perhaps agents would have known endowments of real goods in each state at each time. If the spot prices of real goods which are realized in each state at each future point in time are known at time 0, then we may as well express the initial endowment in terms of wealth in each state. Similarly, the optimal consumption plan is associated with a precise transfer of wealth between states which allows one to realize the consumption plan. So even if utility is typically a function of the real goods (most people like money because of the things it allows them to buy), we can formulate the utility as a function of the wealth available in each state.

Consider[1] an agent who has an endowment $e = (e_1, \ldots, e_S) \in \mathbb{R}^S_+$. This vector represents the random wealth that the agent will have at time 1. The agent has a utility function $U : \mathbb{R}^S_+ \to \mathbb{R}$ which we assume to be concave, differentiable and strictly increasing in each coordinate. Given a financial market represented by the pair (π, D), the agent's problem is

$$\begin{array}{lll} \max_\theta U(e & + & D^\top\theta) \\ \text{s.t. } \pi^\top\theta & \leq & 0. \end{array} \tag{4.1}$$

If we assume that there exists a security with a non-negative payoff which is strictly positive in at least one state, then because the utility function is increasing we can replace the inequality in the constraint by an equality. And then the interpretation is simply that the agent sells endowment in some states to obtain more in other states. But no cash changes hands at time 0. Note that while utility is defined over all (non-negative) consumption vectors, it is the rank of D which decides in which directions the consumer can move away from the initial endowment.

Now make sure that you can prove the following

Proposition 9 *If there exists a portfolio θ with $D^\top\theta > 0$ then the agent can find a solution to the maximization problem if and only if (π, D) is arbitrage-free.*

[1]This closely follows Darrell Duffie: Dynamic Asset Pricing Theory. Princeton University Press. 1996

The 'only if' part of this statement shows how no arbitrage is a necessary condition for existence of a solution to the agent's problem and hence for the existence of equilibrium for economies where agents have increasing utility (no continuity assumptions are needed here). The 'if' part uses continuity and compactness (why?) to ensure existence of a maximum, but of course to discuss equilibrium would require more agents and then we need some more of our general equilibrium apparatus to prove existence.

The important insight is the following (see Proposition 1C in Duffie (1996)):

Proposition 10 *Assume that there exists a portfolio θ with $D^\top\theta > 0$. If there exists a solution θ^* to (4.1) and the associated optimal consumption is given by $c^* := e + D^\top\theta^* >> 0$, then the gradient $\nabla U(c^*)$ (thought of as a column vector) is proportional to a state-price vector. The constant of proportionality is positive.*

Proof. Since c^* is strictly positive, then for any portfolio θ there exists some $k(\theta)$ such that $c^* + \alpha D^\top\theta \geq 0$ for all α in $[-k(\theta), k(\theta)]$. Define

$$g_\theta : [-k(\theta), k(\theta)] \to \mathbb{R}$$

as

$$g_\theta(\alpha) = U(c^* + \alpha D^\top\theta)$$

Now consider a θ with $\pi^\top\theta = 0$. Since c^* is optimal, g_θ must be maximized at $\alpha = 0$ and due to our differentiability assumptions we must have

$$g'_\theta(0) = (\nabla U(c^*))^\top D^\top\theta = 0.$$

We can conclude that any θ with $\pi^\top\theta = 0$ satisfies $(\nabla U(c^*))^\top D^\top\theta = 0$. Transposing the last expression, we may also write $\theta^\top D\nabla U(c^*) = 0$. In words, *any* vector which is orthogonal to π is also orthogonal to $D\nabla U(c^*)$. This means that $\mu\pi = D\nabla U(c^*)$ for some μ showing that $\nabla U(c^*)$ is proportional to a state-price vector. Choosing a θ^+ with $D^\top\theta^+ > 0$ we know from no arbitrage that $\pi^\top\theta^+ > 0$ and from the assumption that the utility function is strictly increasing, we have $\nabla U(c^*)D^\top\theta^+ > 0$. Hence μ must be positive. ■

To understand the implications of this result we turn to the special case where the utility function has an expected utility representation, i.e. where we have a set of probabilities $(p_1, \ldots, p_S)$ and a function u such that

$$U(c) = \sum_{i=1}^{S} p_i u(c_i).$$

In this special case we note that the coordinates of the state-price vector satisfy

$$\psi_i = \lambda p_i u'(c_i^*), \quad i = 1, \ldots, S. \tag{4.2}$$

where λ is some constant of proportionality. Now we can state the economic intuition behind the option example as follows (and it is best to think of a complete market to avoid ambiguities in the interpretation): Given the complete market (π, D) we can find a unique state price vector ψ. This state price vector does not depend on p. Thus if we change p and we are thinking of some agent out there 'justifying' our assumptions on prices of traded securities, it must be the case that the agent has different marginal utilities associated with optimal consumption in each state. The difference must offset the change in p in such a way that (4.2) still holds. We can think of this change in marginal utility as happening in two ways: One way is to change utility functions altogether. Then starting with the same endowment the new utility functions would offset the change in probabilities so that the equality still holds. Another way to think of state prices as being fixed with new probabilities but utility functions unchanged, is to think of a different value of the initial endowment. If the endowment is made very large in one state and very small in the other, then this will offset the large change in probabilities of the two states. The analysis of the single agent can be carried over to an economy with many agents with suitable technical assumptions. Things become particularly easy when the equilibrium can be analyzed by considering the utility of a single, 'representative' agent, whose endowment is the sum of all the agents' endowments. An equilibrium then occurs only if this representative investor has the initial endowment as the solution to the utility maximization problem and hence does not need to trade in the market with the given prices. In this case the aggregate endowment plays a crucial role. Increasing the probability of a state while holding prices and the utility function of the representative investor constant must imply that the aggregate endowment is different with more endowment (low marginal utility) in the states with high probability and low endowment (high marginal utility) in the states with low probability. This intuition is very important when we discuss the Capital Asset Pricing Model later in the course.

A market where we are able to separate out the financial decisions as above is the one we will have in our mind throughout this course. But do keep in mind that this leaves out many interesting issues in the interaction between real markets and financial markets. For example, it is easy to imagine that an incomplete financial market (i.e. one which does not allow any distribution of wealth across states and time periods) makes it impossible for agents to realize consumption plans that they would find optimal in a complete market.

This in turn may change equilibrium prices on real markets because it changes investment behavior. For example, returning to the house market, the fact that financial markets allows young agents to borrow against future income, makes it possible for more consumers to buy a house early in their lives. If all of a sudden we removed the possibility of borrowing we could imagine that house prices would drop significantly, since the demand would suddenly decrease.

Chapter 5

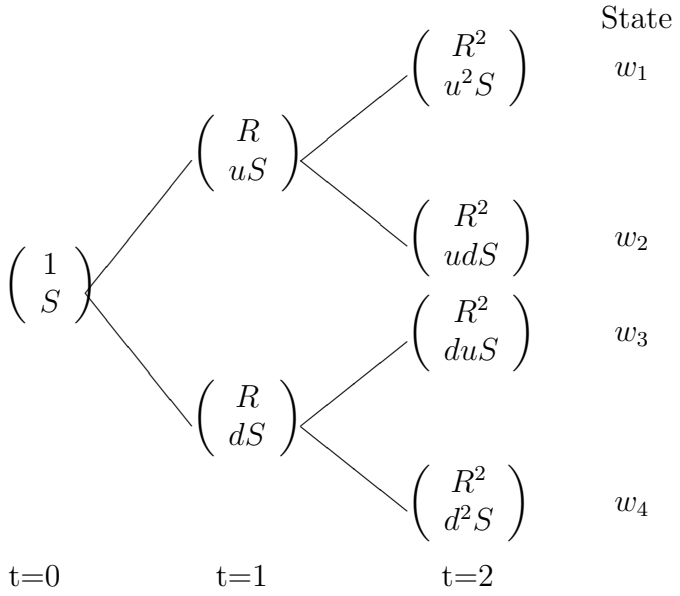
Arbitrage pricing in the multi-period model

5.1 An appetizer

It is fair to argue that to get realism in a model with finite state space we need the number of states to be large. After all, why would the stock take on only two possible values at the expiration date of the option? On the other hand, we know from the previous section that in a model with many states we need many securities to have completeness, which (in arbitrage-free models) is a requirement for pricing every claim. And if we want to price an option using only the underlying stock and a money market account, we only have two securities to work with. Fortunately, there is a clever way out of this.

Assume that over a short time interval the stock can only move to two different values and split up the time interval between 0 and T (the maturity date of an option) into small intervals in which the stock can be traded. Then it turns out that we can have both completeness and therefore arbitrage pricing even if the number of securities is much smaller than the number of states. Again, before we go into the mathematics, we give an example to help with the intuition.

Assume that $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ and that there are three dates: $t \in \{0, 1, 2\}$. We specify the behavior of the stock and the money market account as follows: Assume that $0 < d < R < u$ and that $S > 0$. Consider the following graph:



At time 0 the stock price is S, the money market account is worth 1. At time 1, if the state of the world is ω_1 or ω_2, the prices are uS and R , respectively, whereas if the true state is ω_3 or ω_4, the prices are dS and R. And finally, at time $t = 2$, the prices of the two instruments are as shown in the figure above. Note that $\omega \in \Omega$ describes a whole "sample path" of the stock price process and the money market account, i.e. it tells us not only the final time 2 value, but the entire history of values up to time 2.

Now suppose that we are interested in the price of a European call option on the stock with exercise price K and maturity $T = 2$. At time 2, we know it is worth

$$C_2(\omega) = [S_2(\omega) - K]^+$$

where $S_2(\omega)$ is the value of the stock at time 2 if the true state is ω.

At time 1, if we are in state ω_1 or ω_2, the money market account is worth R and the stock is worth uS, and we know that there are only two possible time 2 values, namely (R^2, u^2S) or (R^2, duS). But then we can use the argument of the one period example to see that at time 1 in state ω_1 or ω_2 we can replicate the calls payoff by choosing a suitable portfolio of stock and money market account: Simply solve the system:

$$au^2S + bR^2 = \left[u^2S - K\right]^+ \equiv C_{uu}$$

$$aduS + bR^2 = [duS - K]^+ \equiv C_{du}$$

for (a, b) and compute the price of forming the portfolio at time 1. We find

$$a = \frac{C_{uu} - C_{du}}{uS\,(u-d)},\ b = \frac{uC_{du} - dC_{uu}}{(u-d)\,R^2}.$$

The price of this portfolio is

$$\begin{aligned} auS + bR &= \frac{R}{R}\frac{(C_{uu} - C_{du})}{(u-d)} + \frac{uC_{du} - dC_{uu}}{(u-d)\,R} \\ &= \frac{1}{R}\left[\frac{(R-d)}{(u-d)}C_{uu} + \frac{(u-R)}{(u-d)}C_{ud}\right] =: C_u \end{aligned}$$

This is clearly what the call is worth at time $t = 1$ if we are in ω_1 or ω_2, i.e. if the stock is worth uS at time 1. Similarly, we may define $C_{ud} := [udS - K]^+$ (which is equal to C_{du}) and $C_{dd} = [d^2S - K]^+$. And now we use the exact same argument to see that if we are in state ω_3 or ω_4, i.e. if the stock is worth dS at time 1, then at time 1 the call should be worth C_d where

$$C_d := \frac{1}{R}\left[\frac{(R-d)}{(u-d)}C_{ud} + \frac{(u-R)}{(u-d)}C_{dd}\right].$$

Now we know what the call is worth at time 1 depending on which state we are in: If we are in a state where the stock is worth uS, the call is worth C_u and if the stock is worth dS, the call is worth C_d.

Looking at time 0 now, we know that all we need at time 1 to be able to "create the call", is to have C_u when the stock goes up to uS and C_d when it goes down. But that we can accomplish again by using the one-period example: The cost of getting $\binom{C_u}{C_d}$ is

$$C_0 := \frac{1}{R}\left[\frac{(R-d)}{(u-d)}C_u + \frac{(u-R)}{(u-d)}C_d\right].$$

If we let $q = \frac{R-d}{u-d}$ and if we insert the expressions for C_u and C_d, noting that $C_{ud} = C_{du}$, we find that

$$C_0 = \frac{1}{R^2}\left[q^2C_{uu} + 2q\,(1-q)\,C_{ud} + (1-q)^2\,C_{dd}\right]$$

which the reader will recognize as a discounted expected value, just as in the one period example. (Note that the representation as an expected value does not hinge on $C_{ud} = C_{du}$.)

The important thing to understand in this example is the following: Starting out with the amount C_0, an investor is able to form a portfolio in the stock and the money market account which produces the payoffs C_u or C_d

at time 1 depending on where the stock goes. Now without any additional costs, the investor can rearrange his/her portfolio at time 1, such that at time 2, the payoff will match that of the option. Therefore, at time 0 the price of the option must be C_0.

This "dynamic hedging" argument is the key to pricing derivative securities in discrete-time, finite state space models. We now want to understand the mathematics behind this example.

5.2 Price processes, trading and arbitrage

Given a probability space $(\Omega, \mathcal{F}, P)$ with Ω finite, let $\mathcal{F} := 2^{\Omega}$ (i.e. the set of all subsets of Ω) and assume that $P(\omega) > 0$ for all $\omega \in \Omega$. Also assume that there are T+1 dates, starting at date 0, ending at date T. To formalize how information is revealed through time, we introduce the notion of a *filtration*:

Definition 23 *A filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t=0}^{T}$ is an increasing sequence of σ-algebras contained in $\mathcal{F}$: $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_T$.*

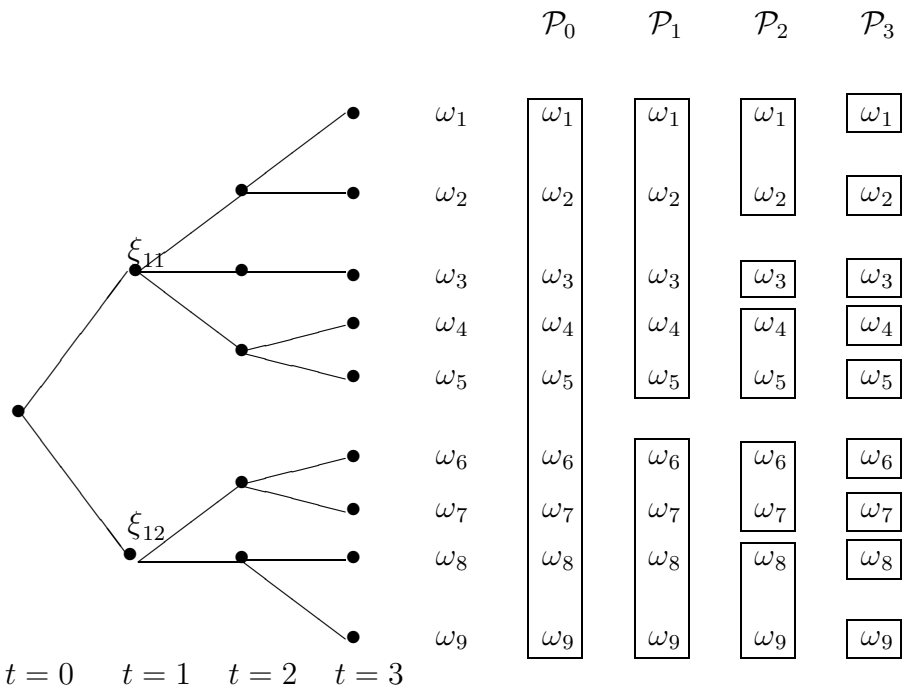
We will always assume that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_T = \mathcal{F}$. Since Ω is finite, it will be easy to think of the $\sigma-$algebras in terms of *partitions:*

Definition 24 *A partition $\mathcal{P}_t$ of Ω is a collection of non-empty subsets of Ω such that*

- $\bigcup_{P_i \in \mathcal{P}_t} P_i = \Omega$
- $P_i \cap P_j = \emptyset$ whenever $i \neq j, P_i, P_j \in \mathcal{P}_t$.

Because Ω is finite, there is a one-to-one correspondence between partitions and $\sigma-$algebras: The elements of $\mathcal{P}_t$ corresponds to *the atoms* of $\mathcal{F}_t$.

The concepts we have just defined are well illustrated in an *event-tree:*



The event tree illustrates the way in which we imagine information about the true state being revealed over time. At time $t = 1$, for example, we may find ourselves in one of two nodes: ξ_{11} or ξ_{12}. If we are in the node ξ_{11}, we know that the true state is in the set $\{\omega_1, \omega_2, \ldots, \omega_5\}$, but we have no more knowledge than that. In ξ_{12}, we know (only) that $\omega \in \{\omega_6, \omega_7, \ldots, \omega_9\}$. At time $t = 2$ we have more detailed knowledge, as represented by the partition $\mathcal{P}_2$. Elements of the partition $\mathcal{P}_t$ are events which we can decide as having occurred or not occurred at time t, *regardless* of what the true ω is. At time 1, we will always know whether $\{\omega_1, \omega_2, \ldots, \omega_5\}$ has occurred or not, regardless of the true ω. If we are at node ξ_{12}, we would be able to rule out

the event $\{\omega_1, \omega_2\}$ also at time 1, but if we are at node ξ_{11}, we will not be able to decide whether this event has occurred or not. Hence $\{\omega_1, \omega_2\}$ is not a member of the partition.

Make sure you understand the following

Remark 2 *A random variable defined on $(\Omega, \mathcal{F}, P)$ is measurable with respect to $\mathcal{F}_t$ precisely when it is constant on each member of $\mathcal{P}_t$.*

A stochastic process $X := (X_t)_{t=0,\ldots,T}$ is a sequence of random variables $X_0, X_1, \ldots, X_T$. The process is *adapted* to the filtration $\mathcal{F}$ if X_t is $\mathcal{F}_t-$measurable (which we will often write: $X_t \in \mathcal{F}_t$) for $t = 0, \ldots, T$. Returning to the event tree setup, it must be the case, for example, that $X_1(\omega_1) = X_1(\omega_5)$ if X is adapted, but we may have $X_1(\omega_1) \neq X_1(\omega_6)$.

Given an event tree, it is easy to construct adapted processes: Just assign the values of the process using the nodes of the tree. For example, at time 1, there are two nodes ξ_{11} and ξ_{12}. You can choose one value for X_1 in ξ_{11} and another in ξ_{12}. The value chosen in ξ_{11} will correspond to the value of X_1 on the set $\{\omega_1, \omega_2, \ldots, \omega_5\}$, the value chosen in ξ_{12} will correspond to the common value of X_1 on the set $\{\omega_6, \ldots, \omega_9\}$. When X_t is constant on an event A_t we will sometimes write $X_t(A_t)$ for this value. At time 2 there are five different values possible for X_2. The value chosen in the top node is the value of X_2 on the set $\{\omega_1, \omega_2\}$.

As we have just seen it is convenient to speak in terms of the event tree associated with the filtration. From now on we will refer to the event tree as the graph Ξ and use ξ to refer to the individual nodes. The notation $p(\xi)$ will denote the probability of the event associated with ξ; for example $P(\xi_{11}) = P(\{\omega_1, \omega_2, \ldots, \omega_5\})$. This graph Ξ will also allow us to identify adapted processes with vectors in $\mathbb{R}^{\Xi}$. The following inner products on the space of adapted processes will become useful later: Let X, Y be adapted processes and define

$$\begin{aligned}
\sum_{\xi\in\Xi} X(\xi)Y(\xi) &\equiv \sum_{\{(t,A_u):A_u\in\mathcal{P}_t, 0\le t\le T\}} X_t(A_u)Y_t(A_u) \\
E\sum_{\xi\in\Xi} X(\xi)Y(\xi) &\equiv \sum_{\xi\in\Xi} P(\xi)X(\xi)Y(\xi) \\
&\equiv \sum_{\{(t,A_u):A_u\in\mathcal{P}_t, 0\le t\le T\}} P(A_u)X_t(A_u)Y_t(A_u)
\end{aligned}$$

Now we are ready to model financial markets in multi-period models.

Given is a vector of adapted *dividend processes*

$$\delta = (\delta^1, \ldots, \delta^N)$$

and a vector of adapted *security price processes*

$$S = (S^1, \ldots, S^N).$$

The interpretation is as follows: $S_t^i(\omega)$ is the price of security i at time t if the state is ω. Buying the $i'th$ security at time t ensures the buyer (and obligates the seller to deliver) the remaining dividends $\delta_{t+1}^i, \delta_{t+2}^i, \ldots, \delta_T^i$.[1] Hence the security price process is to be interpreted as an *ex-dividend* price process and in particular we should think of S_T as 0. In all models considered in these notes we will also assume that there is a money market account which provides locally riskless borrowing and lending. This is modeled as follows: Given an adapted process - *the spot rate process*

$$\rho = (\rho_0, \rho_1, \ldots, \rho_{T-1}).$$

To make the math work, all we need to assume about this process is that it is strictly greater than -1 at all times and in all states, but for modelling purposes it is desirable to have it non-negative. Now we may define the money market account as follows:

Definition 25 *The money market account has the security price process*

$$\begin{aligned} S_t^0 &= 1, \qquad t = 0, 1, \ldots, T-1 \\ S_T^0 &= 0. \end{aligned}$$

and the dividend process

$$\begin{aligned} \delta_t^0(\omega) &= \rho_{t-1}(\omega) \text{ for all } \omega \text{ and } t = 1, \ldots, T-1, \\ \delta_T^0(\omega) &= 1 + \rho_{T-1}(\omega). \end{aligned}$$

This means that if you buy one unit of the money market account at time t you will receive a dividend of ρ_t at time $t+1$. Since ρ_t is known already at time t, the dividend received on the money market account in the next period $t+1$ is known at time t. Since the price is also known to be 1 you know that placing 1 in the money market account at time t and selling the asset at time $t+1$ will give you $1+\rho_t$.This is why we refer to this asset as a locally riskless asset. You may of course also choose to keep the money in the money market account and receive the stream of dividends. Reinvesting the dividends in the money market account will make this account grow according to the process R defined as

$$R_t = (1+\rho_0) \cdots (1+\rho_{t-1}).$$

[1] We will follow the tradition of probability theory and often suppress the ω in the notation.

We will need this process to discount cash flows between arbitrary periods and therefore introduce the following notation:

$$R_{s,t} \equiv (1+\rho_s)\cdots(1+\rho_{t-1}).$$

Definition 26 *A trading strategy is an adapted process*

$$\phi = (\phi_t^0, \ldots, \phi_t^N)_{t=0,\ldots,T-1}.$$

and the interpretation is that $\phi_t^i(\omega)$ is the number of the i'th security held at time t if the state is ω. The requirement that the trading strategy is adapted is very important. It represents the idea that the strategy should not be able to see into the future. Returning again to the event tree, when standing in node ξ_{11}, a trading strategy can base the number of securities on the fact that we are in ξ_{11} (and not in ξ_{12}), but not on whether the true state is ω_1 or ω_2.

The dividend stream generated by the trading strategy ϕ is denoted δ^ϕ and it is defined as

$$\delta_0^\phi = -\phi_0 \cdot S_0$$
$$\delta_t^\phi = \phi_{t-1}\cdot(S_t+\delta_t) - \phi_t \cdot S_t \text{ for } t = 1,\ldots,T.$$

Definition 27 *An arbitrage is a trading strategy for which δ_t^ϕ is a positive process, i.e. always nonnegative and $\delta_t^\phi(\omega) > 0$ for some t and ω. The model is said to be arbitrage-free if it contains no arbitrage opportunities.*

In words, there is arbitrage if we can adopt a trading strategy which at no point in time requires us to pay anything but which at some time in some state gives us a strictly positive payout. Note that since we have included the initial payout as part of the dividend stream generated by a trading strategy, we can capture the definition of arbitrage in this one statement. This one statement captures arbitrage both in the sense of receiving money now with no future obligations and in the sense of paying nothing now but receiving something later.

Definition 28 *A trading strategy ϕ is self-financing if it satisfies*

$$\phi_{t-1}\cdot(S_t+\delta_t) = \phi_t \cdot S_t \quad \textit{for } t = 1,\ldots,T.$$

The interpretation is as follows: Think of forming a portfolio ϕ_{t-1} at time $t-1$. Now as we reach time t, the value of this portfolio is equal to $\phi_{t-1}\cdot(S_t+\delta_t)$, and for a self-financing trading strategy, this is precisely the amount of money which can be used in forming a new portfolio at time t. We will let Φ denote the set of self-financing trading strategies.

5.3 No arbitrage and price functionals

We have seen in the one period model that there is equivalence between the existence of a state price vector and absence of arbitrage. In this section we show the multi-period analogue of this theorem.

The goal of this section is to prove the existence of the multi-period analogue of state-price vectors in the one-period model. Let $\mathbb{L}$ denote the set of adapted processes on the given filtration.

Definition 29 *A pricing functional F is a linear functional*

$$F : \mathbb{L} \to \mathbb{R}$$

which is strictly positive, i.e.

$$\begin{aligned} F(X) &\geq 0 \text{ for } X \geq 0 \\ F(X) &> 0 \text{ for } X > 0. \end{aligned}$$

Definition 30 *A pricing functional F is consistent with security prices if*

$$F(\delta^\phi) = 0 \text{ for all trading strategies } \phi.$$

Note that if there exists a consistent pricing functional we may arbitrarily assume that the value of the process $1_{\{t=0\}}$ (i.e. the process which is 1 at time 0 and 0 thereafter) is 1.

By Riesz' representation theorem we can represent the functional F as

$$F(X) = \sum_{\xi \in \Xi} X(\xi) f(\xi)$$

With the convention $F(1_{\{t=0\}}) = 1$, we then note that if there exists a trading strategy ϕ which is initiated at time 0 and which only pays a dividend of 1 in the node ξ, then

$$\phi_0 \cdot S_0 = f(\xi).$$

Hence $f(\xi)$ is the price at time 0 of having a payout of 1 in the node ξ.

Proposition 11 *The model (δ, S) is arbitrage-free if and only if there exists a consistent pricing functional.*

Proof. First, assume that there exists a consistent pricing functional F. Any dividend stream δ^ϕ generated by a trading strategy which is positive must have $F(\delta^\phi) > 0$ but this contradicts consistency. Hence there is no arbitrage. The other direction requires more work:

Define the sets

$$\begin{aligned}\mathbb{L}^1 &= \left\{ X \in \mathbb{L} \,\middle|\, X > 0 \text{ and } \sum_{\xi \in \Xi} X(\xi) = 1 \right\} \\ \mathbb{L}^0 &= \left\{ \delta^\phi \in \mathbb{L} \,|\, \phi \text{ trading strategy} \right\}\end{aligned}$$

and think of both sets as subsets of $\mathbb{R}^\Xi$. Note that $\mathbb{L}^1$is convex and compact and that $\mathbb{L}^0$ is a linear subspace, hence closed and convex. By the no arbitrage assumption the two sets are disjoint. Therefore, there exists a separating hyperplane $H(f;\alpha) := \{x \in \mathbb{R}^\Xi : f \cdot x = \alpha\}$ which separates the two sets strictly and we may choose the direction of f such that $f \cdot x \leq \alpha$ for $x \in \mathbb{L}^0$. Since $\mathbb{L}^0$ is a linear subspace we must have $f \cdot x = 0$ for $x \in \mathbb{L}^0$ (why?). Strict separation then gives us that $f \cdot x > 0$ for $x \in \mathbb{L}^1$, and that in turn implies $f \gg 0$ (why?). Hence the functional

$$F(X) = \sum_{\xi \in \Xi} f(\xi) X(\xi)$$

is consistent. ■

By using the same geometric intuition as in Chapter 2, we note that there is a connection between completeness of the market and uniqueness of the consistent price functional:

Definition 31 *The security model is complete if for every $X \in \mathbb{L}$ there exists a trading strategy ϕ such that $\delta_t^\phi = X_t$ for $t \geq 1$.*

If the model is complete and arbitrage-free, there can only be one consistent price functional (up to multiplication by a scalar). To see this, assume that if we have two consistent price functionals F, G both normed to have $F(1_{\{t=0\}}) = G(1_{\{t=0\}}) = 1$. Then for any trading strategy ϕ we have

$$\begin{aligned}0 &= -\phi_0 \cdot S_0 + F(1_{\{t>0\}}\delta^\phi) \\ &= -\phi_0 \cdot S_0 + G(1_{\{t>0\}}\delta^\phi)\end{aligned}$$

hence F and G agree on all processes of the form $1_{\{t>0\}}\delta^\phi$. But they also agree on $1_{\{t=0\}}$ and therefore they are the same since by the assumption of completeness every adapted process can be obtained as a linear combination of these processes.

Given a security price system (π, D), the converse is shown in a way very similar to the one-period case. Assume the market is arbitrage-free and incomplete. Then there exists a process π in $\mathbb{L}$, whose restriction to time

$t \geq 1$ is orthogonal to any dividend process generated by a trading strategy. By letting $\pi_0 = 0$ and choosing a sufficiently small $\varepsilon > 0$, the functional defined by

$$(F + \varepsilon\pi)(\delta^\phi) = \sum_{\xi \in \Xi} (f(\xi) + \varepsilon\pi(\xi))\, \delta^\phi(\xi)$$

is consistent. Hence we have shown:

Proposition 12 *If the market is arbitrage-free, then the model is complete if and only if the consistent price functional is unique.*

5.4 Conditional expectations and martingales

Consistent price systems turn out to be less interesting for computation when we look at more general models, and they do not really explain the strange probability measure q which we saw earlier. We are about to remedy both problems, but first we need to make sure that we can handle conditional expectations in our models and that we have a few useful computational rules at our disposal.

Definition 32 *The conditional expectation of an $\mathcal{F}_u-$measurable random variable X_u given $\mathcal{F}_t$, where $\mathcal{F}_t \subseteq \mathcal{F}_u$, is given by*

$$E(X_u \,|\mathcal{F}_t)(\omega) = \frac{1}{P(A_t)} \sum_{A_v \in \mathcal{P}_u : A_v \subseteq A_t} P(A_v) X_u(A_v) \text{ for } \omega \in A_t$$

where we have written $X_u(A_v)$ for the value of $X_u(\omega)$ on the set A_v and where $A_t \in \mathcal{P}_t$.

We will illustrate this definition in the exercises. Note that we obtain an $\mathcal{F}_t-$measurable random variable since it is constant over elements of the partition $\mathcal{P}_t$. The definition above does not work when the probability space becomes uncountable. Then one has to adopt a different definition which we give here and which the reader may check is satisfied by the random variable given above in the case of finite sample space:

Definition 33 *The conditional expectation of an $\mathcal{F}_u-$measurable random variable X_u given $\mathcal{F}_t$ is a random variable $E(X_u \,|\mathcal{F}_t)$ which is $\mathcal{F}_t-$measurable and satisfies*

$$\int_{A_t} E(X_u \,|\mathcal{F}_t) dP = \int_{A_t} X_u dP$$

for all $A_t \in \mathcal{F}_t$.

It is easy to see that the conditional expectation is linear, i.e. if $X_u, Y_u \in \mathcal{F}_u$ and $a, b \in \mathbb{R}$, then

$$E(aX_u + bY_u \,|\mathcal{F}_t) = aE(X_u \,|\mathcal{F}_t) + bE(Y_u \,|\mathcal{F}_t).$$

We will also need the following computational rules for conditional expectations:

$$\begin{aligned} E(E(X_u \,|\mathcal{F}_t)) &= EX_u && (5.1)\\ E(Z_t X_u \,|\mathcal{F}_t) &= Z_t E(X_u \,|\mathcal{F}_t) \text{ whenever } Z_t \in \mathcal{F}_t && (5.2)\\ E(E(X_u \,|\mathcal{F}_t) \,|\mathcal{F}_s) &= E(X_u \,|\mathcal{F}_s) \text{ whenever } s < t < u && (5.3) \end{aligned}$$

Note that a consequence of (5.2) obtained by letting $X_u = 1$, is that

$$E(Z_t \,|\mathcal{F}_t) = Z_t \quad \text{whenever} \quad Z_t \in \mathcal{F}_t. \tag{5.4}$$

Now we can state the important definition:

Definition 34 *A stochastic process X is a martingale with respect to the filtration $\mathbb{F}$ if it satisfies*

$$E(X_t \,|\mathcal{F}_{t-1}) = X_{t-1} \quad \textit{all } t = 1, \ldots, T.$$

You can try out the definition immediately by showing:

Lemma 13 *A stochastic process defined as*

$$X_t = E(X \,|\mathcal{F}_t) \quad t = 0, 1, \ldots, T$$

where $X \in \mathcal{F}_T$, is a martingale.

Proof. Use (5.3)! ■

Let $E^P(Y; A) \equiv \int_A Y dP$ for any random variable Y and $A \in \mathcal{F}$. Using this notation and the definition (33) of a martingale, this lemma says that

$$E(X; A) = E(X_t; A) \qquad \text{for all } t \text{ and } A \in \mathcal{F}_t$$

When there can be no confusion about the underlying filtration we will often write $E_t(X)$ instead of $E(X \,|\mathcal{F}_t)$.

Two probability measures are equivalent when they assign zero probability to the same sets and since we have assumed that $P(\omega) > 0$ for all ω, the measures equivalent to P will be the ones which assign strictly positive probability to all events.

We will need a way to translate conditional expectations under one measure to conditional expectations under an equivalent measure. To do this we need the *density process:*

Definition 35 *Let the density process Z be defined as*

$$Z_T(\omega) = \frac{Q(\omega)}{P(\omega)}$$

and

$$Z_t = E^P(\, Z_T |\, \mathcal{F}_t) \qquad t = 0, 1, \ldots, T.$$

We will need (but will not prove) the following result of called the *Abstract Bayes Formula*.

Proposition 14 *Let X be a random variable on $(\Omega, \mathcal{F})$. Then*

$$E^Q(\, X |\, \mathcal{F}_t) = \frac{1}{Z_t} E^P(\, X Z_T |\, \mathcal{F}_t).$$

5.5 Equivalent martingale measures

In this section we state and prove what is sometimes known as the fundamental theorem of asset pricing. This theorem will explain the mysterious $q-$probabilities which arose earlier and it will provide an indispensable tool for constructing arbitrage-free models and pricing contingent claims in these models.

We maintain the setup with a money market account generated by the spot rate process ρ and N securities with price- and dividend processes $S = (S^1, \ldots, S^N), \delta = (\delta^1, \ldots, \delta^N)$. Define the corresponding discounted processes $\widetilde{S}, \widetilde{\delta}$ by defining for each $i = 1, \ldots, N$

$$\begin{aligned} \widetilde{S}_t^i &= \frac{S_t^i}{R_{0,t}} \qquad t = 0, \ldots, T, \\ \widetilde{\delta}_t^i &= \frac{\delta_t^i}{R_{0,t}} \qquad t = 1, \ldots, T. \end{aligned}$$

Definition 36 *A probability measure Q on $\mathcal{F}$ is an equivalent martingale measure (EMM) if $Q(\omega) > 0$ all ω and for all $i = 1, \ldots, N$*

$$\widetilde{S}_t^i = E_t^Q \left(\sum_{j=t+1}^{T} \widetilde{\delta}_j^i \right) \qquad t = 0, \ldots, T-1. \tag{5.5}$$

The term martingale measure has the following explanation: Given a (one-dimensional) security price process S whose underlying dividend process only pays dividend δ_T at time T. Then the existence of an EMM will give us

$$\widetilde{S}_t = E_t^Q \left(\widetilde{\delta}_T \right) \qquad t = 0, \ldots, T-1.$$

and therefore the process $(\widetilde{S}_0, \widetilde{S}_1, \ldots, \widetilde{S}_{T-1}, \widetilde{\delta}_T)$ is a martingale, cf. Lemma (13).

We are now ready to formulate and prove what is sometimes known as 'the fundamental theorem of asset pricing' in a version with discrete time and finite state space:

Theorem 15 *In our security market model the following statements are equivalent:*

1. *There are no arbitrage opportunities.*
2. *There exists an equivalent martingale measure.*

Proof. We have already seen that no arbitrage is equivalent to the existence of a consistent price functional F. Therefore, what we show in the following is that there is a one-to-one correspondence between consistent price functionals (up to multiplication by a positive scalar) and equivalent martingale measures. We will need the following notation for the restriction of F to an $\mathcal{F}_t-$measurable random variable: Let δ^X be a dividend process whose only payout is X at time t. Define

$$F_t(X) = F(\delta^X).$$

If we assume (as we do from now on) that $F_0(1) = 1$, we may think of $F_t(1_A)$ as the price a time 0 of a claim (if it trades) paying off 1 at time t if $\omega \in A$. Note that since we have a assumed the existence of a money market account, we have

$$F_T(R_{0,T}) = 1 \tag{5.6}$$

First, assume there is no arbitrage and let F be a consistent price functional. Our candidate as equivalent martingale measure is defined as follows:

$$Q(A) = F_T(1_A R_{0,T}) \quad A \in \mathcal{F} \equiv \mathcal{F}_T. \tag{5.7}$$

By the strict positivity, linearity and (5.6) we see that Q is a probability measure which is strictly positive on all ω.We may write (5.7) as

$$E^Q 1_A = F_T(1_A R_{0,T}) \quad A \in \mathcal{F} \equiv \mathcal{F}_T$$

and by writing a random variable X as a sum of constants times indicator functions, we note that

$$E^Q(X) = F_T(X R_{0,T}) \tag{5.8}$$

Now we want to check the condition (5.5). By definition (33) this is equivalent to showing that for every security we have

$$E^Q(1_A\widetilde{S}_t^i) = E^Q\left(1_A\sum_{j=t+1}^{T}\widetilde{\delta}_j^i\right) \qquad t=1,\ldots,T. \tag{5.9}$$

Consider for given $A\in\mathcal{F}_t$ the following trading strategy ϕ:

- Buy one unit of stock i at time 0 (this costs S_t^i). Invest all dividends before time t in the money market account and keep them there at least until time t.
- At time t, if $\omega\in A$ (and this we know at time t since $A\in\mathcal{F}_t$) sell the security and invest the proceeds in the money market account, i.e. buy S_t^i units of the $0'$th security and roll over the money until time T.
- If $\omega\notin A$, then hold the i'th security to time T.

This strategy clearly only requires an initial payment of S_0^i. The dividend process generated by this strategy is non-zero only at time 0 and at time T. At time T the dividend is

$$\begin{aligned}\delta_T^\phi &= 1_A R_{t,T}\left(S_t^i+\sum_{j=1}^{t}\delta_j^i R_{j,t}\right)+1_{A^c}\sum_{j=1}^{T}\delta_j^i R_{j,T}\\ &= 1_A R_{0,T}\left(\widetilde{S}_t^i+\sum_{j=1}^{t}\widetilde{\delta}_j^i\right)+1_{A^c}\sum_{j=1}^{T}\delta_j^i R_{j,T}\end{aligned}$$

One could also choose to just buy the i'th security and then roll over the dividends to time T. Call this strategy ψ.This would generate a terminal dividend which we may write in a complicated but useful way as

$$\begin{aligned}\delta_T^\psi &= 1_A\sum_{j=1}^{T}\delta_j^i R_{j,T}+1_{A^c}\sum_{j=1}^{T}\delta_j^i R_{j,T}\\ &= 1_A R_{0,T}\sum_{j=1}^{T}\widetilde{\delta}_j^i+1_{A^c}\sum_{j=1}^{T}\delta_j^i R_{j,T}\end{aligned}$$

The dividend stream of both strategies at time 0 is $-S_0^i$. We therefore have

$$F_T(\delta_T^\phi) = F_T(\delta_T^\psi)$$

which in turn implies

$$F_T\left(1_A R_{0,T}\left(\widetilde{S}_t^i + \sum_{j=1}^{t}\widetilde{\delta}_j^i\right)\right) = F_T\left(1_A R_{0,T}\sum_{j=1}^{T}\widetilde{\delta}_j^i\right)$$

i.e.

$$F_T\left(1_A R_{0,T}\widetilde{S}_t^i\right) = F_T\left(1_A R_{0,T}\sum_{j=t+1}^{T}\widetilde{\delta}_j^i\right).$$

Now use (5.8) to conclude that

$$E^Q(1_A\widetilde{S}_t^i) = E^Q(1_A\sum_{j=t+1}^{T}\widetilde{\delta}_j^i)$$

and that is what we needed to show. Q is an equivalent martingale measure.

Now assume that Q is an equivalent martingale measure. Define for an arbitrary dividend process δ

$$F(\delta) = E^Q\sum_{j=0}^{T}\widetilde{\delta}_j$$

Clearly, F is linear and strictly positive. Now consider the dividend process δ^ϕ generated by some trading strategy ϕ. To show consistency we need to show that

$$\phi_0\cdot\widetilde{S}_0 = E^Q\sum_{j=1}^{T}\widetilde{\delta}_j^\phi.$$

Notice that we know that for individual securities we have

$$\widetilde{S}_0^i = E^Q\sum_{j=1}^{T}\widetilde{\delta}_j^i.$$

We only need to extend that to portfolios. We do some calculations (where we make good use of the rule $E^Q E_j = E^Q E_{j-1}$)

$$\begin{aligned}
E^Q\sum_{j=1}^{T}\widetilde{\delta}_j^\phi &= E^Q\left(\sum_{j=1}^{T}\phi_{j-1}\cdot(\widetilde{S}_j+\widetilde{\delta}_j) - \phi_j\cdot\widetilde{S}_j\right)\\
&= E^Q\left(\sum_{j=1}^{T}\phi_{j-1}\cdot\left(E_j^Q\left(\sum_{k=j}^{T}\widetilde{\delta}_k\right)\right) - \phi_j\cdot E_j^Q\left(\sum_{k=j+1}^{T}\widetilde{\delta}_k\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&= E^Q\left(\sum_{j=1}^{T}\phi_{j-1}\cdot\left(E_{j-1}^Q\left(\sum_{k=j}^{T}\widetilde{\delta}_k\right)\right)-\sum_{j=2}^{T}\phi_{j-1}\cdot E_{j-1}^Q\left(\sum_{k=j}^{T}\widetilde{\delta}_k\right)\right)\\
&= E^Q\left(\phi_0\cdot\left(E_0^Q\sum_{k=1}^{T}\widetilde{\delta}_k\right)\right)\\
&= E^Q\left(\phi_0\cdot\widetilde{S}_0\right)\\
&= \phi_0\cdot\widetilde{S}_0
\end{aligned}
$$

This completes the proof. ■

Earlier, we established a one-to-one correspondence between consistent price functionals (normed to 1 at date 0) and equivalent martingale measures. Therefore we have also proved the following

Corollary 16 *Assume the security model is arbitrage-free. Then the market is complete if and only if the equivalent martingale measure is unique.*

Another immediate consequence from the definition of consistent price functionals and equivalent martingale measures is the following

Corollary 17 *Let the security model defined by (S,δ) (including the money market account) on $(\Omega,P,\mathcal{F},\mathbb{F})$ be arbitrage-free and complete. Then the augmented model obtained by adding a new pair (S^{N+1},δ^{N+1}) of security price and dividend processes is arbitrage-free if and only if*

$$\widetilde{S}_t^{N+1}=E_t^Q\left(\sum_{j=t+1}^{T}\widetilde{\delta}_j^{N+1}\right) \tag{5.10}$$

i.e.

$$\frac{S_t^{N+1}}{R_{0,t}}=E_t^Q\left(\sum_{j=t+1}^{T}\frac{\delta_j^{N+1}}{R_{0,j}}\right)$$

where Q is the unique equivalent martingale measure for (S,δ).

In the special case where the discount rate is deterministic the expression simplifies somewhat. For ease of notation assume that the spot interest rate is not only deterministic but also constant and let $R=1+\rho$. Then (5.10) becomes

$$
\begin{aligned}
S_t^{N+1} &= R^tE_t^Q\left(\sum_{j=t+1}^{T}\frac{\delta_j^{N+1}}{R_{0,j}}\right) \qquad (5.11)\\
&= E_t^Q\left(\sum_{j=t+1}^{T}\frac{\delta_j^{N+1}}{R^{j-t}}\right)
\end{aligned}
$$

5.6 One-period submodels

Before we turn to applications we note a few results for which we do not give proofs. The results show that the one-period model which we analyzed earlier actually is very useful for analyzing multi-period models as well.

Given the market model with the $N-$dimensional security price process S and dividend process δ and assume that a money market account exists as well. Let $A_t \in \mathcal{P}_t$ and let

$$N(A_t) \equiv |\{B \in \mathcal{P}_{t+1} : B \subseteq A_t\}|.$$

This number is often referred to as the splitting index at A_t. In our graphical representation where the set A_t is represented as a node in a graph, the splitting index at A_t is simply the number of vertices leaving that node. At each such node we can define a one-period submodel as follows: Let

$$\pi(t, A_t) \equiv \left(1, S_t^1(A_t), \ldots, S_t^N(A_t)\right).$$

Denote by $B_1, \ldots, B_{N(A_t)}$ the members of $\mathcal{P}_{t+1}$ which are subsets of A_t and define

$$D(t, A_t) \equiv \begin{pmatrix} 1+\rho_t(A_t) & \cdots & 1+\rho_t(A_t) \\ S_{t+1}^0(B_1)+\delta_{t+1}^0(B_1) & \vdots & S_{t+1}^0(B_{N(A_t)})+\delta_{t+1}^0(B_{N(A_t)}) \\ \vdots & & \vdots \\ S_{t+1}^N(B_1)+\delta_{t+1}^N(B_1) & \cdots & S_{t+1}^N(B_{N(A_t)})+\delta_{t+1}^N(B_{N(A_t)}) \end{pmatrix}.$$

Then the following results hold:

Proposition 18 *The security market model is arbitrage-free if and only if the one-period model* $(\pi(t, A_t), D(t, A_t))$ *is arbitrage-free for all* (t, A_t) *where* $A_t \in \mathcal{P}_t$.

Proposition 19 *The security market model is complete if and only if the one-period model* $(\pi(t, A_t), D(t, A_t))$ *is complete for all* (t, A_t) *where* $A_t \in \mathcal{P}_t$.

In the complete, arbitrage-free case we obtain from each one-period submodel a unique state price vector $\psi(t, A_t)$ and by following the same procedure as outlined in chapter (4) we may decompose this into a discount factor, which will be $1+\rho_t(A_t)$, and a probability measure $q_1, \ldots, q_{N(A_t)}$. The probabilities thus obtained are then the conditional probabilities $q_i = Q(B_i \,|A_t)$ for $i = 1, \ldots, N(A_t)$. From these conditional probabilities the martingale measure can be obtained.

The usefulness of these local results is that we often build multi-period models by repeating the same one-period structure.We may then check absence of arbitrage and completeness by looking at a one-period submodel instead of the whole tree.

5.7 The multi-period model on matrix form

As a final curiosity we note that it is in fact possible to embed a multi-period model into one giant one-period model by stacking the one-period submodels defined above into a giant matrix. Instead of giving the abstract notation for how this is done, we indicate for the two-period model of chapter (5) how this is done. Consider the following table in which we have defined a set of 'elementary securities':

	0	1,1	1,2	2,1	2,2	2,3	2,4
S_1^0	-1	R	R	0	0	0	0
S_1^1	$-S$	uS	dS	0	0	0	0
$S^0(\{\omega_{1,}\omega_2\})$	0	$-R$	0	R^2	R^2	0	0
$S^1(\{\omega_1,\omega_2\})$	0	$-uS$	0	u^2S	duS	0	0
$S^0(\{\omega_3,\omega_4\})$	0	0	$-R$	0	0	R^2	R^2
$S^1(\{\omega_3,\omega_4\})$	0	0	$-dS$	0	0	udS	d^2S

Each elementary security is to be thought of as arising from buying the security at one node and selling at the successor nodes. The pairs 1,1 ; 1,2 etc. in the top row are to be read as date 1, partition element 1; date 1 partition element 2, etc. Note that the setup is very much as in the application of Stiemke's lemma to one-period models in that we include negative prices for one date and positive prices for the subsequent date. Define

$$D_{big} = \begin{pmatrix} -1 & R & R & 0 & 0 & 0 & 0 \\ -S & uS & dS & 0 & 0 & 0 & 0 \\ 0 & -R & 0 & R^2 & R^2 & 0 & 0 \\ 0 & -uS & 0 & u^2S & duS & 0 & 0 \\ 0 & 0 & -R & 0 & 0 & R^2 & R^2 \\ 0 & 0 & -dS & 0 & 0 & udS & d^2S \end{pmatrix}$$

What you can check for yourself now is that we can define a trading strategy as a vector $\theta \in \mathbb{R}^6$ and then interpret $D_{big}^{top}\theta$ as the dividend process generated by the trading strategy. A self-financing strategy would be one for which the dividend process was non-zero at all dates $1, \ldots, T-1$ (although this could easily be relaxed to a definition of self-financing up to a liquidation

date $t < T$). Arbitrage may then be defined as a trading strategy generating a positive, non-zero dividend stream. If the market is arbitrage-free, the corresponding vector of state prices is an element of $\mathbb{R}^7$ and if we normalize the first component to be 1, the state prices correspond to time-zero prices of securities delivering one unit of account at nodes of the tree.

We will not go further into this but note that it may be a useful way of representing a multi-period model when one wants to introduce short-selling constraints into the model.

Chapter 6

Option pricing

The classical application of the arbitrage pricing machinery we have developed is to the pricing of options. The pricing models we obtain are used with minor modifications all over the world as the basis for trading billions of dollars worth of contracts every day. For students planning to become traders of financial derivatives this of course gives plenty of motivation for learning these models. But recent collapses of financial institutions have also reminded us that financial managers and executives must understand the way the derivatives markets work. A manager who understands the markets well may use them for effective risk management and will be able to implement effective control mechanisms within a firm to make sure that traders use the markets in accordance with the firm's overall objectives.

From a theoretical perspective, options are very important in several areas of finance. We will see later in the course how they are indispensable for our understanding of a firm's choice of capital structure. Also, a modern theory of capital budgeting relies critically on recognizing options involved in projects, so-called *real options.* And in actuarial science options appear when modelling reinsurance contracts.

6.1 Terminology

A *European (American) call option* on an *underlying security* S, with *strike price* K and *expiration date* T, gives the owner the right, but **not** the obligation, to buy S at a price of K at (up to and including) time T.

A *European (American) put option* on an *underlying security* S, with *strike price* K and *expiration date* T, gives the owner the right, but **not** the obligation, to sell S at a price of K at (up to and including) time T.

The strike price is also referred to as the *exercise price,* and using the

right to buy or sell is referred to as *exercising* the option.

There is no good reason for the American/European terminology - both types are traded in America and Europe.

In the definition above, we think of the person selling a call option (say), often referred to as the person *writing* an option, as actually delivering the underlying security to the option holder if the option holder decides to exercise. This is referred to as *physical delivery.* In reality, options are often *cash settled.* This means that instead of the option holder paying K to the writer of the call and the writer delivering the stock, the holder merely receives an amount $S_T - K$ from the option writer.

Some common examples of options are *stock options* in which the underlying security is a stock, *currency options* in which the underlying security is a foreign currency and where the strike price is to be thought of as an exchange rate, *bond options* which have bonds as underlying security and *index options* whose underlying security is not really a security but a stock market index (and where the contracts are then typically cash settled.) It will always be assumed that the underlying security has non-negative value.

6.2 Diagrams, strategies and put-call parity

Before we venture into constructing exact pricing models we develop some feel for how these instruments work. In this section we focus on what can be said about options if all we assume is that all securities (stocks, bonds, options) can be bought and sold in arbitrary quantities at the given prices with no transactions costs or taxes. This assumption we will refer to as an assumption of *frictionless* markets. We will also assume that at any time t and for any date $T > t$, there exists a zero coupon bond with maturity T in the market whose price at time t is $d(t, T)$.

An immediate consequence of our frictionless markets assumption is the following

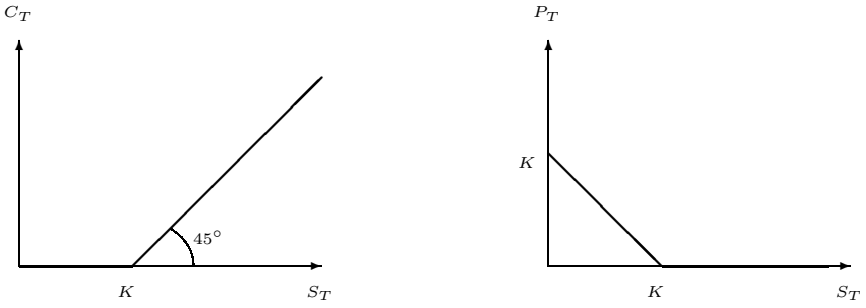
Proposition 20 *The value of an American or European call option at the expiration date T is equal to $C_T = \max(S_T - K, 0)$, where S_T is the price of the underlying security at time T. The value of an American or European put option at the expiration date T is equal to* $\max(K - S_T, 0)$.

Proof. Consider the call option. If $S_T < K$, we must have $C_T = 0$, for if $C_T > 0$ you would sell the option, receive a positive cash flow, and there would be no exercise.[1] If $S_T \geq K$, we must have $C_T = S_T - K$. For if

[1]Actually, here we need to distinguish between whether the person who bought the

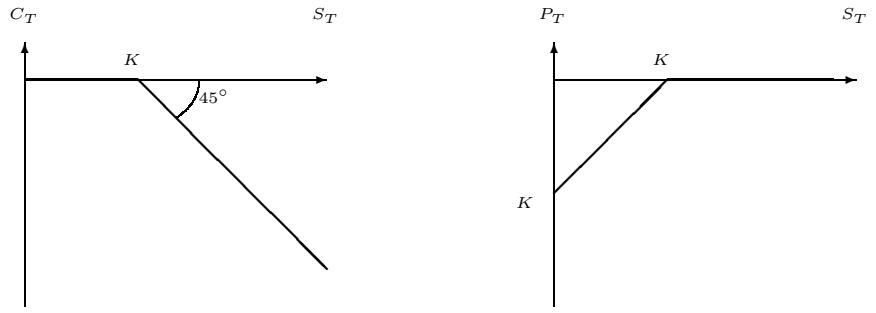
$C_T > S_T - K$ you would sell the option and buy the stock. After the option has been exercised, you are left with a total cash flow of $C_T - S_T + K > 0$, and you would have no future obligations arising from this trade. If $C_T < S_T - K$, buy the option, exercise it immediately, and sell the stock. The total cash flow is $-C_T + S_T - K > 0$, and again there would be no future obligations arising from this trade. The argument for the put option is similar.

We often represent payoffs of options at an exercise date using payoff diagrams, which show the value of the option as a function of the value of the underlying:

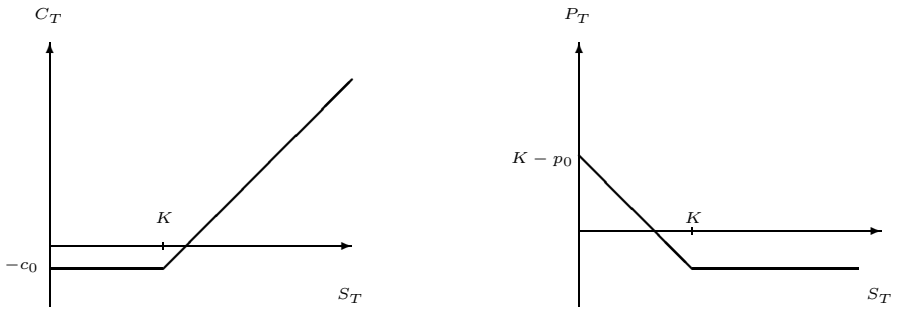


option is an idiot or a complete idiot. Both types are not very smart to pay something for the option at time T. The idiot, however, would realize that there is no reason to pay K to receive the stock which can be bought for less in the market. The complete idiot would exercise the option. Then you as the person having sold the option would have to buy the stock in the market for S_T , but that would be more than financed by the K you received from the complete idiot.

Of course, you can turn these hockey sticks around in which case you are looking at the value of a written option:



Note that we are only looking at the situation at an exercise date (i.e. date T for a European option). Sometimes we wish to take into account that the option had an initial cost at date 0, c_0 for a call, p_0 for a put, in which case we get the following profit diagrams:

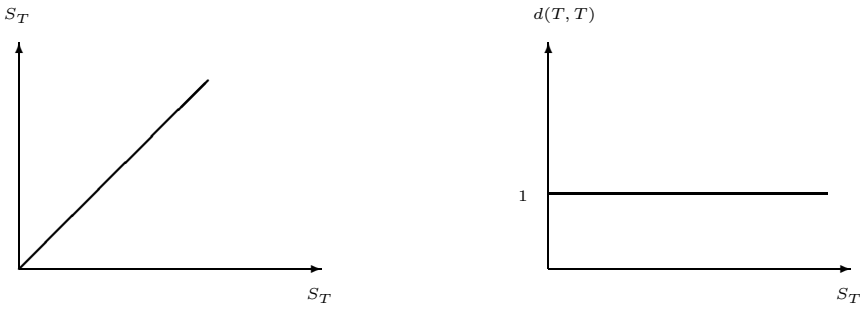


Of course, we are slightly allergic to subtracting payments occurring at different dates without performing some sort of discounting. Therefore, one may also choose to represent the prices of options by their time T forward discounted values $\frac{c_0}{d(0,T)}$ and $\frac{p_0}{d(0,T)}$.

The world of derivative securities is filled with special terminology and here are a few additions to your vocabulary: A call option with strike price K is said to be *(deep) in-the-money* at time t if $S_t > K$ ($S_t \gg K$).The opposite situation $S_t < K$ ($S_t \ll K$) is referred to as the call option being *(deep) out-of-the-money.* If $S_t \approx K$, the option is said to be *at-the-money.* The same terminology applies to put options but with 'opposite signs': A

put option is in-the-money if $S_t < K$.

The diagrams we have seen so far considered positions consisting of just one option. We considered a *long position,* i.e. a position corresponding to holding the option, and we considered a *short position,* i.e. a position corresponding to having written an option. One of the attractive features of options is that they can be combined with positions in other options, the underlying security and bonds to produce more complicated payoffs than those illustrated in the profit diagrams above. We will see examples of this in the exercises. Note that you should think of the payoff diagram for holding the stock and the diagram for holding the bond as being represented by:



Until further notice we will assume that the stock does not pay any dividends in the time interval $[0, T]$. This means that if you own the stock you will not receive any cash unless you decide to sell the stock. With this assumption and the maintained assumption of frictionless markets we will give some restrictions on option prices which follow solely from arbitrage considerations.

The most important relation is the so-called *put-call parity* for European options. Consider the portfolio strategy depicted in the table below and the associated cash flows at time t and time T. Assume that both options are European, expire at date T and have strike price equal to K :

strategy\cashflow	date t	date $T, S_T \leq K$	date $T, S_T > K$
sell 1 call	c_t	0	$K - S_T$
buy 1 put	$-p_t$	$K - S_T$	0
buy stock	$-S_t$	S_T	S_T
sell K bonds	$Kd(t,T)$	$-K$	$-K$
total cash flow	must be 0	0	0

Note that we have constructed a portfolio which gives a payoff of 0 at time T no matter what the value of S_T. Since the options are European we need not consider any time points in (t, T). This portfolio must have price 0, or else there would be an obvious arbitrage strategy. If, for example, the portfolio

had positive value, we would sell the portfolio (corresponding to reversing the strategy in the table) and have no future obligations. In other words we have proved that in a frictionless market we have the following

Proposition 21 *(Put-call parity) The price c_t of a European call and the price p_t of a European put option with expiration date T and exercise price K must satisfy*

$$c_t - p_t = S_t - Kd(t,T).$$

Note one simple but powerful consequence of this result: When deciding which parameters may influence call and put prices the put-call parity gives a very useful way of testing intuitive arguments. If S_t, K and $d(t,T)$ are fixed, then a change in a parameter which produces a higher call price, must produce a higher put-price as well. One would easily for example be tricked into believing that in a model where S_T is stochastic, a higher mean value of S_T given S_t would result in a higher call price since the call option is more likely to finish in-the-money and that it would result in a lower put price since the put is more likely then to finish out-of-the money. But if we assume that S_t and the interest rate are held fixed, put-call parity tells us that this line of reasoning is wrong.

Also note that for $K = \frac{S_t}{d(t,T)}$ we have $c_t = p_t$. This expresses the fact that the exercise price for which $c_t = p_t$ is equal to the *forward* price of S at time t. A *forward* contract is an agreement to buy the underlying security at the expiration date T of the contract at a price of F_t. Note that F_t is specified at time t and that the contract unlike an option forces the holder to buy. In other words you can lose money at expiration on a forward contract. The *forward price* F_t is decided so that the value of the forward contract at date t is 0. Hence the forward price is *not* a price to be paid for the contract at date t. It is more like the exercise price of an option. Which value of F_t then gives the contract a value of 0 at date t? Consider the following portfolio argument:

strategy\cashflow	date t	date T
buy 1 stock	$-S_t$	S_T
sell $\frac{S_t}{d(t,T)}$ bonds	S_t	$-\frac{S_t}{d(t,T)}$
sell 1 forward	0	$F_t - S_T$
total cash flow	0	$F_t - \frac{S_t}{d(t,T)}$

Note that the cash flow at time T is known at time t and since the cash flow by definition of the forward price is equal to 0 at date t, the cash flow at

date T must be 0 as well. Hence

$$F_t = \frac{S_t}{d(t,T)}.$$

Note that buying a call and selling a put, both with exercise price K and expiration date T, is equivalent to buying forward at the price K. Therefore the convention that the forward contract has value 0 at date t is exactly equivalent to specifying K so that $c_t = p_t$.

6.3 Restrictions on option prices

In this section we derive some bounds on call prices which much be satisfied in frictionless markets. The line of reasoning used may of course be used on put options as well.

Consider a European call option with expiration date T and exercise price K. Assume that the underlying security does not pay any dividends during the life of the option. Then the value of the option c_t satisfies

$$S_t \geq c_t \geq \max\left(0, S_t - Kd(t,T)\right). \tag{6.1}$$

Proof. Clearly, $c_t \geq 0$. Also, the corresponding put option satisfies $p_t \geq 0$. Hence

$$c_t \geq c_t - p_t = S_t - Kd(t,T) \tag{6.2}$$

where we have used put-call parity. To see that $S_t \geq c_t$, assume that $S_t < c_t$ and consider the strategy of buying the stock and selling the option. That gives a positive cash flow at time t. If at time T, $S_T > K$ and the option is exercised the stock is delivered to the option holder and K is received. If the option is not exercised, the stock can be sold at non-negative value.

■

It is clear that an American option is more valuable than the corresponding European option, hence we note that the price C_t of an American option also satisfies $C_t \geq S_t - Kd(t,T)$. If interest rates are positive, i.e. $d(t,T) < 1$, this produces the interesting result that the value of the American call is always strictly greater than the immediate exercise value $S_t - K$ when $t < T$. This shows the important result that an American option on a non-dividend paying stock should never be exercised early. Our inequalities above show that it will be better to sell the option. A corresponding result does *not* hold for put options. This is perhaps not so surprising considering that postponing the exercise of a put postpones the receipt of K, whereas delaying the exercise of a call delays the payment of K.

Typically, stocks pay dividends and it is important to take this into account when pricing options. It will often be the case that the option contract does not take into account whether the underlying stock pays dividends. A dividend payment will normally produce a drop in the stock price and an owner of a call option will be hurt by this drop without receiving the benefit of a dividend. A date t is denoted an *ex-dividend* date if purchasing the stock at time $s < t$ gives the new owner part in the next dividend payment whereas a purchase at time t does not. For simplicity, we assume in the following that the dividend payment takes place at the ex-dividend date. Furthermore, we will assume that the size of the dividend is known some time before the dividend date. In a world with no taxes it ought to be the case then that the drop in the stock price around the dividend date is equal to the size of the dividend. Assume, for example, that the drop in the stock price is less than D. Then buying the stock right before the dividend date for a price of S_{t-} and selling it for S_{t+} immediately after the dividend date will produce a cash flow of $S_{t+} + D - S_{t-} > 0$. This resembles an arbitrage opportunity and it is our explanation for assuming in the following that $S_{t-} = S_{t+} + D$.

Now let us consider the price at time 0 of a European call option on a stock which is known to pay one dividend D at time t. Then

$$c_0 \geq \max\left(0, S_0 - Kd(0,T) - Dd(0,t)\right).$$

Again, $c_0 \geq 0$ is trivial. Assume $c_0 < S_0 - Kd(0,T) - Dd(0,t)$. Then buy the left hand side and sell the right hand side. At time t, we must pay dividend D on the stock we have sold, but that dividend is exactly received from the D zero coupon bonds with maturity t. At time T the value of the option we have sold is equal to $\max\left(0, S_T - K\right)$. The value of the right hand side is equal to $S_T - K$. If $S_T \geq K$ the total position is 0.If $S_T < K$ the total position has value $K - S_T$. Hence we have constructed a positive cash flow while also receiving money initially. This is an arbitrage opportunity and hence we rule out $c_0 < S_0 - Kd(0,T) - Dd(0,t)$.

There are many possible variations on the dividend theme. If dividends are not known at time 0 we may assume that they fall within a certain interval and then use the endpoints of this interval to bound calls and puts. The reader may verify that the maximal dividend is important for bounding calls and the minimum dividend for bounding put prices.

However, we maintain the assumption of a known dividend and finish this section by another important observation on the early exercise of American calls on dividend paying stocks. Assume that the stock pays a dividend at time t and that we are at time $0 < t$. It is then not optimal to exercise the option at time 0 whereas it may be optimal right before time t. To see that it

is not optimal at time 0, note that the American option contains as a part of its rights an option with expiration date $s \in (0, t)$, and since this option is an option on a non-dividend paying stock we know that its value is larger than $S_0 - K$, which is the value of immediate exercise. Therefore, the American option is also more worth than $S_0 - K$ and there is no point in exercising before t. To see that it may be optimal to exercise right before t, consider a firm which pays a liquidating dividend to all its shareholders. The stock will be worthless after the liquidation and so will the call option. Certainly, the option holder is better off to exercise right before the dividend date to receive part of the liquidating dividend.

The picture is much more complicated for puts. In the next section we will see how to compute prices for American puts in binomial models and this will give us the optimal exercise strategy as well.

6.4 Binomial models for stock options

In this section we will go through the binomial model for pricing stock options. Our primary focus is the case where the underlying security is a non-dividend paying stock but it should be transparent that the binomial framework is highly flexible and will easily handle the pricing and hedging of derivative securities with more complicated underlying securities.

We consider a model with T periods and assume throughout that the following two securities trade:

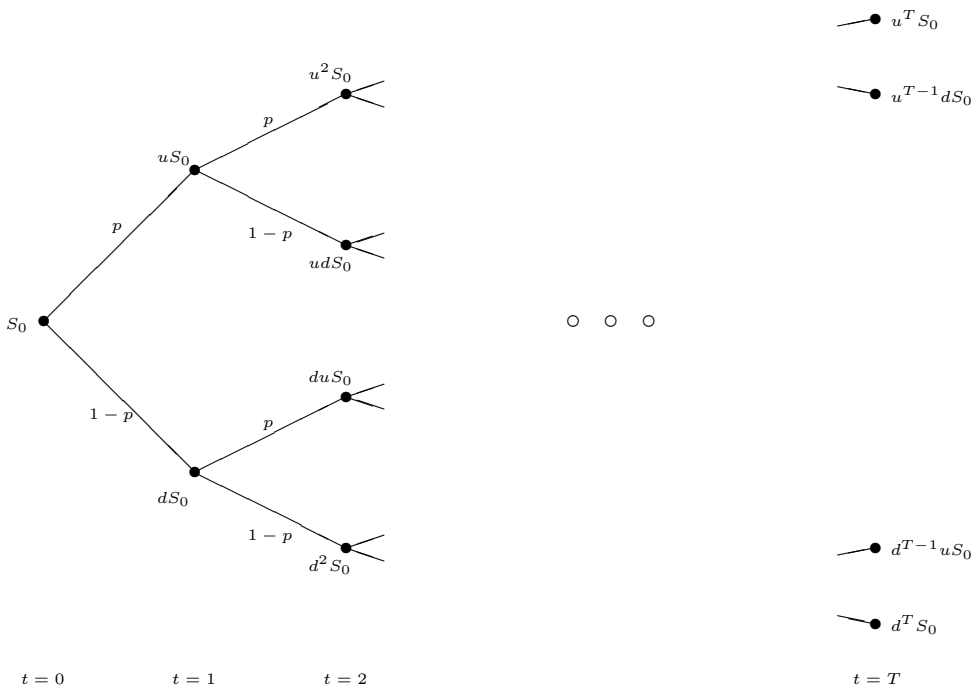
1. A money market account with a constant spot rate process ρ. Let $1 + \rho_t = R$, where $R \geq 1$. Hence we have for $s < t$

$$R_{s,t} = R^{t-s}.$$

2. A stock[2] S, which pays no dividends[3], whose price at time 0 is S_0 and whose evolution under the measure P is described in the tree (where we have assumed that $u > R > d > 0$) shown below.

[2]Since there is only one stock we will write S instead of S^1.

[3]To comply with the mathematical model of the previous chapter we should actually say that the stock pays a liquidating dividend of S_T at time T. We will however speak of S_T as the price at time T of the stock.



The mathematical description of the process is as follows: Let $U_1, \ldots, U_T$ be a sequence of i.i.d. Bernoulli variables, let $p = P(U_1 = 1)$ and define

$$N_t = \sum_{i=1}^{t} U_i.$$

Think of N_t as the number of up-jumps that the stock has had between time 0 and t. Clearly, this is a binomially distributed random variable. Let $u > R > d > 0$ be constants. Later, we will see how these parameters are chosen in practice. Then

$$S_t = S_0 u^{N_t} d^{t-N_t}. \tag{6.3}$$

Using the results on one-period submodels it is clear that the model is arbitrage free and complete and that the equivalent martingale measure is given in terms of conditional probabilities as

$$\begin{aligned} Q(S_t = uS_{t-1} | S_{t-1}) &\equiv q = \frac{R-d}{u-d} \\ Q(S_t = dS_{t-1} | S_{t-1}) &= 1-q = \frac{u-R}{u-d}. \end{aligned}$$

6.5 Pricing the European call

We now have the martingale measure Q in place and hence the value at time t of a European call with maturity T is given in an arbitrage-free model by

$$C_t = \frac{1}{R^{T-t}} E^Q \left(\max(0, S_T - K) \, | \mathcal{F}_t \right).$$

Using this fact we get the following

Proposition 22 *Let the stock and money market account be as described in section 6.4. Then the price of a European call option with exercise price K and maturity date T is given as*

$$C_t = \frac{1}{R^{T-t}} \sum_{i=0}^{T-t} \binom{T-t}{i} q^i (1-q)^{T-t-i} \max(0, S_t u^i d^{T-t-i} - K).$$

Proof. Since the money market account and S_0 are deterministic, we have that we get all information by observing just stock-prices, or equivalently the U's, i.e. $\mathcal{F}_t = \sigma(S_1, \ldots, S_t) = \sigma(U_1, \ldots, U_t)$. By using (6.3) twice we can write

$$S_T = S_t u^{(N_T - N_t)} d^{(T-t)-(N_T-N_t)} = S_t u^Z d^{(T-t)-Z},$$

where $Z = N_T - N_t = \sum_{j=t+1}^{T} U_j \overset{Q}{\sim} \text{bi}(q; (T-t))$, and Z is independent of $\mathcal{F}_t$ (because the U's are independent). Therefore

$$R^{T-t} C_t = E^Q((S_T - K)^+ | \mathcal{F}_t) = E^Q((S_t u^Z d^{(T-t)-Z} - K)^+ | \mathcal{F}_t).$$

At this point in the narrative we need something called "the useful rule". It states the following: Suppose we are given a function $f : \mathbb{R}^2 \mapsto \mathbb{R}$, a σ-algebra $\mathcal{F}$, an $\mathcal{F}$-measurable random variable X and a random variable Y that is independent of $\mathcal{F}$. Define the function $g : \mathbb{R} \mapsto \mathbb{R}$ by $g(x) = E(f(x, Y))$. Then $E(f(X, Y)|\mathcal{F}) = g(X)$. We then use this in the above expression with S_t playing the role of X, Z as Y, and $f(x, y) = (x u^y d^{(T-t)-y} - K)^+$. By using the general transformation rule for discrete random variables $E(h(Y)) = \sum_{y_i} h(y_i) P(Y = y_i)$, and the fact that Z is Q-binomially distributed we get in the notation of "the useful rule" that

$$g(x) = \sum_{i=0}^{T-t} \left(\begin{array}{c} T-t \\ i \end{array} \right) q^i (1-q)^{(T-t)-i} (x u^i d^{(T-t)-i} - K)^+,$$

and the desired result follows. ■

We rewrite the expression for C_0 using some handy notation. Let a be the smallest number of upward jumps needed for the option to finish in the money, i.e.

$$\begin{aligned} a &= \min_{j\in\mathbf{N}}\{j|S_0u^jd^{T-j} > K\} \\ &= \min_{j\in\mathbf{N}}\{j|j\ln u + (T-j)\ln d > \ln(K/S_0)\} \\ &= \min_{j\in\mathbf{N}}\{j|j > \ln(K/(S_0d^T))/\ln(u/d)\} \\ &= \left\lfloor \frac{\ln\left(\frac{K}{S_0d^T}\right)}{\ln\left(\frac{u}{d}\right)} \right\rfloor + 1. \end{aligned}$$

Letting

$$\Psi(a;T,q) = \sum_{i=a}^{T} \left(\begin{array}{c} T \\ i \end{array} \right) q^i(1-q)^{T-i},$$

we may write (you may want to check the first term on the RHS)

$$C_0 = S_0\Psi\left(a;T,q'\right) - \frac{K}{R^T}\Psi\left(a;T,q\right) \tag{6.4}$$

where

$$q' = \frac{u}{R}q.$$

Using put-call parity gives us the price of the European put:

Corollary 23 *The price of a European put option with T periods to maturity, exercise price T and the stocks as underlying security has a price at time 0 given by*

$$P_0 = \frac{K}{R^T}\left(1-\Psi\left(a;T,q\right)\right) - S_0\left(1-\Psi\left(a;T,q'\right)\right)$$

Note that our option pricing formulae use T to denote the number of periods until maturity. Later, we will be more explicit in relating this to actual calendar time.

6.6 Hedging the European call

We have already seen in a two period model how the trading strategy replicating a European call option may be constructed. In this section we simply state the result for the case with T periods and we then note an interesting

way of expressing the result. We consider the case with a money market account and one risky asset S and assume that the market is complete and arbitrage-free. The European call option has a payout at maturity of

$$\delta_T^c = \max(S_T - K, 0).$$

Proposition 24 *A self-financing trading strategy replicating the dividend process of the option from time* 1 *to* T *is constructed recursively as follows: Find* $\phi_{T-1} = (\phi_{T-1}^0, \phi_{T-1}^1)$ *such that*

$$\phi_{T-1}^0 R + \phi_{T-1}^1 S_T = \delta_T^c.$$

For $t = T-2, T-3, \ldots, 1$ *find* $\phi_t = (\phi_t^0, \phi_t^1)$ *such that*

$$\phi_t^0 R + \phi_t^1 S_{t+1} = \phi_{t+1}^0 + \phi_{t+1}^1 S_{t+1}.$$

The trading strategy is self-financing by definition, replicates the call and its initial price of $\phi_0^0 + \phi_0^1 S_0$ is equal to the arbitrage-free price of the option. We may easily extend to the case where both the underlying and the contingent claim have dividends other than the one dividend of the option considered above. In that case the procedure is the following:Find $\phi_{T-1} = (\phi_{T-1}^0, \phi_{T-1}^1)$ such that

$$\phi_{T-1}^0 R + \phi_{T-1}^1 (S_T + \delta_T) = \delta_T^c.$$

For $t = T-2, T-3, \ldots, 1$ find $\phi_t = (\phi_t^0, \phi_t^1)$ such that

$$\phi_t^0 R + \phi_t^1 (S_{t+1} + \delta_{t+1}) = \phi_{t+1}^0 + \phi_{t+1}^1 S_{t+1} + \delta_{t+1}^c.$$

In this case the trading strategy is not self-financing in general but it matches the dividend process of the contingent claim, and the initial price of the contingent claim is still $\phi_0^0 + \phi_0^1 S_0$.

An additional insight into the hedging strategy is given by the proposition below.

Recall the notation

$$\widetilde{S}_t = \frac{S_t}{R_{0,t}}$$

for the discounted price process of the stock. Let C_t denote the price process of a contingent claim whose dividend process is δ^c and let

$$\begin{aligned} \widetilde{C}_t &= \frac{C_t}{R_{0,t}} \\ \widetilde{\delta}_t^c &= \frac{\delta_t^c}{R_{0,t}} \end{aligned}$$

denote the discounted price and dividend processes of the contingent claim. Define the conditional covariance under the martingale measure Q as follows:

$$\mathrm{Cov}^Q\left(X_{t+1}, Y_{t+1}\,|\mathcal{F}_t\right) = E^Q\left(\left(X_{t+1}-X_t\right)\left(Y_{t+1}-Y_t\right)|\,\mathcal{F}_t\right)$$

One may then show the following (but we will omit the proof):

Proposition 25 *Assume that the stock pays no dividends during the life of the option. The hedging strategy which replicates δ^c is computed as follows:*

$$\phi_t^1 = \frac{\mathrm{Cov}^Q\left(\widetilde{S}_{t+1}, \widetilde{C}_{t+1}+\widetilde{\delta}^c_{t+1}\Big|\,\mathcal{F}_t\right)}{\mathrm{VAR}^Q\left(\widetilde{S}_{t+1}\,|\mathcal{F}_t\right)} \qquad t=0,1,\ldots,T-1$$

$$\phi_t^0 = \widetilde{C}_t - \phi_t^1\widetilde{S}_t \qquad t=0,1,\ldots,T-1$$

Note the similarity with regression analysis! We will not go further into this at this stage. But this way of looking at hedging is important when defining so-called risk minimal trading strategies in incomplete markets.

The number of stocks held at time t in the replicating strategy is called the *hedge ratio.* The hedge ratio for a call option is a number between 0 and 1, and it is larger the more in-the-money the call is.

6.7 Recombining tree representation

If the number of time periods T is large it the tree representing the stock price evolution grows very rapidly. The number of nodes at time t is equal to 2^t, and since for example $2^{20} = 1048576$ we see that when you implement this model in a spreadsheet and you wish to follow C_t and the associated hedging strategy over time, you may soon run out of space. Fortunately, in many cases there is a way around this problem: If your security price process is *Markov* and the contingent claim you wish to price is *path-independent*, you can use a *recombining* tree to do all of your calculations. Let us look at each property in turn[4]: The process S is a *Markov chain* under Q if it satisfies

$$Q(S_{t+1}=s_{t+1}\,|S_t=s_t,\ldots S_1=s_1,S_0=s_0) = Q(S_{t+1}=s_{t+1}\,|S_t=s_t)$$

for all t and all $(s_{t+1},s_t,\ldots,s_1,s_0)$. Intuitively, standing at time t, the current value of the process s_t is sufficient for describing the distribution of the

[4]These properties are interesting to consider for the stock only since the money market account trivially has all nice properties discussed in the following.

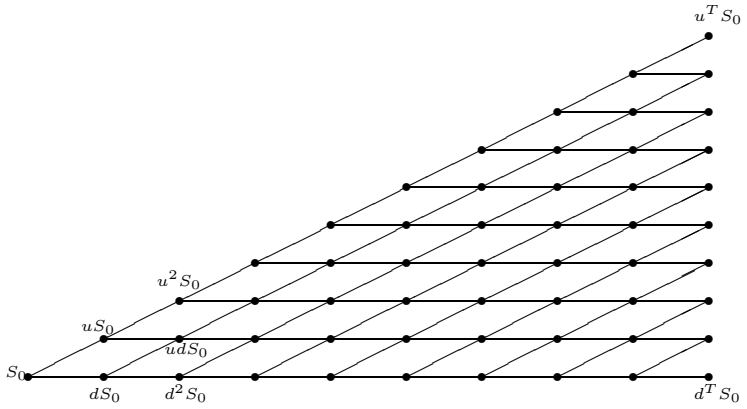


Figure 6.1: A lattice, i.e. a recombining tree.

process at time $t+1$. The binomial model of this chapter is clearly a Markov chain. An important consequence of this is that when $\mathcal{F}_t = \sigma(S_0, \ldots, S_t)$ then for any (measurable) function f and time points $t < u$ there exists a function g such that

$$E^Q\left(f(S_u)\,|\mathcal{F}_t\right) = g(S_t). \tag{6.5}$$

In other words, conditional expectations of functions of future values given everything we know at time t can be expressed as a function of the value of S_t at time t. The way S arrived at S_t is not important. We used this fact in the formula for the price of the European call: There, the conditional expectation given time t information became a function of S_t. The past did not enter into the formula. We can therefore represent the behavior of the process S in a *recombining tree*, also known as a *lattice*, as shown in Figure 6.1 in which one node at time t represents exactly one value of S_t. Another way of stating this is to say that the tree keeps track of the number of up-jumps that have occurred, not the order in which they occurred. A full event tree would keep track of the exact timing of the up-jumps.

To see what can go wrong, Figure 6.2 shows a process that is not Markov.

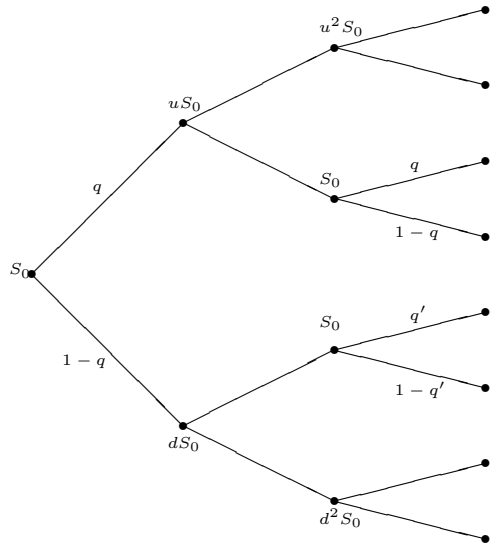


Figure 6.2: A tree that's not a lattice.

The problem is at time 2 when the value of the process is S_0, we need to know the pre-history of S to decide whether the probability of going up to uS is equal to q or q'. In standard binomial models such behavior is normally precluded.

Note that now the number of nodes required at time t is only $t+1$, and then using several hundred time periods is no problem for a spreadsheet.

A technical issue which we will not address here is the following: Normally we specify the process under the measure P, and it need not be the case that the Markov property is preserved under a change of measure. However, one may show that if the price process is Markov under P and the model is complete and arbitrage-free, then the price process is Markov under the equivalent martingale measure Q as well.

A second condition for using a recombining tree to price a contingent claim is a condition on the contingent claim itself:

Definition 37 *A contingent claim with dividend process δ^c is path independent if $\delta_t = f_t(S_t)$ for some (measurable) function f.*

Indeed if the claim is path independent and the underlying process is Markov, we have

$$C_t = R_{0,t}E\left(\sum_{i=t+1}^{T}\widetilde{\delta}_i^c \middle| \mathcal{F}_t\right)$$

$$= R_{0,t}E\left(\sum_{i=t+1}^{T} f_i(S_i)\middle|\mathcal{F}_t\right)$$
$$= R_{0,t}E\left(\sum_{i=t+1}^{T} f_i(S_i)\middle|S_t\right)$$

and the last expression is a function of S_t by the Markov property. A European option with expiration date T is path-independent since its only dividend payment is at time T and is given as $\max(S_T - K, 0)$.

The *Asian option* is an example of a contingent claim which is *not* path-independent. An Asian option on the stock, initiated at time 0, expiration date T and exercise price K has a payoff at date T given by

$$C_T^{asian} = \max\left(0, \left(\frac{1}{T+1}\sum_{t=0}^{T} S_t\right) - K\right)$$

Hence the average of the stock price over the period determines the option price. Clearly, S_T is not sufficient to describe the value of the Asian option at maturity. To compute the average value one needs the whole path of S. As noted above, even in a binomial model keeping track of the whole path for, say, 50 periods becomes intractable.

6.8 The binomial model for American puts

We describe in this section a simple way of pricing the American put option in a binomial model. Strictly speaking, an American put is not a contingent claim in the sense we have thought of contingent claims earlier. Generally, we have thought of contingent claims as random variables or sometimes as processes but a put is actually not specified until an exercise policy is associated with the put. What we will do in the following is to simultaneously solve for the optimal exercise policy, i.e. the one that maximizes the expected, discounted value of the cash flows under the martingale measure, and the price of the option. The argument given is not a proof but should be enough to convince the reader that the right solution is obtained (it is fairly easy to show that another exercise policy will create arbitrage opportunities for the option writer).

The value of an American put at its maturity is easy enough:[5]

$$P_T = \max(0, K - S_T). \tag{6.6}$$

[5]Or is it? As it stands P_t is really the value at time t given that the put has not been exercised at times $0, 1, t-1$. But that will most often be exactly what we are interested in; if we exercised the put to years ago, we really don't care about it anymore.

Now consider the situation one period before maturity. If the put has not been exercised at that date, the put option holder has two possibilities: Exercise the put at time $T-1$ or hold the put to maturity. The value of holding the put to maturity is given as the discounted (back to time $T-1$) value of (6.6), whereas the value at time $T-1$ of exercising immediately is $K - S_{T-1}$ something only to be considered of course if $K > S_{T-1}$. Clearly, the put option holder has a contract whose value is given by the maximal value of these two strategies, i.e.[6]

$$P_{T-1} = \max\left(K - S_{T-1}, E^Q\left(\frac{P_T}{R}\middle|\mathcal{F}_{T-1}\right)\right).$$

Now continue in this fashion by working backwards through the tree to obtain the price process of the American put option given by the recursion

$$P_{t-1} = \max\left(K - S_{t-1}, E^Q\left(\frac{P_t}{R}\middle|\mathcal{F}_{t-1}\right)\right) \qquad t = 1, \ldots, T.$$

Once this price process is given we see that the optimal exercise strategy is to exercise the put the first time t for which

$$K - S_t > E^Q\left(\frac{P_{t+1}}{R}\middle|\mathcal{F}_t\right).$$

This way of thinking is easily translated to American call options on dividend paying stocks for which early exercise is something to consider.

6.9 Implied volatility

We assume in this section that the Black-Scholes formula is known to the reader: The price at time t of a European call option maturing at time T, when the exercise price is K and the underlying security is a non-dividend paying stock with a price of S_t, is given in the Black-Scholes framework by

$$C_t = S_t\Phi(d_1) - Ke^{-r(T-t)}\Phi(d_2)$$

where

$$d_1 = \frac{\log\left(\frac{S_t}{K}\right) + \left(r + \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}}$$

and

$$d_2 = d_1 - \sigma\sqrt{T-t}$$

[6] We do not need 0 in the list of arguments of max since positivity is assured by $P_T \geq 0$.

where Φ is the cumulative distribution function of a standard normal distribution.

Consider the Black-Scholes formula for the price of a European call on an underlying security whose value at time 0 is S_0: Recall that Φ is a distribution function, hence $\Phi(x) \to 1$ as $x \to \infty$ and $\Phi(x) \to 0$ as $x \to -\infty$. Assume throughout that $T > 0$. From this it is easy to see that $c_0 \to S_0$ as $\sigma \to \infty$. By considering the cases $S_0 < K\exp(-rT)$, $S_0 = K\exp(-rT)$ and $S_0 > K\exp(-rT)$ separately, it is easy to see that as $\sigma \to 0$, we have $c_0 \to \max(0, S_0 - K\exp(-rT))$. By differentiating c_0 with respect to σ, one may verify that c_0 is strictly increasing in σ. Therefore, the following definition makes sense:

Definition 38 *Given a security with price S_0. Assume that the risk free rate (i.e. the rate of the money market account) is equal to r. Assume that the price of a call option on the security with exercise price K and time to maturity T is observed to have a price of c^{obs} with*

$$\max(0, S_0 - K\exp(-rT)) < c^{obs} < S_0.$$

Then the implied volatility of the option is the unique value of σ for which

$$c_0(S_0, K, T, \sigma, r) = c^{obs}.$$

In other words, the implied volatility is the unique value of the volatility which makes the Black-Scholes model 'fit' c^{obs}. Clearly, we may also associate an implied volatility to a put option whose observed price respects the appropriate arbitrage bounds.

A very important reason for the popularity of implied volatility is the way in which it allows a transformation of option prices which are hard to compare into a common scale. Assume that the price of a stock is 100 and the riskfree rate is 0.1. If one observed a price of 9.58 on a call option on the stock with exercise price 100 and 6 months to maturity and a price of 2.81 on a put option on the stock with exercise price 95 and 3 months to maturity then it would require a very good knowledge of the Black-Scholes model to see if one price was in some way higher than the other. However, if we are told that the implied volatility of the call is 0.25 and the implied volatility of the put is 0.30, then at least we know that compared to the Black-Scholes model, the put is more expensive than the call. This way of comparing is in fact so popular that traders in option markets typically do not quote prices in (say) dollars, but use 'vols' instead.

If the Black-Scholes model were true the implied volatility of all options written on the same underlying security should be the same, namely equal

to the volatility of the stock and this volatility would be a quantity we could estimate from historical data. In short, in a world where the Black-Scholes model holds, historical volatility (of the stock) is equal to implied volatility (of options written on the stock). In practice this is not the case - after all the Black-Scholes model is only a model. The expenses of hedging an option depend on the volatility of the stock during the life of the option. If, for example, it is known that, after a long and quiet period, important news about the underlying stock will arrive during the life of the option, the option price should reflect the fact that future fluctuations in the stock price might be bigger than the historical ones. In this case the implied volatility would be higher than the historical.

However, taking this knowledge of future volatility into account one could still imagine that all implied volatilities of options on the same underlying were the same (and equal to the 'anticipated' volatility). In practice this is not observed either. To get an idea of why, we consider the notion of portfolio insurance.

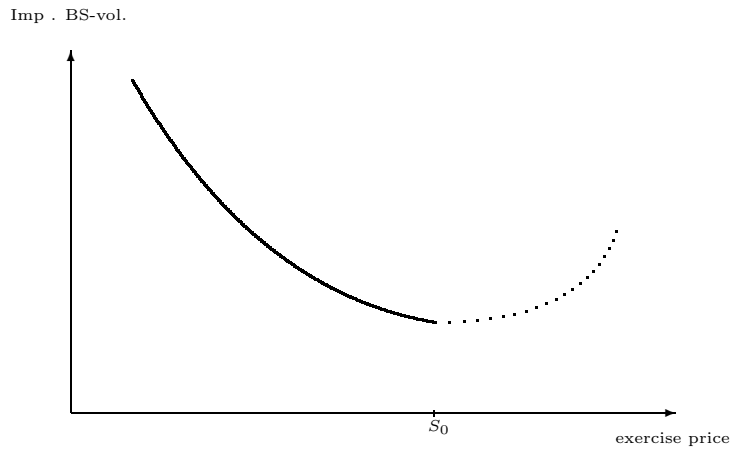
6.10 Portfolio insurance, implied volatility and crash fears

Consider a portfolio manager who manages a portfolio which is diversified so that the value of her portfolio follows that of the market stock index. Assume that the value of her portfolio is 1000 times the value of the index which is assumed to be at 110. The portfolio manager is very worried about losing a large portion of the value of the portfolio over the next year - she thinks that there is a distinct possibility that the market will crash. On the other hand she is far from certain. If she were certain, she could just move the money to a bank at a lower but safer expected return than in the stock market. But she does not want to exclude herself from the gains that a surge in the index would bring. She therefore decides to buy *portfolio insurance* in such a way that the value of her portfolio will never fall below a level of (say) 90.000. More specifically, she decides to buy 1000 put options with one year to maturity and an exercise price of 90 on the underlying index. Now consider the value of the portfolio after a year as a function of the level of the index S_T :

value of index	$S_T \geq 90$	$S_T < 90$
value of stocks	$S_T \times 1000$	$S_T \times 1000$
value of puts	0	$1000 \times (90 - S_T)$
total value	$S_T \times 1000 > 90.000$	90.000

Although it has of course not been costless to buy put options, the portfolio manager has succeeded in preventing the value of her portfolio from falling below 90.000. Since the put options are far out-of-the-money (such contracts are often called "lottery tickets") at the time of purchase they are probably not that expensive. And if the market booms she will still be a successful portfolio manager.

But what if she is not alone with her fear of crashes. We may then imagine a lot of portfolio managers interested in buying out-of-the-money put options hence pushing up the price of these contracts. This is equivalent to saying that the implied volatility goes up and we may experience the scenario shown in the graph below, in which the implied volatility of put options is higher for low exercise price puts:



This phenomenon is called a "smirk". If (as it is often seen from data) the implied volatility is increasing (the dotted part of the curve) for puts that are in the money, then we have what is known as a "smile". Actually options that are deeply in-the-money are rarely traded, so the implied volatility figures used to draw "the other half" of the smile typically comes from out-of-the-money calls. (Why/how? Recall the put-call parity.)

A smirk has been observed before crashes and it is indicative of a situation where the Black-Scholes model is not a good model to use. The typical modification allows for stock prices to jump discontinuously but you will have to wait for future courses to learn about this.

6.11 Debt and equity as options on firm value

In this section we consider a very important application of option pricing. Our goal is to learn a somewhat simplified but extremely useful way of think-

ing about a firm which is financed by debt and equity (see below). A fundamental assumption in this section is that a firm has a market value given by a stochastic process V. In Arrow-Debreu economies in which we know prices and production plans adopted by the firms, it is easy to define the value of a firm as the (net) value of its production. In reality things are of course a lot more complicated. It is hard to know, for example, what the value of Novo-Nordisk is - i.e. what is the market value of the firm's assets (including know-how, goodwill etc.). Part of the problem is of course that it is extremely difficult to model future prices and production levels. But in a sense the actual value does not matter for this section in that the 'sign' of the results that we derive does not depend on what the value of the firm is - only the "magnitude' does.

The fundamental simplification concerns the *capital structure* of the firm. Assume that the firm has raised capital to finance its activities in two ways: It has issued stocks (also referred to as *equity)* and *debt.* The debt consists of zero coupon bonds with face value D maturing at time T. Legally what distinguishes the debt holders from the stock holders is the following: The stock holders control the firm and they decide at time T whether the firm should repay its debt to the bondholders. If the bondholders are not repaid in full they can force the firm into bankruptcy and take over the remaining assets of the firm (which means both controlling and owning it). The stocks will then be worthless. If the stockholders pay back D at maturity to the bondholders, they own the firm entirely. They may then of course decide to issue new debt to finance new projects but we will not worry about that now.

It is clear that the stockholders will have an interest in repaying the bondholders precisely when $V_T > D$. Only then will the expense in paying back the debt be more than outweighed by the value of the firm. If $V_T < D$ (and there are no bankruptcy costs) the stockholders will default on their debt, the firm will go into bankruptcy and the bondholders will take over. In short, we may write the value of debt and equity at time T as

$$\begin{aligned} B_T &= \min(D, V_T) = D - \max(D - V_T, 0) \\ S_T &= \max(V_T - D, 0). \end{aligned}$$

In other words, we may think of equity as a call option on the value of the firm and debt as a zero coupon bond minus a put option on the value of the firm. Assuming then that V behaves like the underlying security in the Black-Scholes model and that there exists a money market account with interest rate r, we can use the Black-Scholes model to price debt and equity at time 0 :

$$B_0 = D\exp(-rT) - p_0(V_0, D, T, \sigma, r)$$

$$S_0 = c_0(V_0, D, T, \sigma, r)$$

where p_0, c_0 are Black-Scholes put and call functions.

Let us illustrate a potential conflict between stockholders and bondholders in this model. Assume that at time 0 the firm has the possibility of adopting a project which will not alter the value of the firm at time 0, but which will have the effect of increasing the volatility of the process V. Since both the value of the call and the put increases when σ increases we see that the stockholders will like this project since it increases the value of the equity whereas the bondholders will not like the project since the put option which they have in a sense written will be a greater liability to them. This is a very clear and very important illustration of so-called *asset substitution,* a source of conflict which exists between stock-and bondholders of a firm. This setup of analyzing the value of debt and equity is useful in a number of contexts and you should make sure that you understand it completely. We will return to this towards the end of the course when discussing corporate finance.

Chapter 7

The Black-Scholes formula

7.1 Black-Scholes as a limit of binomial models

So far we have not specified the parameters p, u, d and R which are of course critical for the option pricing model. Also, it seems reasonable that if we want the binomial model to be a realistic model for stock prices over a certain interval of time we should use a binomial model which divides the (calendar) time interval into many sub-periods. In this chapter we will first show that if one divides the interval into finer and finer periods and choose the parameters carefully, the value of the option converges to a limiting formula, the Black-Scholes formula, which was originally derived in a continuous time framework. We then describe that framework and show how to derive the formula in it.

Our starting point is an observed stock price whose logarithmic return satisfies

$$E^P\left[\ln\left(\frac{S_t}{S_{t-1}}\right)\right] = \mu$$

and

$$V^P\left(\ln\left(\frac{S_t}{S_{t-1}}\right)\right) = \sigma^2,$$

where S_t is the price of the stock t *years* after the starting date 0. Also, assume that the money market account has a continuously compounded return of r, i.e. an amount of 1 placed in the money market account grows to $\exp(r)$ in one year. Note that since $R^T = \exp(T\ln(R))$, a yearly rate of $R = 1.1$ (corresponding to a yearly rate of 10%) translates into the continuous compounding analogue $r = \ln(1.1)$ and this will be a number smaller than 0.1.

Consider pricing an option on this stock with time to maturity T years in a binomial model. Divide each year into n periods. This gives a binomial model with nT periods. In this tree, which we label the nth tree, choose

$$\begin{aligned} u_n &= \exp\left(\sigma\sqrt{\frac{1}{n}}\right), \\ d_n &= \exp\left(-\sigma\sqrt{\frac{1}{n}}\right) = \frac{1}{u_n}, \\ R_n &= \exp\left(\frac{r}{n}\right), \end{aligned}$$

and

$$p_n = \frac{1}{2} + \frac{1}{2}\frac{\mu}{\sigma}\sqrt{\frac{1}{n}}.$$

With the setup in the nth model specified above you may show by simple computation that the one-year logarithmic return satisfies

$$E^P\left[\ln\left(\frac{S_1}{S_0}\right)\right] = n\left\{p_n \ln(u_n) + (1-p_n)\ln(d_n)\right\} = \mu$$

and

$$V^P\left(\ln\left(\frac{S_1}{S_0}\right)\right) = \sigma^2 - \frac{1}{n}\mu^2,$$

so the log-return of the price process has the same mean and almost the same variance as the process we have observed. And since

$$V^P\left(\ln\left(\frac{S_1}{S_0}\right)\right) \to \sigma^2 \text{ for } n \to \infty,$$

it is presumably so that large values of n brings us closer to to "desired" model.

The above story was primarily motivational. Let us now investigate precisely what happens to stock and call prices when n tends to infinity. For each n we may compute the price of a call option with maturity T in the binomial model and we know that it is given as

$$C^n = S_0\Psi\left(a_n; nT; q_n^{'}\right) - \frac{K}{(R_n)^T}\Psi\left(a_n; nT; q_n\right) \tag{7.1}$$

where

$$q_n = \frac{R_n - d_n}{u_n - d_n}, \quad q_n' = \frac{u_n}{R_n}q_n$$

and a_n is the smallest integer larger than $\ln\left(K/(S_0 d_n^{Tn})\right)/\ln(u_n/d_n)$. Note that alternatively we may write(7.1) as

$$C^n = S_0 Q'(S_n(T) > K) - Ke^{-rT} Q(S_n(T) > K) \tag{7.2}$$

where $S_n(T) = S_0 u_n^j d_n^{Tn-j}$ and $j \overset{Q}{\sim} \text{bi}(Tn, q_n)$ and $j \overset{Q'}{\sim} \text{bi}(Tn, q_n')$. It is easy to see that

$$\begin{aligned} M_n^Q &:= E^Q(\ln S_n(T)) = \ln S_0 + Tn(q_n \ln u_n + (1-q_n)\ln d_n) \\ V_n^Q &:= V^Q(\ln S_n(T)) = Tnq_n(1-q_n)(\ln u_n - \ln d_n)^2, \end{aligned}$$

and that similar expressions (with q_n' instead of q_n) hold for Q'-moments. Now rewrite the expression for M_n^Q in the following way:

$$\begin{aligned} M_n^Q - \ln S_0 &= Tn\left(\frac{\sigma}{\sqrt{n}}\frac{e^{r/n} - e^{-\sigma/\sqrt{n}}}{e^{\sigma/\sqrt{n}} - e^{-\sigma/\sqrt{n}}} - \frac{\sigma}{\sqrt{n}}\frac{e^{\sigma/\sqrt{n}} - e^{r/n}}{e^{\sigma/\sqrt{n}} - e^{-\sigma/\sqrt{n}}}\right) \\ &= T\sqrt{n}\sigma\left(\frac{2e^{r/n} - e^{\sigma/\sqrt{n}} - e^{-\sigma/\sqrt{n}}}{e^{\sigma/\sqrt{n}} - e^{-\sigma/\sqrt{n}}}\right). \end{aligned}$$

Recall the Taylor-expansion to the second order for the exponential function: $\exp(\pm x) = 1 \pm x + x^2/2 + \text{o}(x^2)$. From this we get

$$\begin{aligned} e^{r/n} &= 1 + r/n + \text{o}(1/n) \\ e^{\pm\sigma/\sqrt{n}} &= 1 \pm \sigma/\sqrt{n} + \sigma^2/(2n) + \text{o}(1/n). \end{aligned}$$

Inserting this in the M_n^Q expression yields

$$\begin{aligned} M_n^Q - \ln S_0 &= T\sqrt{n}\sigma\left(\frac{2r/n - \sigma^2/n + \text{o}(1/n)}{2\sigma/\sqrt{n} + \text{o}(1/n)}\right) \\ &= T\sigma\left(\frac{2r - \sigma^2 + \text{o}(1)}{2\sigma + \text{o}(1/\sqrt{n})}\right) \\ &\to T\left(r - \frac{\sigma^2}{2}\right) \quad \text{for } n \to \infty. \end{aligned}$$

Similar Taylor expansions for V_n^Q, $M_n^{Q'}$ and $V_n^{Q'}$ show that

$$\begin{aligned} V_n^Q &\to \sigma^2 T, \\ M_n^{Q'} - \ln S_0 &\to T\left(r + \frac{\sigma^2}{2}\right) \quad \text{(note the change of sign on } \sigma^2\text{)}, \\ V_n^{Q'} &\to \sigma^2 T. \end{aligned}$$

So now we know what the Q/Q' moments converge to. Yet another way to think of $\ln S_n(T)$ is as a sum of Tn independent Bernoulli-variables with possible outcomes $(\ln d_n, \ln u_n)$ and probability parameter q_n (or q'_n). This means that we have a sum of (well-behaved) independent random variables for which the first and second moments converge. Therefore we can use a version of the Central Limit Theorem[1] to conclude that the limit of the sum is normally distributed, i.e.

$$\ln S_n(T) \overset{Q/Q'}{\to} N(\ln S_0 + (r \pm \sigma^2/2)T, \sigma^2 T).$$

This means (almost by definition of the form of convergence implied by CLT) that when determining the limit of the probabilities on the right hand side of (7.2) we can (or: have to) substitute $\ln S_n(T)$ by a random variable X such that

$$X \overset{Q/Q'}{\sim} N(\ln S_0 + (r \pm \sigma^2/2)T, \sigma^2 T) \Leftrightarrow \frac{X - \ln S_0 - (r \pm \sigma^2/2)T}{\sigma\sqrt{T}} \overset{Q/Q'}{\sim} N(0,1).$$

The final analysis:

$$\begin{aligned}
\lim_{n\to\infty} C^n &= \lim_{n\to\infty} \left(S_0 Q'(\ln S_n(T) > \ln K) - K e^{-rT} Q(\ln S_n(T) > \ln K) \right) \\
&= S_0 Q'(X > \ln K) - K e^{-rT} Q(X > \ln K) \\
&= S_0 Q' \left(\frac{X - \ln S_0 - (r + \sigma^2/2)T}{\sigma\sqrt{T}} > \frac{\ln K - \ln S_0 - (r + \sigma^2/2)T}{\sigma\sqrt{T}} \right) \\
&\quad -K e^{-rT} Q \left(\frac{X - \ln S_0 - (r - \sigma^2/2)T}{\sigma\sqrt{T}} > \frac{\ln K - \ln S_0 - (r - \sigma^2/2)T}{\sigma\sqrt{T}} \right)
\end{aligned}$$

Now multiply by -1 inside the Q's (hence reversing the inequalities), use that the $N(0,1)$-variables on the left hand sides are symmetric and continuous, and that $\ln(x/y) = \ln x - \ln y$. This shows that

$$\lim_{n\to\infty} C^n = S_0 \Phi(d_1) - K e^{-rT} \Phi(d_2),$$

where Φ is the standard normal distribution function and

$$\begin{aligned}
d_1 &= \frac{\ln\left(\frac{S_0}{K}\right) + \left(r + \frac{1}{2}\sigma^2\right) T}{\sigma\sqrt{T}}, \\
d_2 &= \frac{\ln\left(\frac{S_0}{K}\right) + \left(r - \frac{1}{2}\sigma^2\right) T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T}.
\end{aligned}$$

[1] Actually you cannot quite make do with the De Moivre-version that you know from Stat 0 because we do not have a scaled sum of *identically* distributed random variables. You need the notion of a triangular array and the Lindeberg-Feller-version of the Central Limit Theorem. Yet another reason to take Stat 2b.

This formula for the call price is called the Black-Scholes formula. So far we can see it just as an artifact of going to the limit in a particular way in a binomial model. But the formula is so strikingly beautiful and simple that there must be more to it than that. In particular, we are interested in the question: Does that exist a "limiting" model in which the above formula is the exact call option price? The answer is: Yes. In the next section we describe what this "limiting" model looks like, and show that the Black-Scholes formula gives the exact call price in the model. That does involve a number of concepts, objects and results that we cannot possibly make rigorous in this course, but the reader should still get a "net benefit" and hopefully an appetite for future courses in financial mathematics.

7.2 The Black-Scholes model

The Black-Scholes formula for the price of a call option on a non-dividend paying stock is one of the most celebrated results in financial economics. In this chapter we will indicate how the formula is derived. A rigorous derivation requires some fairly advanced mathematics which is beyond the scope of this course. Fortunately, the formula is easy to interpret and to apply. Even if there are some technical details left over for a future course, the rigorous understanding we have from our discrete-time models of how arbitrage pricing works will allow us to apply the formula safely.

The formula is formulated in a continuous time framework with random variables that have continuous distribution. The continuous-time and infinite state space setup will not be used elsewhere in the course.[2] But let us mention that if one wants to develop a theory which allows random variables with continuous distribution and if one wants to obtain results similar to those of the previous chapters, then one has to allow continuous trading as well. By 'continuous trading' we mean that agents are allowed to readjust portfolios continuously through time.

If X is normally distributed $X \sim N(\alpha, \sigma^2)$, then we say that $Y := \exp(X)$ is *lognormally* distributed and write $Y \sim LN(\alpha, \sigma^2)$. There is one thing you must always remember about lognormal distributions:

$$\text{If } Y \sim LN(\alpha, \sigma^2) \text{ then } E(Y) = \exp\left(\alpha + \frac{\sigma^2}{2}\right).$$

[2] A setup which combines *discrete time* and continuous distributions will be encountered later when discussing CAPM and APT, but the primary focus of these models will be to explain stock price behavior and not – as we are now doing – determining option prices for a given behavior of stock prices

If you have not seen this before, then you are strongly urged to check it. (With that result you should also be able to see why there is no need to use "brain RAM" remembering the variance of a lognormally distributed variable.) Often the lognormal distribution is preferred as a model for stock price distributions since it conforms better with the institutional fact that prices of a stock are non-negative and the empirical observation that the logarithm of stock prices seem to show a better fit to a normal distribution than do prices themselves. However, specifying a distribution of the stock price at time t, say, is not enough. We need to specify the whole process of stock prices, i.e. we need to state what the joint distribution $(S_{t_1}, \ldots, S_{t_N})$ is for any $0 \leq t_1 < \ldots < t_N$. To do this the following object is central.

Definition 39 *A (standard) Brownian motion ((S)BM) is a stochastic process $B = (B_t)_{t \in [0;\infty[}$ -i.e. a sequence of random variables indexed by t such that:*

1. $B_0 = 0$

2. $B_t - B_s \sim N(0, t-s) \; \forall \, s < t$

3. *B has independent increments, i.e. for every N and a set of N time points $t_1 < \ldots < t_N$, $B_{t_1}, B_{t_2} - B_{t_1}, B_{t_3} - B_{t_2}, \ldots, B_{t_N} - B_{t_{N-1}}$ are independent random variables.*

That these demands on a process can be satisfied simultaneously is not trivial. But don't worry, Brownian motion does exist. It is, however, a fairly "wild" object. The sample paths (formally the mapping $t \mapsto B_t$ and intuitively simply the graph you get by plotting "temperature/stock price/..." against time) of BM are continuous everywhere but differentiable nowhere. The figure shows a simulated sample path of a BM and should give an indica-

tion of this.

A useful fact following from the independent increment property is that for any measurable $f : \mathbb{R} \to \mathbb{R}$ for which $E[|f(B_t - B_s)|] < \infty$ we have

$$E[f(B_t - B_s)|\mathcal{F}_s] = E[f(B_t - B_s)] \tag{7.3}$$

where $\mathcal{F}_s = \sigma\{B_u : 0 \leq u \leq s\}$.

The fundamental assumption of the Black-Scholes model is that the stock price can be represented by

$$S_t = S_0 \exp(\alpha t + \sigma B_t) \tag{7.4}$$

where B_t is a SBM. Such a process is called a geometric BM (with drift). Furthermore, it assumes that there exists a riskless asset (a money market account). One dollar invested in the money market account will grow as

$$\beta_t = \exp(rt) \tag{7.5}$$

where r is a constant (typically $r > 0$). Hence β_t is the continuous time analogue of $R_{0,t}$.

What does (7.4) mean? Note that since $B_t \sim N(0, t)$, S_t has a lognormal distribution and

$$\ln\left(\frac{S_{t_1}}{S_0}\right) = \alpha t_1 + \sigma B_{t_1},$$

$$\ln\left(\frac{S_{t_2}}{S_{t_1}}\right) = \alpha\left(t_2 - t_1\right) + \sigma\left(B_{t_2} - B_{t_1}\right)$$

Since αt, $\alpha\left(t_2 - t_1\right)$, and σ are constant, we see that $\ln\left(\frac{S_{t_1}}{S_0}\right)$ and $\ln\left(\frac{S_{t_2}}{S_{t_1}}\right)$ are independent. The *return,* defined in this section as the logarithm of the price relative, that the stock earns between time t_1 and t_2 is independent of the return earned between time 0 and time t_1, and both are normally distributed. We refer to σ as the *volatility* of the stock - but note that it really describes a property of the logarithmic return of the stock. There are several reasons for modelling the stock price as geometric BM with drift or equivalently all logarithmic returns as independent and normal. First of all, unless it is blatantly unreasonable, modelling "random objects" as "*niid*" is *the* way to start. Empirically it is often a good approximation to model the logarithmic returns as being normal with fixed mean and fixed variance through time.[3] From a probabilistic point of view, it can be shown that if we want a stock price process with continuous sample paths and we want returns to be independent and stationary (but not necessarily normal from the outset), then geometric BM is the only possibility. And last but not least: It gives rise to beautiful financial theory.

If you invest one dollar in the money market account at time 0, it will grow as $\beta_t = \exp(rt)$. Holding one dollar in the stock will give an uncertain amount at time t of $\exp(\alpha t + \sigma B_t)$ and this amount has an expected value of

$$E\exp(\alpha t + \sigma B_t) = \exp(\alpha t + \frac{1}{2}\sigma^2 t).$$

The quantity $\mu = \alpha + \frac{1}{2}\sigma^2$ is often referred to as the *drift* of the stock. We have not yet discussed (even in our discrete models) how agents determine μ and σ^2, but for now think of it this way: Risk averse agents will demand μ to be greater than r to compensate for the uncertainty in the stock's return. The higher σ^2 is, the higher should μ be.

7.3 A derivation of the Black-Scholes formula

In this section we derive the Black-Scholes model taking as given some facts from continuous time finance theory. The main assertion is that the fundamental theorem of asset pricing holds in continuous time and, in particular, in the Black-Scholes setup:

$$S_t \;=\; S_0 \exp\left(\alpha t + \sigma B_t\right)$$

[3] But skeptics would say many empirical analyses of financial data is a case of "believing is seeing"rather than the other way around.

$$\beta_t = \exp(rt)$$

What you are asked to believe in this section are the following facts:

- There is no arbitrage in the model and therefore there exists an *equivalent martingale measure* Q such that the discounted stock price $\frac{S_t}{\beta_t}$ is a martingale under Q. (Recall that this means that $E^Q\left[\frac{S_t}{\beta_t}\,|\mathcal{F}_s\right] = \frac{S_s}{\beta_s}$). The probabilistic behavior of S_t under Q is given by

$$S_t = S_0 \exp\left(\left(r - \frac{1}{2}\sigma^2\right)t + \sigma\widetilde{B_t}\right), \tag{7.6}$$

  where $\widetilde{B_t}$ is a SBM under the measure Q.

- To compute the price of a call option on S with expiration date T and exercise price K, we take the discounted expected value of $C_T = [S_T - K]^+$ assuming the behavior of S_t given by (7.6).

Recall that in the binomial model we also found that the expected return of the stock under the martingale measure was equal to that of the riskless asset.(7.6) is the equivalent of this fact in the continuous time setup. Before sketching how this expectation is computed note that we have not defined the notion of arbitrage in continuous time. Also we have not justified the form of S_t under Q. But let us check at least that the martingale behavior of $\frac{S_t}{\beta_t}$ seems to be OK (this may explain the "$-\frac{1}{2}\sigma^2 t$"-term which is in the expression for S_t). Note that

$$\begin{aligned} E^Q\left[\frac{S_t}{\beta_t}\right] &= E^Q\left[S_0 \exp\left(-\frac{1}{2}\sigma^2 t + \sigma\widetilde{B}_t\right)\right] \\ &= S_0 \exp\left(-\frac{1}{2}\sigma^2 t\right) E^Q\left[\exp\left(\sigma\widetilde{B}_t\right)\right]. \end{aligned}$$

But $\sigma\widetilde{B}_t \sim N(0, \sigma^2 t)$ and since we know how to compute the mean of the lognormal distribution we get that

$$E^Q\left[\frac{S_t}{\beta_t}\right] = S_0 = \frac{S_0}{\beta_0}, \text{ since } \beta_0 = 1.$$

By using the property (7.3) of the Brownian motion one can verify that

$$E^Q\left[\frac{S_t}{\beta_t}\middle|\,\mathcal{F}_s\right] = \frac{S_s}{\beta_s}, \ (\mathcal{F}_s = \text{"information at time s"}).$$

but we will not do that here.[4]

Accepting the fact that the call price at time 0 is

$$C_0 = \exp(-rT)\, E^Q \left[S_0 \exp\left(\left(r - \frac{1}{2}\sigma^2 \right) T + \sigma \widetilde{B}_T \right) - K \right]^+$$

we can get the Black-Scholes formula: We know that $\sigma B_T \sim N(0, \sigma^2 T)$ and also "the rule of the unconscious statistician", which tells us that to compute $E[f(X)]$ for some random variable X which has a density $p(x)$, we compute $\int f(x)\, p(x)\, dx$. This gives us

$$C_0 = e^{-rT} \int_{\mathbb{R}} \left[S_0 e^{(r-\sigma^2/2)T+x} - K \right]^+ \frac{1}{\sqrt{2\pi}\sigma\sqrt{T}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2 T}} dx$$

The integrand is different from 0 when

$$S_0 e^{(r-\sigma^2/2)T+x} > K$$

i.e. when[5]

$$x > \ln(K/S_0) - \left(r - \sigma^2/2\right) T \equiv d$$

So

$$\begin{aligned} C_0 &= e^{-rT} \int_d^\infty \left(S_0 e^{(r-\frac{1}{2}\sigma^2)T+x} - K \right) \frac{1}{\sqrt{2\pi}\sigma\sqrt{T}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2 T}} dx \\ &= \underbrace{e^{-rT} S_0 \int_d^\infty \frac{1}{\sqrt{2\pi}\sigma\sqrt{T}} e^{(r-\frac{1}{2}\sigma^2)T+x} e^{-\frac{1}{2}\frac{x^2}{\sigma^2 T}} dx}_{:=A} - \underbrace{K e^{-rT} \int_d^\infty \frac{1}{\sqrt{2\pi}\sigma\sqrt{T}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2 T}} dx}_{:=B}. \end{aligned}$$

It is easy to see that $B = Ke^{-rT}\text{Prob}(Z > d)$, where $Z \sim N(0, \sigma^2 T)$. So by using symmetry and scaling with $\sigma\sqrt{T}$ we get that

$$B = Ke^{-rT}\Phi(d_2),$$

[4]If you want to try it yourself, use

$$\begin{aligned} E\left[\frac{S_t}{\beta_t} \middle| \mathcal{F}_s \right] &= E\left[\frac{S_t \beta_s S_s}{S_s \beta_t \beta_s} \middle| \mathcal{F}_s \right] \\ &= \frac{S_s}{\beta_s} E\left[\frac{S_t \beta_s}{S_s \beta_t} \middle| \mathcal{F}_s \right] \end{aligned}$$

and then see if you can bring (7.3) into play and use

$$E[\exp(\sigma(B_t - B_s))] = \exp\left(\frac{1}{2}\sigma^2 (t-s) \right).$$

[5]This should bring up memories of the quantity a which we defined in the binomial model.

where (as before)

$$d_2 = -\frac{d}{\sigma\sqrt{T}} = \frac{\ln\left(\frac{S_0}{K}\right) + \left(r - \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}.$$

So "we have half the Black-Scholes formula". The A-term requires a little more work. First we use the change of variable $y = x/(\sigma\sqrt{T})$ to get (with a few rearrangements, a completion of the square, and a further change of variable ($z = y - \sigma\sqrt{T}$))

$$\begin{aligned} A &= S_0 e^{-T\sigma^2/2} \int_{-d_2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\sigma\sqrt{T}y - y^2/2} dy \\ &= S_0 e^{-T\sigma^2/2} \int_{-d_2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(y-\sigma\sqrt{T})^2/2 + T\sigma^2/2} dy \\ &= S_0 \int_{-d_1}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz, \end{aligned}$$

where as per usual $d_1 = d_2 + \sigma\sqrt{T}$. But the last integral we can write as $\text{Prob}(Z > d_1)$ for a random variable $Z \sim N(0,1)$, and by symmetry we get

$$A = S_0 \Phi(d_1),$$

which yields the "promised" result.

Theorem 26 *The unique arbitrage-free price of a European call option on a non-dividend paying stock in the Black-Scholes framework is given by*

$$C_0 = S_0 \Phi\left(d_1\right) - K e^{-rT} \Phi\left(d_2\right)$$

where

$$d_1 = \frac{\ln\left(\frac{S_0}{K}\right) + \left(r + \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}}$$

and

$$d_2 = d_1 - \sigma\sqrt{T},$$

where Φ is the cumulative distribution function of a standard normal distribution.

As stated, the Black-Scholes formula says only what the call price is at time 0. But it is not hard to guess what happens if we want the price at some time $t \in [0;T]$: The same formula applies with S_0 substituted by S_t and T substituted by $T - t$. You may want to "try your hand" with conditional expectations and properties of BM by proving this.

7.3.1 Hedging the call

There is one last thing about the Black-Scholes model/formula you should know. Just as in the binomial model the call option can be hedged in the Black-Scholes model. This means that there exists a self-financing trading strategy involving the stock and the bond such that the value of the strategy at time T is exactly equal to the payoff of the call, $(S_T - K)^+$. (This is in fact the very reason we can talk about a *unique* arbitrage-free price for the call.) It is a general fact that if we have a contract whose price at time t can be written as

$$\pi(t) = F(t, S_t)$$

for some deterministic function F, then the contract is hedged by a strategy consisting of

$$\phi^1(t) = \frac{\partial F}{\partial x}(t, x)\bigg|_{x=S_t}$$

units of the stock and $\phi^0(t) = \pi(t) - \phi^1(t)S_t$ \$ in the bank account. Note that this is a strategy that is continuously adjusted.

For the Black-Scholes model this applies to the call with

$$\begin{aligned} F^{BScall}(t,x) &= x\Phi\left(\frac{\ln\left(\frac{x}{K}\right) + (r + \sigma^2/2)\,(T-t)}{\sigma\sqrt{T-t}}\right) \\ &\quad -Ke^{-r(T-t)}\Phi\left(\frac{\ln\left(\frac{x}{K}\right) + (r - \sigma^2/2)\,(T-t)}{\sigma\sqrt{T-t}}\right). \end{aligned}$$

The remarkable result (and what you must forever remember) is that the partial derivative (wrt. x) of this lengthy expression is simple:[6]

$$\frac{\partial F^{BScall}}{\partial x}(t,x) = \Phi\left(\frac{\ln\left(\frac{x}{K}\right) + (r + \sigma^2/2)\,(T-t)}{\sigma\sqrt{T-t}}\right) = \Phi(d_1),$$

where the last part is standard and understandable but slightly sloppy notation. So to hedge the call option in a Black-Scholes economy you have to hold (at any time t) $\Phi(d_1)$ units of the stock. This quantity is called *the delta* (or: $\triangle$) *hedge ratio* for the call option. The "lingo" comes about because of the intimate relation to partial derivatives; $\triangle$ is approximately the amount that the call price changes, when the stock price changes by 1. In this course we will use computer simulations to illustrate, justify, and hopefully to some degree understand the result.

[6] At one time or another you are bound to be asked to verify this, so you may as well do it right away. Note that if you just look at the B-S formula, forget that S_0 (or x) also appears inside the Φ's, and differentiate, then you get the right result with a wrong proof.

Chapter 8

Some notes on term structure modelling

8.1 Introduction

After the brief encounter with continuous time modelling in Chapter 7 we now return to the discrete time, finite state space models of Chapter 5. They still have a great deal to offer.

One of the most widespread applications of arbitrage pricing in the multi-period finite state space model is in the area of term structure modelling. We saw in Chapter 3 how the term structure could be defined in several equivalent ways through the discount function, the yields of zero coupon bonds and by looking at forward rates. In this chapter we will think of the term structure as the yield of zero coupon bonds as a function of time to maturity. In Chapter 3 we considered the term structure at a fixed point in time. In this chapter our goal is to look at dynamic modelling of the evolution of the term structure. This topic could easily occupy a whole course in itself so here we focus merely on explaining a fundamental method of constructing arbitrage-free systems of bond prices. Once this method is understood the reader will be able to build models for the evolution of the term structure and price interest rate related contingent claims.

We also consider a few topics which are related to term structure modelling and which we can discuss rigorously with our arbitrage pricing technology. These topics are the difference between forwards and futures and the role of 'convexity effects' - or Jensen's inequality - can rule out various properties of term structure evolutions. We also look briefly at so-called swap contracts which are quite important in bond markets.

8.2 Constructing an arbitrage free model

Our goal is to model *prices of zero coupon bonds* of different maturities and through time. Let $P(t,T_i)$, $0 \leq t \leq T_i \leq T$, denote the price at time t of a zero coupon bond with maturity T_i. To follow the notation which is most commonly used in the literature we will deviate slightly from the notation of Chapter 5. To be consistent with Chapter 5 we should write $P(t,T_i)$ for the price of the bond prior to maturity. i.e. when $t < T_i$ and then have a dividend payment $\delta(T_i) = 1$ at maturity and a price process satisfying $P(t,T_i) = 0$ for $t \geq T_i$. We will instead write the dividend into the price and let

$$P(t,t) = 1$$

for all t. (You should have gotten used to this deceptive notation in Chapters 6 and 7.)

We will consider models of bond prices which use the spot rate process $\rho = (\rho_t)_{t=0,\ldots,T-1}$ as the fundamental modelling variable. Recall that the money market account is a process with value 1 and dividend at date $t < T$ given by ρ_{t-1} and a dividend of $1+\rho_T$ at time T. We will need our simple notation for returns obtained by holding money over several periods in the money market account:

Definition 40 *The return of the money market account from period t to u is*

$$R_{t,u} = (1+\rho_t)(1+\rho_{t+1})\cdots(1+\rho_{u-1}), \quad \textit{for } t < u$$

Make sure you understand that $R_{t,t+1}$ is known at time t, whereas $R_{t,t+2}$ is not!

From the fundamental theorem of asset pricing (Theorem 15) we know that the system consisting of the money market account and zero coupon bonds will be arbitrage free if and only if

$$\left(\frac{P(t,T_i)}{R_{0,t}}\right)_{0\leq t\leq T_i}$$

is a martingale for every T_i under some measure Q. Here, we use the fact that the zero coupon bonds only pay one dividend at maturity and we have denoted this dividend $P(T_i,T_i)$ for the bond maturing at date T_i. It is not easy, however, to specify a family of sensible and consistent bond prices. If T is large there are many maturities of zero coupon bonds to keep track of. They all should end up having price 1 at maturity, but that is about all we know. How do we ensure that the large system of prices admits no arbitrage opportunities?

What is often done is the following: We simply construct bond prices as expected discounted values of their terminal price 1 under a measure Q which we specify in advance (as opposed to derive from bond prices). More precisely:

Proposition 27 *Given a spot rate process $\rho = (\rho_t)_{t=0,\ldots,T-1}$. Let*

$$\mathcal{F}_t = \sigma(\rho_0, \rho_1, \ldots, \rho_T).$$

For a given Q define

$$P(t, T_i) = E_t^Q \left[\frac{1}{R_{t,T_i}} \right] \text{ for } 0 \leq t \leq T_i \leq T,$$

where $E_t^Q[\cdot]$ is short hand for $E^Q[\cdot \mid \mathcal{F}_t]$. Then the system consisting of the money market account and the bond price processes $(P(t, T_i))_{t=0,\ldots,T}$ is arbitrage free.

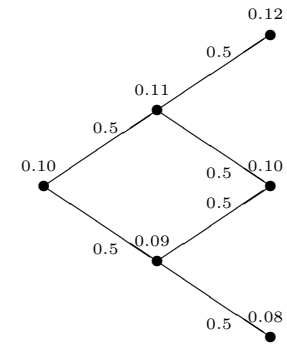
Proof. The proof is an immediate consequence of the definition of prices, since

$$\frac{P(t, T_i)}{R_{0,t}} = \frac{1}{R_{0,t}} E_t^Q \left[\frac{1}{R_{t,T_i}} \right] = E_t^Q \left[\frac{1}{R_{0,T_i}} \right]$$

and this we know defines a martingale for each T_i by Lemma 13. ■

It is important to note that we take Q as given. Another way of putting this is that a P-specification of the short rate (however well it may fit the data) is not enough to determine Q, bond prices and the Q-dynamics of the short rate. If you only have a short rate process, the only traded asset is the bank account and you cannot replicate bonds with that. Later courses will explain this in more detail.

Example 10 Here is a simple illustration of the procedure in a model where the spot rate follows a binomial process.



A ρ-tree with Q-probabilties

$$\begin{pmatrix} 0.8929 \end{pmatrix}$$

$$\begin{pmatrix} 0.9009 \\ 0.8117 \end{pmatrix}$$

$$\begin{pmatrix} 0.9091 \\ 0.8265 \\ 0.7516 \end{pmatrix} \qquad \begin{pmatrix} 0.9091 \end{pmatrix}$$

$$\begin{pmatrix} 0.9174 \\ 0.8418 \end{pmatrix}$$

$$\begin{pmatrix} 0.9259 \end{pmatrix}$$

ZCB prices

The spot rate at time 0 is 0.10. At time 1 it becomes 0.11 with probability $\frac{1}{2}$ and 0.09 with probability $\frac{1}{2}$ (both probabilities under Q) Given that it is 0.09 at time 1, it becomes either 0.10 or 0.08 at time 2, both with probability $\frac{1}{2}$. The bond prices have been computed using Proposition 27. Note that a consequence of Proposition 27 is that (check it!)

$$P(t,T_i) = \frac{1}{1+\rho_t} E_t^Q \left[P(t+1,T_i) \right]$$

and therefore the way to use the proposition is to construct bond prices working backwards through the tree. For a certain maturity T_i we know $P(T_i,T_i) = 1$ regardless of the state. Now the price of this bond at time $T_i - 1$ can be computed as a function of ρ_{T_i-1}, and so forth. The term structure at time 0 is now computed as follows

$$r(0,1) = \frac{1}{P(0,1)} - 1 = 0.1$$

$$r(0,2) = \left(\frac{1}{P(0,2)} \right)^{\frac{1}{2}} - 1 = 0.09995$$

$$r(0,3) = \left(\frac{1}{P(0,3)} \right)^{\frac{1}{3}} - 1 = 0.0998$$

using definitions in Chapter 3. So the term structure in this example is decreasing in t - which is not what is normally seen in the market (but it does happen, for instance in Denmark in 1993 and in the U.S. in 2000). In fact, one calls the term structure *"inverted"* in this case. Note that when the Q-behavior of r has been specified we can determine not only the current term structure, we can find the term structure in any node of the tree. (Since the model only contains two non-trivial zero-coupon bonds at time 1, the term structure only has two points at time 1.)

So Example 10 shows how the term structure is calculated from a Q-tree of the short rate. But what we (or: practitioners) are really interested in is the reverse question: Given todays (observed) term structure, how do we construct a Q-tree of the short rate that is consistent with the term structure? (By consistent we mean that if we use the tree for ρ in Example 10-fashion we match the observed term structure at the first node.) Such a tree is needed for pricing more complicated contracts (options, for instance).

First, it is easy to see that generally such an "inversion" is in no way unique; a wide variety of ρ-trees give the same term structure. But that is not bad; it means that we impose a convenient structure on the ρ-process and still fit observed term structures. Two such conveniences are that the development of ρ can be represented in a recombining tree (a lattice), or in other words that ρ is Markovian, and that the Q-probability 1/2 is attached to all branches. (It may not be totally clear that we can do that, but it is easily seen from the next example/subsection.)

8.2.1 Constructing a Q-tree for the short rate that fits the initial term structure

Imagine a situation where two things have been thrust upon us.

1. The almighty ("God "or "The Market") has determined todays term structure,
$$(P(0,1), P(0,2), \ldots, P(0,T)).$$
2. Our not-so-almighty boss has difficulties understanding probability beyond the tossing of a fair coin and wants answers fast, so he(s secretary) has drawn the ρ-lattice in Figure 8.1.

All we have to do is "fill in the blanks'. Optimistically we start, and in the box corresponding to $(t = 0, i = 0)$ we have no choice but to put

$$\rho_0(0) = \frac{1}{P(0,1)} - 1.$$

To fill out boxes corresponding to $(t = 1, i = 0)$ and $(t = 1, i = 1)$ we have the equation

$$P(0,2) = \frac{1}{\rho_0(0)} \left(\frac{1}{2} \times \frac{1}{1+\rho_1(0)} + \frac{1}{2} \times \frac{1}{1+\rho_1(1)} \right), \tag{8.1}$$

which of course has many solutions. (Even many sensible ones.) So we can/have to put more structure on the problem. Two very popular ways of

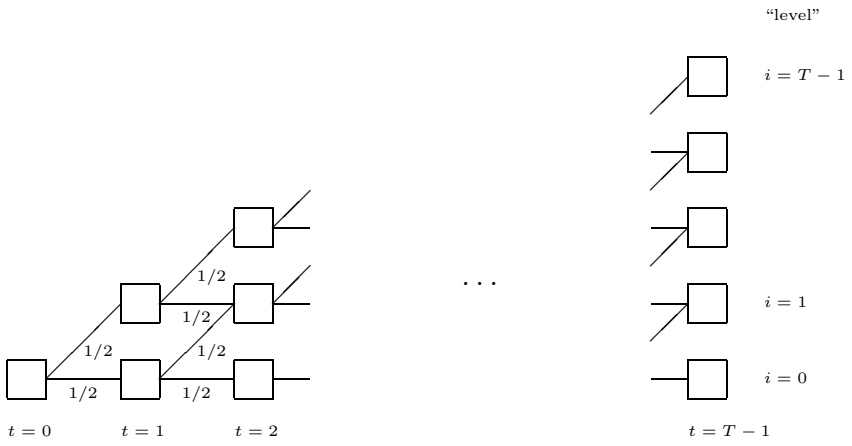


Figure 8.1: The ρ-lattice we must complete.

doing this are these functional forms: [1]

$$\begin{aligned} \text{Ho/Lee-specification:} \quad & \rho_t(i) = a_{imp}(t) + b_{hist} i \\ \text{Black/Derman/Toy-specification:} \quad & \rho_t(i) = a_{imp}(t) \exp(b_{hist} i) \end{aligned}$$

For each t we fit by choosing an appropriate a_{imp}, while b_{hist} is considered a known constant. b_{hist} is called a volatility parameter and is closely related (as you should be able to see) to the conditional variance of the short rate (or its logarithm). This means that it is fairly easy to estimate from historical time series data of the short rate. With b_{hist} fixed, (8.1) can be solved hence determining what goes in the two "$t = 1$"-boxes. We may have to solve the equation determining $a_{imp}(1)$ numerically, but monotonicity makes this an easy task (by bisection or Newton-Raphson, for instance).

And now can can do the same for $t = 2, \ldots, T-1$ and we can put our computer to work and go to lunch. Well, yes and no. Even though we take a long lunch there is a good chance that the computer is not finished when we get back. Why? Note that as it stands, every time we make a guess at $a_{imp}(t)$ (and since a numerical solution is involved we are likely to be making a number of these) we have to work our way backward trough the lattice all the way down to 0. And this we have to do for each t. While not a computational catastrophe (a small calculation shows that the computation time grows as T^3), it does not seem totally efficient. We would like to go through the lattice only once (as it was the case when the initial term structure was determined

[1] Of course there is a reason for the names attached. As so often before, this is for later courses to explain.

from a known ρ-lattice). Fortunately there is a way of doing this. We need the following lemma.

Lemma 28 *Consider the binomial ρ-lattice in Figure 8.1. Let $\psi(t,i)$ be the price at time 0 of a security that pays 1 at time t if state/level i occurs at that time. Then $\psi(0,0)=1$, $\psi(0,i)=0$ for $i>0$ and the following* forward equation *holds:*

$$\psi(t+1,i)=\begin{cases}\frac{\psi(t,i)}{2(1+\rho_t(i))}+\frac{\psi(t,i-1)}{2(1+\rho_t(i-1))} & 0<i<t+1,\\ \frac{\psi(t,i-1)}{2(1+\rho_t(i-1))} & i=t+1,\\ \frac{\psi(t,i)}{2(1+\rho_t(i))} & i=0.\end{cases}$$

Proof. We do the proof only for the "$0<i<t+1$"-case, the others are similar. Recall that we can think of $\mathcal{F}_t$-measurable random variables (of the type considered here) as vectors in in $\mathbb{R}^{t+1}$. Since conditional expectation is linear, we can (for $s\leq t$) think of the $\mathcal{F}_s$-conditional expectation of an $\mathcal{F}_t$-measurable random variable as a linear mapping from $\mathbb{R}^{t+1}$ to $\mathbb{R}^{s+1}$. In other words it can be represented by a $(s+1)\times(t+1)$-matrix. In particular the time $t-1$ price of a contract with time t price X can be represented as

$$E_t^Q\left(\frac{X}{1+\rho_{t-1}}\right)=\Pi_{t-1}X$$

Now note that in the binomial model there are only two places to go from a given point, so the Π_{t-1}-matrices have the form

$$\Pi_{t-1}=\underbrace{\left.\begin{bmatrix}\frac{1-q}{1+\rho_{t-1}(0)} & \frac{q}{1+\rho_{t-1}(0)} & & & 0\\ & \frac{1-q}{1+\rho_{t-1}(1)} & \frac{q}{1+\rho_{t-1}(1)} & & \\ & & \ddots & \ddots & \\ 0 & & & \frac{1-q}{1+\rho_{t-1}(t-1)} & \frac{q}{1+\rho_{t-1}(t-1)}\end{bmatrix}\right\}t\text{ rows}}_{t+1\text{ columns}}$$

Let $e_i(t)$ be the i'th vector of the standard base in $\mathbb{R}^t$. The claim that pays 1 in state i at time $t+1$ can be represented in the lattice by $e_{i+1}(t+2)$ and by iterated expectations we have

$$\psi(t+1,i)=\Pi_0\Pi_1\cdots\Pi_{t-1}\Pi_t e_{i+1}(t+2).$$

But we know that multiplying a matrix by $e_i(t)$ from the right picks out the i'th column. For $0<i<t+1$ we may write the $i+1$'st column of Π_t as (look at $i=1$)

$$\frac{1-q}{1+\rho_t(i-1)}e_i(t+1)+\frac{q}{1+\rho_t(i)}e_{i+1}(t+1).$$

Hence we get

$$\begin{aligned}\psi(t+1,i) &= \Pi_0\Pi_1\cdots\Pi_{t-1}\left(\frac{1-q}{1+\rho_t(i-1)}e_i(t+1)+\frac{q}{1+\rho_t(i)}e_{i+1}(t+1)\right)\\ &= \frac{1-q}{1+\rho_t(i-1)}\underbrace{\Pi_0\Pi_1\cdots\Pi_{t-1}e_i(t+1)}_{\psi(t,i-1)}\\ &\quad+\frac{q}{1+\rho_t(i)}\underbrace{\Pi_0\Pi_1\cdots\Pi_{t-1}e_{i+1}(t+1)}_{\psi(t,i)},\end{aligned}$$

and since $q = 1/2$, this ends the proof. ■

Since $P(0,t) = \sum_{i=0}^{t}\psi(t,i)$, we can use the following algorithm to fit the initial term structure.

1. Let $\psi(0,0) = 1$ and put $t = 1$.
2. Let $\lambda_t(a_{imp}(t-1)) = \sum_{i=0}^{t}\psi(t,i)$ where $\psi(t,i)$ is calculated from the $\psi(t-1,\cdot)$'s using the specified $a_{imp}(t-1)$-value in the forward equation from Lemma 28.
   Solve $\lambda_t(a_{imp}(t-1)) = P(0,t)$ numerically for $a_{imp}(t-1)$.
3. Increase t by one. If $t \le T$ then go to 2., otherwise stop.

An inspection reveals that the computation time of this procedure only grows as T^2, so we have "gained an order", which can be quite significant when T is large. And don't worry: There will be exercises to help you understand and implement this algorithm.

8.3 On the impossibility of flat shifts of flat term structures

Now let us demonstrate that in our term structure modelling framework it is *impossible to have only parallel shifts of a flat term structure*. In other words, in a model with no arbitrage we cannot have bond prices at time 0 given as

$$P(0,t) = \frac{1}{(1+r)^t}$$

for some $r \ge 0$, $t = 1,\ldots,T$ and

$$P(1,t) = \frac{1}{(1+\widetilde{r})^{t-1}},\ t = 2,\ldots,T,$$

where $\widetilde{r}$ is a random variable (which takes on at least two different values with positive probability). To assign meaning to a "flat term structure" at time 1 we should have $T \geq 3$.

Now consider the zero-coupon bonds with maturity dates 2 and 3. If the term structure is flat at time 0 we have for some $r \geq 0$

$$P(0,2) = \frac{1}{(1+r)^2} \text{ and } P(0,3) = \frac{1}{(1+r)^3}$$

and if it remains flat at time 1, there exist a random variable $\widetilde{r}$ such that

$$P(1,2) = \frac{1}{1+\widetilde{r}} \text{ and } P(1,3) = \frac{1}{(1+\widetilde{r})^2}.$$

Furthermore, in an arbitrage-free model it will be the case that

$$\begin{aligned} P(0,2) &= \frac{1}{1+r} E^Q \left[P(1,2) \right] \\ &= \frac{1}{1+r} E^Q \left[\frac{1}{1+\widetilde{r}} \right] \end{aligned}$$

and

$$\begin{aligned} P(0,3) &= \frac{1}{1+r} E^Q \left[P(1,3) \right] \\ &= \frac{1}{1+r} E^Q \left[\frac{1}{(1+\widetilde{r})^2} \right] \end{aligned}$$

Combining these results, we have

$$\frac{1}{1+r} = E^Q \left[\frac{1}{(1+\widetilde{r})} \right]$$

and

$$\frac{1}{(1+r)^2} = E^Q \left[\frac{1}{(1+\widetilde{r})^2} \right]$$

which contradicts Jensen's inequality, for if

$$\frac{1}{1+r} = E^Q \left[\frac{1}{(1+\widetilde{r})} \right]$$

then since $u \longmapsto u^2$ is strictly convex and $\widetilde{r}$ not constant we must have

$$\frac{1}{(1+r)^2} < E^Q \left[\frac{1}{(1+\widetilde{r})^2} \right].$$

Note that the result does not say that it is impossible for the term structure to be flat. But it is inconsistent with no arbitrage to have a flat term structure and *only* have the possibility of moves to other flat term structures.

This explains what goes "wrong" in the example in Section 3.5.3. There the term structure was flat. We then created a position that had a value of 0 at that level of interest rates, but a strictly positive value with at flat term structure at any other level. But if interest rates are really stochastic then an arbitrage-free model cannot have only flat shifts of flat structure.

8.4 On forwards and futures

A forward and a futures contract are very similar contracts: The buyer (seller) of either type of contract is obligated to buy (sell) a certain asset at some specified date in the future for a price - the delivery price - agreed upon today. The forward/futures price of a certain asset is the delivery price which makes the forward/futures contract have zero value initially. It is very important to see that a forward/futures price is closer in spirit to the exercise price of an option than to the price of an option contract. Whereas an option always has positive value (and usually strictly positive) initially, both futures and forwards have zero value initially because the delivery price is used as a balancing tool.

The following example might clarify this: If a stock trades at $100 today and we were to consider buying a futures contract on the stock with delivery in three months and if we had an idea that this stock would not move a lot over the next three months, then we would be happy to pay something for a contract which obligated us to buy the stock in three months for, say, $50. Even though things could go wrong and the stock fall below $50 in three months we consider that a much smaller risk of loss than the chance of gaining a lot from the contract. Similarly, we would not obligate ourselves to buying the stock in three months for, say, $150 without receiving some money now. Somewhere in between $50 and $150 is a delivery price at which we would neither pay nor insist on receiving money to enter into the contract.

In a market with many potential buyers and sellers there is an equilibrium price at which supply meets demand: The number of contracts with that delivery price offered at zero initial cost equals the number of contracts demanded. This equilibrium price is the forward/futures price (depending on which contract we consider). In the following we will look at this definition in a more mathematical way and we will explain in what sense futures and forwards are different. Although they produce different cash flows (see below) that only results in a price difference when interest rates are stochas-

tic. Therefore, we will illustrate this difference with an example involving futures/forwards on bonds. We will ignore margin payments (i.e. payments that one or both sides of the contract have to make initially to guarantee future payments) in this presentation.

First, let us look at the key difference between forwards and futures by illustrating the cash flows involved in both types of contracts: Let F_t denote the forward price at time t for delivery of an underlying asset at time T and let Φ_t denote the futures price of the same asset for delivery at T, where $t \leq T$. Strictly speaking, we should write $F_{t,T}$ and $\Phi_{t,T}$ instead of F_t and Φ_t respectively, since it is important to keep track of both the date at which the contract is entered into and the delivery date. But we have chosen to consider the particular delivery date T and then keep track of how the futures and forward prices change as a function of t. The cash flows produced by the two types of contracts, if bought at time t, are as follows:

	t	t+1	t+2	$\cdots$	T-1	T
Forward	0	0	0	$\cdots$	0	$S_T - F_t$
Futures	0	$\Phi_{t+1} - \Phi_t$	$\Phi_{t+2} - \Phi_{t+1}$	$\cdots$	$\Phi_{T-1} - \Phi_{T-2}$	$S_T - \Phi_{T-1}$

where S_T is the price of the underlying asset at time T. The forward cash flow is self-explanatory. The futures cash flow can be explained as follows: If you buy a futures contract at date t you agree to buy the underlying asset at time T for Φ_t. At time $t+1$ markets may have changed and the price at which futures trade changed to Φ_{t+1}. What happens is now a *resettlement of the futures contract*. If Φ_{t+1} is bigger than Φ_t you (the buyer of the futures at time t) receive the amount $\Phi_{t+1} - \Phi_t$ from the seller at time t+1 whereas you pay the difference between Φ_{t+1} and Φ_t to the seller if $\Phi_{t+1} < \Phi_t$. The story continues as shown in the figure.

We have already seen that if the underlying asset trades at time t and a zero coupon bond with maturity T also trades then the forward price is given as

$$F_t = \frac{S_t}{P(t,T)}$$

i.e.

$$F_t = S_t \left(1 + r\left(t,T\right)\right)^{T-t} \tag{8.2}$$

where $r(t,T)$ is the internal rate of return on the zero coupon bond.

To see what Φ_t is requires a little more work: First of all to avoid arbitrage we must have $\Phi_T = S_T$. Now consider Φ_{T-1}. In an arbitrage free system there exists an equivalent martingale measure Q. The futures price Φ_{T-1} is such

that the cash flow promised by the contract (bought at $T-1$) has value 0. We must therefore have

$$0 = E^Q_{T-1}\left[\frac{S_T - \Phi_{T-1}}{R_{T-1,T}}\right]$$

but since $R_{T-1,T}$ is $\mathcal{F}_{T-1}$-measurable this implies

$$0 = \frac{1}{R_{T-1,T}} E^Q_{T-1}\left[S_T - \Phi_{T-1}\right]$$

i.e.

$$\Phi_{T-1} = E^Q_{T-1}\left[S_T\right] \tag{8.3}$$

Since Q is a martingale measure recall that

$$\frac{S_{T-1}}{R_{0,T-1}} = E^Q_{T-1}\left[\frac{S_T}{R_{0,T}}\right]$$

i.e.

$$S_{T-1} = \frac{1}{1+\rho_{T-1}} E^Q_{T-1}\left[S_T\right]$$

hence we can write (8.3) as

$$\Phi_{T-1} = \left(1+\rho_{T-1}\right) S_{T-1}$$

and that is the same as (8.2) since the yield on a one period zero coupon bond is precisely the spot rate. So we note that with one time period remaining we have $\Phi_{T-1} = F_{T-1}$. But that also follows trivially since with one period remaining the difference in cash flows between forwards and futures does not have time to materialize.

Now consider Φ_{T-2}. By definition Φ_{T-2} should be set such that the cash flow of the futures contract signed at $T-2$ has zero value:

$$0 = E^Q_{T-2}\left[\frac{\Phi_{T-1} - \Phi_{T-2}}{R_{T-2,T-1}} + \frac{S_T - \Phi_{T-1}}{R_{T-2,T}}\right] \tag{8.4}$$

Now note that using the rule of iterated expectations and the expression for Φ_{T-1} we find

$$\begin{aligned} & E^Q_{T-2}\left[\frac{S_T - \Phi_{T-1}}{R_{T-2,T}}\right] \\ = \; & \frac{1}{R_{T-2,T-1}} E^Q_{T-2}\left[E^Q_{T-1}\left[\frac{S_T - \Phi_{T-1}}{R_{T-1,T}}\right]\right] \\ = \; & 0 \end{aligned}$$

so (8.4) holds precisely when

$$\begin{aligned} 0 &= E_{T-2}^{Q}\left[\frac{\Phi_{T-1}-\Phi_{T-2}}{R_{T-2,T-1}}\right] \\ &= \frac{1}{R_{T-2,T-1}} E_{T-2}^{Q}\left[\Phi_{T-1}-\Phi_{T-2}\right] \end{aligned}$$

i.e. we have

$$\Phi_{T-2} = E_{T-2}^{Q}\left[\Phi_{T-1}\right] = E_{T-2}^{Q}\left[S_T\right].$$

This argument can be continued backwards and we arrive at the expression

$$\Phi_t = E_t^Q\left[S_T\right] \tag{8.5}$$

Note that (8.5) is not in general equal to (8.2):

Under Q, we have $S_t = E_t^Q\left[\frac{S_T}{R_{t,T}}\right]$ so if $\frac{1}{R_{t,T}}$ and S_T are uncorrelated under Q we may write

$$S_t = E_t^Q\left[\frac{1}{R_{t,T}}\right] E_t^Q\left[S_T\right] = P(t,T)\Phi_t$$

which would imply that

$$\Phi_t = \frac{S_t}{P(t,T)} = F_t$$

Hence, *if* $\frac{1}{R_{t,T}}$ *and* S_T *are uncorrelated under* Q, *the forward price* F_t *and the futures price* Φ_t *are the same.* A special case of this is *when interest rates are deterministic*, i.e. all future spot rates and hence $R_{t,T}$ are known at time t.

Note that in general,

$$\begin{aligned} \Phi_t - F_t &= \frac{1}{P(t,T)}\left(P(t,T)E_t^Q\left[S_T\right] - S_t\right) \\ &= \frac{1}{P(t,T)}\left(E_t^Q\left[\frac{1}{R_{t,T}}\right] E_t^Q\left[S_T\right] - S_t\right) \\ &= \frac{1}{P(t,T)}\left(E_t^Q\left(\frac{S_T}{R_{t,T}}\right) - Cov_t^Q\left(\frac{1}{R_{t,T}}, S_T\right) - S_t\right) \\ &= \frac{-1}{P(t,T)}\left(Cov_t^Q\left(\frac{1}{R_{t,T}}, S_T\right)\right). \end{aligned}$$

Note that margin payments go to the holder of a futures contract when spot prices rise, i.e. in states where S_T is high. If $\frac{1}{R(t,T)}$ is negatively correlated with S_T, then interest rates tend to be high when the spot price is high and hence the holder of a futures contract will receive cash when interest rates are high. Hence a futures contract is more valuable in that case and the futures price should therefore be set higher to keep the contract value at 0.

8.5 On swap contracts

A swap contract is an agreement to exchange one stream of payments for another. A wide variety of swaps exists in financial markets; they are often tailor-made to the specific need of a company/an investor and can be highly complex. However, we consider only the valuation of the simplest[2] interest rate swap where fixed interest payments are exchanged for floating rate interest payments.

This swap you may see referred to as anything from "basis" to "forward starting *???*monthly payer swap settled in arrears". Fortunately the payments are easier to describe. For a set of equidistant dates $(T_i)_{i=0}^n$, say δ apart, it is a contract with cash flow (per unit of notational principal)

$$\left(\underbrace{\frac{1}{P(T_{i-1},T_i)} - 1}_{\text{floating leg}} - \underbrace{\delta\kappa}_{\text{fixed leg}} \right) \quad \text{at date } T_i \quad \text{for } i = 1, \ldots, n,$$

where κ is a constant (an interest rate with δ-compounding quoted on yearly basis.) You should convince yourself why the so-called floating leg does in fact correspond to receiving floating interest rate payments. The term $(1/P(T_{i-1},T_i) - 1)/\delta$ is often called the $(12^*\delta)$-month LIBOR (which an acronym for London Interbank Offer Rate, and does not really mean anything nowadays, it is just easy to pronounce). Note that the payment made at T_i is known at T_{i-1}.

It is clear that since the payments in the fixed leg are deterministic, they have a value of

$$\delta\kappa \sum_{i=1}^{n} P(t,T_i).$$

The payments in the floating leg are not deterministic. But despite this, we can find their value without a stochastic model for bond prices/interest rates. Consider the following simple portfolio strategy:

Time	Action	Net cash flow
t	Sell 1 T_i-ZCB Buy 1 T_{i-1}-ZCB	$P(t,T_i) - P(t,T_{i-1})$
T_{i-1}	Use principal received from T_{i-1}-ZCB to buy $1/P(T_{i-1},T_i)$ T_i-ZCBs	0
T_i	Close position	$1/P(T_{i-1},T_i) - 1$

[2]Simple *objects* are often referred to as plain vanilla *objects*. But what is seen as simple depends very much on who is looking.

This means that the T_i-payment in the floating leg has a value of $P(t, T_{i-1}) - P(t, T_i)$, so when summing over i see that the value of the floating leg is

$$P(t, T_0) - P(t, T_n).$$

In the case where $t = T_0$ this is easy to remember/interpret. A bullet-like bond that has a principal of 1 pays a coupon that is the short rate must have a price of 1 (lingo: "it is trading at par"). The only difference between this contract and the floating leg is the payment of the principal at time T_n; the time t value of this is $P(t, T_n)$ hence the value of the floating leg is $1 - P(t, T_n)$.

All in all the swap has a value of

$$V = P(t, T_0) - P(t, T_n) - \delta\kappa \sum_{i=1}^{n} P(t, T_i)).$$

But there is a further twist; these basis swaps are only traded with one κ (for each length; each n), namely the one that makes the value 0. This rate is called the swap rate (at a given date for a given maturity)

$$\kappa_n(t) = \frac{P(t, T_0) - P(t, T_n)}{\delta \sum_{i=1}^{n} P(t, T_i)}. \tag{8.6}$$

In practice (8.6) is often used "backwards", meaning that swap rates for swaps of different lengths (called the "swap curve") are used to infer discount factors/the term structure. Note that this is easy to do recursively if we can "get started", which is clearly the case if $t = T_0$.[3]

The main point is that the basis swap can be priced without using a full dynamic model, we only need today's term structure. But it takes only minor changes in the contract specification for this conclusion to break down. For instance different dynamic models with same current term structure give different swap values if the ith payment in the basis swap is transferred to date T_{i-1} (where it is first known; this is called settlement in advance) or if we swap every 3 months against the 6-month LIBOR.

The need for a swap-market can also be motivated by the following example showing swaps can offer comparative advantages. In its swap-formulation it is very inspired by Hull's book, but you you should recognize the idea from introductory economics courses (or David Ricardo's work of 1817, whichever came first). Consider two firms, A and B, each of which wants to borrow

[3]There should be a "don't try this at work" disclaimer here. In the market different day count conventions are often used on the two swap legs, so things may not be quite what they seem.

$10M for 5 years. Firm A prefers to pay a floating rate, say one that is adjusted every year. It could be that the cash-flows generated by the investment (that it presumably needs the $10M for) depend (positively) on the interest rate market conditions. So from their point of view a floating rate loan removes risk. Firm B prefers to borrow at a fixed rate. In this way is knows in advance exactly how much it has to pay over the 5 years, which it is quite concievable that someone would want. The firms contact their banks and recieve the following loan offers: (Lingo: "bp" means basispoints (pronuonced "beeps" if you're really cool) and is one hundredth of a percentage point, i.e. "100bp = 1%")

Firm	Fixed	Floating
A	5Y-ZCB-rate + 50bp	1Y-ZCB-rate + 30bp
B	5Y-ZCB-rate + 170bp	1Y-ZCB-rate + 100bp

So B gets a systematically "worse deal" than A, which could be because is of lower credit quality than A. But "less worse" for a floating rate loan, where they only have to pay 70bp more than A compared to 120bp for a fixed rate loan. So A could take the floating rate offer and B the fixed rate offer, and everybody is mildly happy. But consider the following arrangement: A takes the fixed rate offer from the bank and B the floating rate. A then offers to lend B the 10M as a fixed rate loan "at the 5Y-ZCB-rate + 45bp", whereas B offers to lend A its 10M floating rate loan "at the 1Y-ZCB-rate" (and would maybe add "flat" to indicate that there *is* no spread). In other words A and B are exchanging, or swapping, their bank loans. The result:

A: Pays (5Y-ZCB-rate + 50bp) (to bank), Pays 1Y-ZCB-rate (to B) and receives (5Y-ZCB-rate + 45bp) (from B). In net-terms: Pays 1Y-ZCB-rate+5bp

B: Pays (1Y-ZCB-rate + 100bp) (to bank), Pays (5Y-ZCB-rate + 45bp) (to A) and receives (1Y-ZCB-rate) (from A). In net-terms: Pays 5Y-ZCB-rate+145bp

So this swap-arrangement has put both A and B in a better position (by 25bp) than they would have been had they only used the bank.

But when used in the finance/interest rate context, there is somewhat of a snag in this story. We argued that the loans offered reflected differences in credit quality. If that is so, then it must mean that default ("going broke") is a possibility that cannot be ignored. It is this risk that the bank is "charging extra" for. With this point of view the reason why the firms get better deals after swapping is that each chooses to take on the credit risk from the other party. If firm B defaults, firm A can forget about (at least part of) what's in the "receives from B"-column, but will (certainly with this construction)

only be able to get out of its obligations to B to a much lesser extent. So the firms are getting lower rates by taking on default risk, which a risk of the type "a large loss with a small probability". One can quite sensibly ask if that is the kind of risks that individual firms want to take.

One could try to remedy the problem by saying that we set up a financial institution through which the swapping takes place. This institution should ensure payments to the non-defaulting party (hence taking "credit risk" $\times$ 2), in return for a share of the possible "lower rate"-gain from the swap, and hope for some "law of large numbers"-diversification effect. But that story is questionable; isn't that what the bank is doing in the first place?

So the morale is two-fold: *i*) If something seems to be too good to be true it usually is. Also in credit risk models. *ii*) The only way to see if the spreads offered to firms A and B are set such that there is no gain without extra risk, i.e. consistent with no arbitrage, is to set up a real dynamic stochastic model of the defaults (something that subsequent courses will do), just as stochastic term structure models help us realize that non-flat yield curves do not imply arbitrage.

8.6 On expectation hypotheses

Recall that the spot rate in our term structure models is a stochastic process. At time 0 we do not know what the spot rate will be at time 1. We may however from current bond prices compute the one-period forward rate $f(1,2)$ and it is natural to think that this rate at least carries some information about the level of the spot rate at time 1. For example, one type of *expectation hypothesis* would argue, that the expected value of spot rates is equal to the corresponding forward rates. As we shall see shortly, there is little reason to think that this is satisfied in arbitrage-free models. There are a number of other *expectation hypotheses* that one can formulate concerning future levels of interest rates, bond prices, yields and forward rates. Although we will no go through all of these in great detail, one point should be clear after this: There is essentially only one expectation hypothesis which follows as a simple consequence of no arbitrage (and an assumption of risk neutral agents). Many other form of expectation hypothesis have little mathematical justification, often there are inconsistent with each other, and very often the same form of the expectation hypotheses cannot hold for different maturities.

But let us begin with the good news. We know that in an arbitrage-free model, we have for any zero coupon bond with maturity T_i that

$$P(t,T_i) = \frac{1}{1+\rho_t} E_t^Q \left[P(t+1,T_i) \right].$$

Hence standing at time t, the expected return under Q of holding a bond in one period is

$$\frac{E_t^Q\left[P(t+1,T_i)\right]}{P(t,T_i)} - 1 = \rho_t$$

and this does not depend on the maturity of the bond. Hence under Q, the one-period return on all bonds is the same. This is a mathematical consequence of no arbitrage. It becomes a hypothesis, which we may call the *local expectations hypothesis,* once we claim that this also holds under the measure P which governs the evolution of interest rates in the real world. This would of course be true if $P = Q$, something which only holds in an economy in which all agents are risk-neutral.

Let us assume that $P = Q$ and consider an extension (called the "return to maturity hypothesis") of this local hypothesis to n periods which equates the expected return from rolling over the money market account in n periods with that of holding an $n-$period bond. This would be equivalent to stating that

$$E_t^Q(R_{t,t+n}) = (1 + y(t, t+n))^n$$

where $y(t, t+n)$ is the yield at time t of a bond maturing at time $t+n$. What if we claim that this holds for all n? Then Jensen's inequality brings us into trouble since from our fundamental pricing relationship we have

$$\begin{aligned} P(t,t+n) &= \\ \frac{1}{(1+y(t,t+n))^n} &= E_t^Q\left[\frac{1}{R_{t,t+n}}\right] \end{aligned}$$

and unless interest rates are deterministic we have

$$E_t^Q\left[\frac{1}{R_{t,t+n}}\right] > \frac{1}{E_t^Q R_{t,t+n}}.$$

Finally let us consider another popular hypothesis about the term structure of interest rates, which states that forward rates are unbiased predictors of spot rates. Our discussion of this hypothesis will be much clearer if we have at our disposal the concept of forward measures.

Proposition 29 *Given a term structure model with Q as the martingale measure. Define the random variable Z_T^T as*

$$Z_T^T = \frac{1}{R_{0,T+1}P(0,T+1)}.$$

Then a new probability measure Q^T is defined by letting

$$Q^T(A) = E^Q\left(Z_T^T 1_A\right),\ A \in \mathcal{F}. \tag{8.7}$$

Under this measure, the forward rate process $(f(t,T))_{t=0,\ldots,T}$ is a martingale.
Proof. First note that since $\rho > -1$, $Z_T^T > 0$. Also,

$$E^Q\left(Z_T^T\right) = \frac{1}{P(0,T+1)} E^Q \frac{1}{R_{0,T+1}} = 1$$

and therefore (8.7) defines a new probability measure on Ω. Let

$$\begin{aligned} Z_t^T &= E_t^Q\left(Z_T^T\right) \\ &= \frac{1}{R(0,t)P(0,T+1)} E_t^Q\left(\frac{1}{R_{t,T+1}}\right) \\ &= \frac{P(t,T+1)}{R(0,t)P(0,T+1)}. \end{aligned}$$

Now note that

$$\begin{aligned} E_t^Q\left(\frac{\rho_T}{R_{t,T+1}}\right) &= P(0,T+1)R_{0,t}E_t^Q\left(\frac{\rho_T}{R_{0,T+1}P(0,T+1)}\right) \\ &= \frac{P(t,T+1)}{Z_t^T} E_t^Q\left(Z_T^T \rho_T\right) \\ &= P(t,T+1)E_t^{Q^T}\left(\rho_T\right). \end{aligned}$$

Therefore,

$$\begin{aligned} E_t^{Q^T}\left(\rho_T\right) &= \frac{1}{P(t,T+1)} E_t^Q\left(\frac{\rho_T}{R_{t,T+1}}\right) \\ &= \frac{1}{P(t,T+1)} E_t^Q\left(\frac{1+\rho_T}{R_{t,T+1}} - \frac{1}{R_{t,T+1}}\right) \\ &= \frac{1}{P(t,T+1)} E_t^Q\left(\frac{1}{R_{t,T}} - \frac{1}{R_{t,T+1}}\right) \\ &= \frac{P(t,T)}{P(t,T+1)} - 1 \\ &= f(t,T). \end{aligned}$$

This proves the martingale property. ■

This proposition shows that there exists a measure (and this measure is called the $T-$forward measure) under which the expected spot rate at

time T is equal to the forward rate. Typically, the forward measure is not equal to P, and it is not equal to Q unless interest rates are deterministic. Furthermore, one may check that for $f(t,T)$ and $f(t,T+1)$ to be unbiased estimators of r_T and r_{T+1},respectively, the spot interest rate at time T must be deterministic. The moral of all this, is that viewing the forward rate as unbiased estimators of future spot rates is problematic.

8.7 Why $P = Q$ means risk neutrality

In this section we will keep referring to the measure P which is the measure determining the actual evolution of prices. To make sure that the meaning of P is clear, we can say that a statistician estimating parameters of prices is trying to find P. We have seen that the version of the expectations hypothesis known as the local expectations hypothesis holds under the martingale measure Q used for pricing. Recall that the measure Q is a measure which allows us to give convenient expressions for prices of claims and derivative securities but not a measure governing the actual movement of prices.

We have stated earlier somewhat loosely that P and Q are actually the same when agents are risk neutral. Since we have not seen many agents this statement needs some elaboration. A quick sketch of this line of reasoning is the following: Recall that under Q all securities have the same one period returns: They are equal to the short rate. If $Q = P$ it would be the case that *actual* expected returns were the same for all assets, regardless of their variances. This would only be possible in a world where agents are risk neutral and therefore do not care about risk (variance, say) but look only at expected returns and prefer more expected return to less. In fact, if there is as much as one risk neutral agent in the economy and two assets have different expected returns, then this one agent would ruin the equilibrium by demanding infinitely much of the asset with the high expected return and financing the purchase by selling the asset with low expected return in infinite quantities. Therefore, we may say that $Q = P$ follows from risk neutrality of at least one agent. The argument can be made more precise by explicitly modelling the inter-temporal optimization problem of a representative agent who maximizes an additively separable expected utility of consumption over a certain time period. When this is done we can interpret the pricing relation

$$P(t,T_i) = \frac{1}{1+\rho_t} E_t^Q \left[P(t+1,T_i) \right]$$

in terms of marginal utilities. The key result is that in equilibrium the prices of bonds adjust in such a way that the increase in marginal utility for the

agent obtained by selling the bond at date t and using the proceeds for consumption is exactly equal to the marginal loss of expected utility at date $t+1$ resulting from the smaller amount of money for consumption available by selling the position in that bond at time $t+1$. Let us consider a one-period case. If we denote by C_0 (known at time 0) and C_1 (stochastic viewed from time 0) the optimal consumption of the agent at dates 0 and 1, it will be the case in equilibrium that the price of the $i'th$ asset satisfies

$$P^i(0)u'(C_0) = E_0^P\left[P^i(1)u'(C_1)\right]$$

i.e.

$$\begin{aligned} P^i(0) &= E_0^P\left[\frac{P^i(1)u'(C_1)}{u'(C_0)}\right] \\ &= E_t^P\left[\frac{P^i(1)Z_1}{1+\rho_0}\right] \end{aligned}$$

where

$$\begin{aligned} Z_1 &= \frac{u'(C_1)}{E_0^P u'(C_1)} \\ 1+\rho_0 &= \frac{u'(C_0)}{E_t^P u'(C_1)} \end{aligned}$$

and this we may then write as

$$P^i(0) = E_t^Q\left[\frac{P^i(1)}{1+\rho_0}\right]$$

where Q is defined by

$$Q(A) = E^P(1_A Z_1).$$

This establishes the connection between utility maximization and the equivalent martingale measure. An agent who is risk neutral will have an affine utility function, and hence for such an agent $u'(C_1)$ is constant (i.e. does not vary with ω as C_1 does). In that case $Z_1 = 1$ and $P = Q$.

It is clear that $P = Q$ is sufficient for the local expectation hypothesis to hold but it may seem to be too strong a requirement. After all, it is only an expectation of one random variable that we are referring to and one could imagine that a measure change would not alter this particular expectation. To analyze this question a little further, consider the fundamental definition of a new measure through the random variable Z_1 :

$$Q(A) = E^P(1_A Z_1).$$

For some random variable X, which could be the spot rate at some future date, we have

$$E^Q(X) = E^P(XZ)$$

and therefore $E^Q(X) = E^P(X)$ if and only if

$$E^P(X(Z-1)) = 0.$$

Since $E(Z-1) = 0$ this is the same as requiring

$$Cov(X, Z) = 0.$$

Therefore, for the change of measure to preserve a mean value we must have that the variable in question is uncorrelated with the change of measure variable Z, and this will typically not hold in the term structure models we consider.

Chapter 9

Portfolio Theory

Matrix Algebra

First we need a few things about matrices. (A very useful reference for mathematical results in the large class imprecisely defined as "well-known"is Berck & Sydsæter (1992), "Economists' Mathematical Handbook", Springer.)

- When $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ then

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{V}\mathbf{x}) = (\mathbf{V} + \mathbf{V}^\top)\mathbf{x}$$

- A matrix $\mathbf{V} \in \mathbb{R}^{n \times n}$ is said to be *positive definite* if $\mathbf{z}^\top \mathbf{V}\mathbf{z} > 0$ for all $\mathbf{z} \neq \mathbf{0}$. If $\mathbf{V}$ is positive definite then $\mathbf{V}^{-1}$ exists and is also positive definite.

- Multiplying (appropriately) partitioned matrices is just like multiplying 2×2-matrices.

- When X is an n-dimensional random variable with covariance matrix $\mathbf{\Sigma}$ then

$$\mathrm{Cov}(\mathbf{A}X + \mathbf{B}, \mathbf{C}X + \mathbf{D}) = \mathbf{A}\mathbf{\Sigma}\mathbf{C}^\top,$$

  where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are deterministic matrices such that the multiplications involved are well-defined.

Basic Definitions & Justification of Mean-Variance Analysis

We will consider an agent who wants to invest in the financial markets. We look at a simple model with only two time-points, 0 and 1. The agent has an initial wealth of W_0 to invest. We are not interested in how the agent

determined this amount, it's just there. There are n financial assets to choose from and these have prices

$$S_{i,t} \text{ for } i = 1, \ldots, n \text{ and } t = 0, 1,$$

where $S_{i,1}$ is stochastic and not known until time 1. The rate of return on asset i is defined as

$$r_i = \frac{S_{i,1} - S_{i,0}}{S_{i,0}},$$

and $r = (r_1, \ldots, r_n)^\top$ is the vector of rates of return. Note that r is stochastic.

At time 0 the agent chooses a portfolio, that is he buys a_i units of asset i and since all in all W_0 is invested we have

$$W_0 = \sum_{i=1}^{n} a_i S_{i,0}.$$

(If $a_i < 0$ the agent is selling some of asset i; in most of our analysis short-selling will be allowed.)

Rather than working with the absolute number of assets held, it is more convenient to work with relative portfolio weights. This means that for the ith asset we measure the value of the investment in that asset relative to total investment and call this w_i, i.e.

$$w_i = \frac{a_i S_{i,0}}{\sum_{i=1}^{n} a_i S_{i,0}} = \frac{a_i S_{i,0}}{W_0}.$$

We put $\mathbf{w} = (w_1, \ldots w_n)^\top$, and have that $\mathbf{w}^\top \mathbf{1} = 1$. In fact, *any* vector satisfying this condition identifies an investment strategy. Hence in the following a portfolio is a vector whose coordinate sum to 1. Note that in this one period model a portfolio $\mathbf{w}$ is not a stochastic variable (in the sense of being unknown at time 0).

The terminal wealth is

$$\begin{aligned} W_1 &= \sum_{i=1}^{n} a_i S_{i,1} = \sum_{i=1}^{n} a_i (S_{i,1} - S_{i,0}) + \sum_{i=1}^{n} a_i S_{i,0} \\ &= W_0 \left(1 + \sum_{i=1}^{n} \frac{S_{i,0} a_i}{W_0} \frac{S_{i,1} - S_{i,0}}{S_{i,0}} \right) \\ &= W_0 (1 + \mathbf{w}^\top r), \end{aligned} \tag{9.1}$$

so if we know the relative portfolio weights and the realized rates of return, we know terminal wealth. We also see that

$$E(W_1) = W_0 (1 + \mathbf{w}^\top E(r))$$

and

$$\text{Var}(W_1) \quad = \quad W_0^2 \text{Cov}(\mathbf{w}^\top r, \mathbf{w}^\top r) = W_0^2 \mathbf{w}^\top \text{Var}(r)\mathbf{w}.$$

In this chapter we will look at how agents should choose $\mathbf{w}$. We will focus on how to choose $\mathbf{w}$ such that for a given expected rate of return, the variance on the rate of return is minimized. This is called mean-variance analysis. Intuitively, it sounds reasonable enough, but can it be justified?

An agent has a utility function, u, and let us for simplicity say that he derives utility from directly from terminal wealth. (So in fact we are saying that we can eat money.) We can expand u in a Taylor series around the expected terminal wealth,

$$\begin{aligned} u(W_1) \quad &= \quad u(E(W_1)) + u'(E(W_1))(W_1 - E(W_1)) \\ &\quad + \frac{1}{2}u''(E(W_1))(W_1 - E(W_1))^2 + R_3, \end{aligned}$$

where the remainder term R_3 is

$$R_3 = \sum_{i=3}^{\infty} \frac{1}{i!} u^{(i)}(E(W_1))(W_1 - E(W_1))^i,$$

"and hopefully small". With appropriate (weak) regularity condition this means that expected terminal wealth can be written as

$$E(u(W_1)) = u(E(W_1)) + \frac{1}{2}u''(E(W_1))\text{Var}(W_1) + E(R_3),$$

where the remainder term involves higher order central moments. As usual we consider agents with increasing, concave (i.e. $u'' < 0$) utility functions who maximize expected wealth. This then shows that to a second order approximation there is a preference for expected wealth (and thus, by (9.1), to expected rate of return), and an aversion towards variance of wealth (and thus to variance of rates of return).

But we also see that mean-variance analysis cannot be a completely general model of portfolio choice. A sensible question to ask is: What restrictions can we impose (on u and/or on r) to ensure that mean-variance analysis is fully consistent with maximization of expected utility?

An obvious way to do this is to assume that utility is quadratic. Then the remainder term is identically 0. But quadratic utility does not go too well with the assumption that utility is increasing and concave. If u is concave (which it has to be for mean-variance analysis to hold ; otherwise our interest would be in maximizing variance) there will be a point of satiation beyond

which utility decreases. Despite this, quadratic utility is often used with a "happy-go-lucky" assumption that when maximizing, we do not end up in an area where it is decreasing.

We can also justify mean-variance analysis by putting distributional restrictions on rates of return. If rates of return on individual assets are normally distributed then the rate of return on a portfolio is also normal, and the higher order moments in the remainder can be expressed in terms of the variance. In general we are still not sure of the signs and magnitudes of the higher order derivatives of u, but for large classes of reasonable utility functions, mean-variance analysis can be formally justified.

9.1 The Mathematics of the Efficient Frontier

9.1.1 The case with no riskfree asset

First we consider a market with no riskfree asset and n risky assets. Later we will include a riskfree asset, and it will become apparent that we have done things in the right order.

The risky assets have a vector of rates of return of r, and we assume that

$$E(r) = \boldsymbol{\mu}, \tag{9.2}$$
$$\mathrm{Var}(r) = \boldsymbol{\Sigma}, \tag{9.3}$$

where $\boldsymbol{\Sigma}$ is positive definite (hence invertible) and not all coordinates of $\boldsymbol{\mu}$ are equal. As a covariance matrix $\boldsymbol{\Sigma}$ is always positive semidefinite, the definiteness means that there does not exist an asset whose rate of return can be written as an affine function of the other $n-1$ assets' rates of return. Note that the existence of a riskfree asset would violate this.

Consider the following problem:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^\top \boldsymbol{\Sigma}\mathbf{w} := \sigma_P^2 \quad \text{subject to} \quad \begin{aligned} \mathbf{w}^\top \boldsymbol{\mu} &= r_P \\ \mathbf{w}^\top \mathbf{1} &= 1 \end{aligned}$$

First note that our assumptions on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ ensure that a unique finite solution exits for any value of r_P. Second note that the problem can be interpreted as choosing portfolio weights (the second constraint ensures that $\mathbf{w}$ is a vector of portfolio weights) such that the variance on the return on the portfolio ($\mathbf{w}^\top \boldsymbol{\Sigma}\mathbf{w}$; the "1/2" is just there for convenience) is minimized given that we want a specific expected rate of return (r_P; "P is for portfolio").

To solve the problem we set up the Lagrange-function with multipliers

$$\mathcal{L}(\mathbf{w}, \lambda_1, \lambda_2) = \frac{1}{2}\mathbf{w}^\top \boldsymbol{\Sigma}\mathbf{w} - \lambda_1(\mathbf{w}^\top \boldsymbol{\mu} - r_P) - \lambda_2(\mathbf{w}^\top \mathbf{1} - 1).$$

The first-order conditions for optimality are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \boldsymbol{\Sigma}\mathbf{w} - \lambda_1\boldsymbol{\mu} - \lambda_2\mathbf{1} &= 0, & (9.4)\\ \mathbf{w}^\top\boldsymbol{\mu} - r_P &= 0, & (9.5)\\ \mathbf{w}^\top\mathbf{1} - 1 &= 0. & (9.6) \end{aligned}$$

Usually we might say "and these are linear equations that can easily be solved", but working on them algebraically leads to a much deeper understanding and intuition about the model. Note that invertibility gives that we can write (9.4) as (check for yourself)

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu} \;\; \mathbf{1}] \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}, \tag{9.7}$$

and (9.5)-(9.6) as

$$[\boldsymbol{\mu} \;\; \mathbf{1}]^\top \mathbf{w} = \begin{bmatrix} r_P \\ 1 \end{bmatrix}. \tag{9.8}$$

Multiplying both sides of (9.7) by $[\boldsymbol{\mu} \;\; \mathbf{1}]^\top$ and using (9.8) gives

$$\begin{bmatrix} r_P \\ 1 \end{bmatrix} = [\boldsymbol{\mu}\; \mathbf{1}]^\top \mathbf{w} = \underbrace{[\boldsymbol{\mu}\; \mathbf{1}]^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu}\; \mathbf{1}]}_{:=\mathbf{A}} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}. \tag{9.9}$$

By using the multiplication rules for partitioned matrices we see that

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} & \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\mathbf{1} \\ \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\mathbf{1} & \mathbf{1}^\top\boldsymbol{\Sigma}^{-1}\mathbf{1} \end{bmatrix} := \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

We now show that $\mathbf{A}$ is positive definite, in particular it is invertible. To this end let $\mathbf{z}^\top = (z_1, z_2) \neq \mathbf{0}$ be an arbitrary non-zero vector in $\mathbb{R}^2$. Then

$$\mathbf{y} = [\boldsymbol{\mu}\; \mathbf{1}] \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = [z_1\boldsymbol{\mu} \quad z_2\mathbf{1}] \neq \mathbf{0},$$

because the coordinates of $\boldsymbol{\mu}$ are not all equal. From the definition of $\mathbf{A}$ we get

$$\forall \mathbf{z} \neq \mathbf{0} \;\; : \;\; \mathbf{z}^\top\mathbf{A}\mathbf{z} = \mathbf{y}^\top\boldsymbol{\Sigma}^{-1}\mathbf{y} > 0,$$

because $\mathbf{\Sigma}^{-1}$ is positive definite (because $\mathbf{\Sigma}$ is). In other words, $\mathbf{A}$ is positive definite. Hence we can solve (9.9) for the λ's,

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \mathbf{A}^{-1} \begin{bmatrix} r_P \\ 1 \end{bmatrix},$$

and insert this into (9.7) in order to determine the optimal portfolio weights

$$\widehat{\mathbf{w}} = \mathbf{\Sigma}^{-1}[\boldsymbol{\mu}\ \mathbf{1}]\mathbf{A}^{-1} \begin{bmatrix} r_P \\ 1 \end{bmatrix}. \tag{9.10}$$

The portfolio $\widehat{\mathbf{w}}$ is called the minimum variance portfolio for a given mean r_P (So we can't be bothered to say the correct full phrase: "minimum variance on rate of return for a given mean rate on return r_P ".) Twice the optimal value (i.e. the minimal portfolio return variance) is

$$\begin{aligned} \widehat{\sigma}_P^2 &= \widehat{\mathbf{w}}^\top \mathbf{\Sigma} \widehat{\mathbf{w}} \\ &= [r_P\ \ 1]\mathbf{A}^{-1}[\boldsymbol{\mu}\ \ \mathbf{1}]^\top \mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}[\boldsymbol{\mu}\ \ \mathbf{1}]\mathbf{A}^{-1}[r_P\ \ 1]^\top \\ &= [r_P\ \ 1]\mathbf{A}^{-1} \underbrace{\left([\boldsymbol{\mu}\ \ \mathbf{1}]^\top \mathbf{\Sigma}^{-1}[\boldsymbol{\mu}\ \ \mathbf{1}]\right)}_{=\mathbf{A} \text{ by def.}} \mathbf{A}^{-1}[r_P\ \ 1]^\top \\ &= [r_P\ \ 1]\mathbf{A}^{-1} \begin{bmatrix} r_P \\ 1 \end{bmatrix}, \end{aligned}$$

where symmetry (of $\mathbf{\Sigma}$ and $\mathbf{A}$ and their inverses) was used to obtain the second line. But note that

$$\mathbf{A}^{-1} = \frac{1}{ac - b^2} \begin{bmatrix} c & -b \\ -b & a \end{bmatrix},$$

which gives us

$$\widehat{\sigma}_P^2 = \frac{a - 2br_P + cr_P^2}{ac - b^2}. \tag{9.11}$$

In (9.11) the relation between the variance of the minimum variance portfolio for a given r_p, $\widehat{\sigma}_P^2$, is expressed as a parabola and is called the *variance portfolio frontier* or *locus*. In mean-standard deviation-space the relation is expressed as a hyperbola. Figure 9.1 illustrates what things look like in mean-variance-space. (When using graphical arguments you should be quite careful to use "the right space"; for instance lines that are straight in one space, are not straight in the other.) The upper half of the curve in Figure 9.1 (the solid line) identifies the set of portfolios that have the highest mean return for a given variance; these are called mean-variance *efficient portfolios*. The portfolios on the bottom half (the dotted part) are called *inefficient*

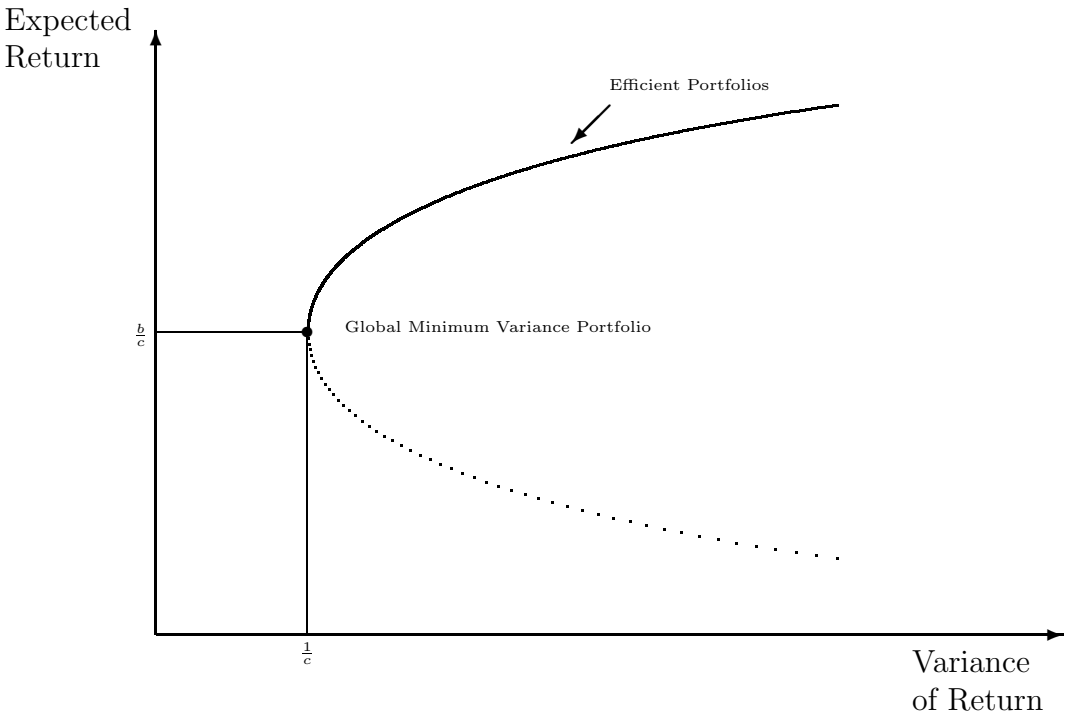


Figure 9.1: The minimum variance portfolio frontier.

portfolios. Figure 9.1 also shows the *global minimum variance portfolio*, the portfolio with the smallest possible variance for any given mean return. Its mean, r_G, is found by minimizing (9.11) with respect to r_P, and is $r_{gmv} = \frac{b}{c}$. By substituting this in the general $\widehat{\sigma}^2$-expression we obtain

$$\widehat{\sigma}^2_{gmv} = \frac{a - 2br_{gmv} + cr^2_{gmv}}{ac - b^2} = \frac{a - 2b(b/c) + c(b/c)^2}{ac - b^2} = \frac{1}{c},$$

while the general formula for portfolio weights gives us

$$\widehat{\mathbf{w}}_{gmv} = \frac{1}{c}\mathbf{\Sigma}^{-1}\mathbf{1}.$$

Example 11 (A Recurrent Numerical Example) Consider the case with 3 assets (referred to as A, B, and C) and

$$\boldsymbol{\mu} = \begin{bmatrix} 0.1 \\ 0.12 \\ 0.15 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} 0.25 & 0.10 & -0.10 \\ 0.10 & 0.36 & -0.30 \\ -0.10 & -0.30 & 0.49 \end{bmatrix}.$$

Figure 9.2: The minimum variance frontiers and individual assets for Example 11

The all-important **A**-matrix is then

$$\mathbf{A} = \begin{bmatrix} 0.33236 & 2.56596 \\ 2.565960 & 20.04712 \end{bmatrix},$$

which means that the locus of mean-variance portfolios is given by

$$\widehat{\sigma}_P^2 = 4.22918 - 65.3031 r_P + 255.097 r_P^2.$$

The locus is illustrated in Figure 9.2 in both in (variance, expected return)-space and (standard deviation, expected return)-space.

An important property of the set of minimum variance portfolios is is so-called two-fund separation. This means that the minimum variance portfolio frontier can be generated by any two distinct frontier portfolios.

Proposition 30 *Let* $\mathbf{x}_a$ *and* $\mathbf{x}_b$ *be two minimum variance portfolios with mean returns* r_a *and* r_b*,* $r_a \neq r_b$*. Then every minimum variance portfolio,* $\mathbf{x}_c$ *is a linear combination of* $\mathbf{x}_a$ *and* $\mathbf{x}_b$*. Conversely, every portfolio that is a linear combination of* $\mathbf{x}_a$ *and* $\mathbf{x}_b$ *(i.e. can be written as* $\alpha\mathbf{x}_a + (1-\alpha)\mathbf{x}_b$*) is a minimum variance portfolio. In particular, if* $\mathbf{x}_a$ *and* $\mathbf{x}_b$ *are efficient portfolios, then* $\alpha\mathbf{x}_a + (1-\alpha)\mathbf{x}_b$ *is an efficient portfolio for* $\alpha \in [0;1]$*.*

Proof. To prove the first part let r_c denote the mean return on a given minimum variance portfolio $\mathbf{x}_c$. Now choose α such that $r_c = \alpha r_a + (1-\alpha) r_b$, that is $\alpha = (r_c - r_b)/(r_a - r_b)$ (which is well-defined because $r_a \neq r_b$). But since $\mathbf{x}_c$ is a minimum variance portfolio we know that (9.10) holds, so

$$\begin{aligned}
\mathbf{x}_c &= \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu} \;\; \mathbf{1}]\mathbf{A}^{-1} \begin{bmatrix} r_c \\ 1 \end{bmatrix} \\
&= \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu} \;\; \mathbf{1}]\mathbf{A}^{-1} \begin{bmatrix} \alpha r_a + (1-\alpha) r_b \\ \alpha + (1-\alpha) \end{bmatrix} \\
&= \alpha \mathbf{x}_a + (1-\alpha)\mathbf{x}_b,
\end{aligned}$$

where the third line is obtained because $\mathbf{x}_a$ and $\mathbf{x}_b$ also fulfill (9.10). This proves the first statement. The second statement is proved by "reading from right to left" in the above equations. This shows that $\mathbf{x}_c = \alpha \mathbf{x}_a + (1-\alpha)\mathbf{x}_b$ is the minimum variance portfolio with expected return $\alpha r_a + (1-\alpha) r_b$. From this, the validity of the third statement is clear. ■

Another important notion is *orthogonality* of portfolios. We say that two portfolios $\mathbf{x}_P$ and $\mathbf{x}_{zP}$ ("z is for zero") are orthogonal if the covariance of their rates of return is 0, i.e.

$$\mathbf{x}_{zP}^\top \boldsymbol{\Sigma} \mathbf{x}_P = 0. \tag{9.12}$$

Often $\mathbf{x}_{zP}$ is called $\mathbf{x}_P$'s 0-β portfolio (we'll see why later).

Proposition 31 *For every minimum variance portfolio, except the global minimum variance portfolio, there exists a unique orthogonal minimum variance portfolio. Furthermore, if the first portfolio has mean rate of return r_P, its orthogonal one has mean*

$$r_{zP} = \frac{a - b r_P}{b - c r_P}.$$

Proof. First note that r_{zP} is well-defined for any portfolio except the global minimum variance portfolio. By (9.10) we know how to find the minimum variance portfolios with means r_P and $r_{zP} = (a - b r_P)/(b - c r_P)$. This leads to

$$\begin{aligned}
\mathbf{x}_{zP}^\top \boldsymbol{\Sigma} \mathbf{x}_P &= [r_{zP} \;\; 1]\mathbf{A}^{-1}[\boldsymbol{\mu} \;\; \mathbf{1}]^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu} \;\; \mathbf{1}]\mathbf{A}^{-1}[r_P \;\; 1]^\top \\
&= [r_{zP} \;\; 1]\mathbf{A}^{-1} \underbrace{\left([\boldsymbol{\mu} \;\; \mathbf{1}]^\top \boldsymbol{\Sigma}^{-1}[\boldsymbol{\mu} \;\; \mathbf{1}]\right)}_{=\mathbf{A} \text{ by def.}} \mathbf{A}^{-1}[r_P \;\; 1]^\top
\end{aligned}$$

$$
\begin{aligned}
&= [r_{zP} \;\; 1]\mathbf{A}^{-1}\begin{bmatrix} r_P \\ 1 \end{bmatrix} \qquad (9.13)\\
&= \begin{bmatrix} \dfrac{a-br_P}{b-cr_P} & 1 \end{bmatrix} \frac{1}{ac-b^2}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}\begin{bmatrix} r_P \\ 1 \end{bmatrix}\\
&= \frac{1}{ac-b^2}\begin{bmatrix} \dfrac{a-br_P}{b-cr_P} & 1 \end{bmatrix}\begin{bmatrix} cr_P-b \\ a-br_P \end{bmatrix}\\
&= 0,
\end{aligned}
$$

which was the desired result. ■

Proposition 32 *Let $\mathbf{x}_{mv}$ ($\neq \mathbf{x}_{gmv}$, the global minimum variance portfolio) be a portfolio on the mean-variance frontier with rate of return r_{mv}, expected rate of return μ_{mv} and variance σ^2_{mv}. Let $\mathbf{x}_{zmv}$ be the corresponding orthogonal portfolio, $\mathbf{x}_P$ be an arbitrary portfolio, and use similar notation for rates of return on these portfolios. Then the following holds:*

$$\mu_P - \mu_{zmv} = \beta_{P,mv}(\mu_{mv} - \mu_{zmv}),$$

where

$$\beta_{P,mv} = \frac{Cov(r_P, r_{mv})}{\sigma^2_{mv}}.$$

Proof. Consider first the covariance between return on asset i and $\mathbf{x}_{mv}$. By using (9.10) we get

$$
\begin{aligned}
\mathrm{Cov}(r_i, r_{mv}) &= \mathbf{e}_i^\top \mathbf{\Sigma}\mathbf{x}_{mv}\\
&= \mathbf{e}_i^\top[\boldsymbol{\mu} \;\; \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} \mu_{mv} \\ 1 \end{bmatrix}\\
&= [\mu_i \;\; 1]\mathbf{A}^{-1}\begin{bmatrix} \mu_{mv} \\ 1 \end{bmatrix}.
\end{aligned}
$$

From calculations in the proof of Proposition 31 we know that the covariance between $\mathbf{x}_{mv}$ and $\mathbf{x}_{zvp}$ is given by (9.13). We also know that it is 0. Subtracting this 0 from the above equation gives

$$
\begin{aligned}
\mathrm{Cov}(r_i, r_{mv}) &= [\mu_i - \mu_{zmv} \;\; 0]\mathbf{A}^{-1}\begin{bmatrix} \mu_{mv} \\ 1 \end{bmatrix}\\
&= (\mu_i - \mu_{zmv})\underbrace{\frac{c\mu_{mv}-b}{ac-b^2}}_{:=\gamma}, \qquad (9.14)
\end{aligned}
$$

where we have used the formula for $\mathbf{A}^{-1}$. Since this holds for all individual assets and covariance is bilinear, it also holds for portfolios. In particular for $\mathbf{x}_{mv}$,

$$\sigma_{mv}^2 = \gamma(\mu_{mv} - \mu_{zmv}),$$

so $\gamma = \sigma_{mv}^2/(\mu_{mv} - \mu_{zmv})$. By substituting this into (9.14) we get the desired result for individual assets. But then linearity ensures that it holds for all portfolios. ■

Proposition 32 says that the expected excess return on any portfolio (over the expected return on a certain portfolio) *is a linear function* of the expected excess return on a minimum variance portfolio. It also says that the expected excess return is proportional to covariance.

9.1.2 The case with a riskfree asset

We now consider a portfolio selection problem with $n+1$ assets. These are indexed by $0, 1, \ldots, n$, and 0 corresponds to the riskfree asset with (deterministic) rate of return r_0. For the risky assets we let ${r_i}^e$ denote the *excess* rate of return over the riskfree asset, i.e. the actual rate of return less r_0. We let $\boldsymbol{\mu}^e$ denote the mean excess rate of return, and $\boldsymbol{\Sigma}$ the variance (which is of course unaffected). A portfolio is now a $n+1$-dimensional vector whose coordinate sum to unity. But in the calculations we let $\mathbf{w}$ denote the vector of weights $w_1, \ldots, w_n$ corresponding to the risky assets and write $w_0 = 1 - \mathbf{w}^\top \mathbf{1}$.

With these conventions the mean excess rate of return on a portfolio P is

$$r_P^e = \mathbf{w}^\top \boldsymbol{\mu}^e$$

and the variance is

$$\sigma_P^2 = \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}.$$

Therefore the mean-variance portfolio selection problem with a riskless asset can be stated as

$$\min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^\top \boldsymbol{\mu}^e = r_P^e.$$

Note that $\mathbf{w}^\top \mathbf{1} = 1$ is not a constraint; some wealth may be held in the riskless asset.

As in the previous section we can set up the Lagrange-function, differentiate it, at solve to first order conditions. This gives the optimal weights

$$\widehat{\mathbf{w}} = \frac{r_P^e}{(\boldsymbol{\mu}^e)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^e} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^e, \tag{9.15}$$

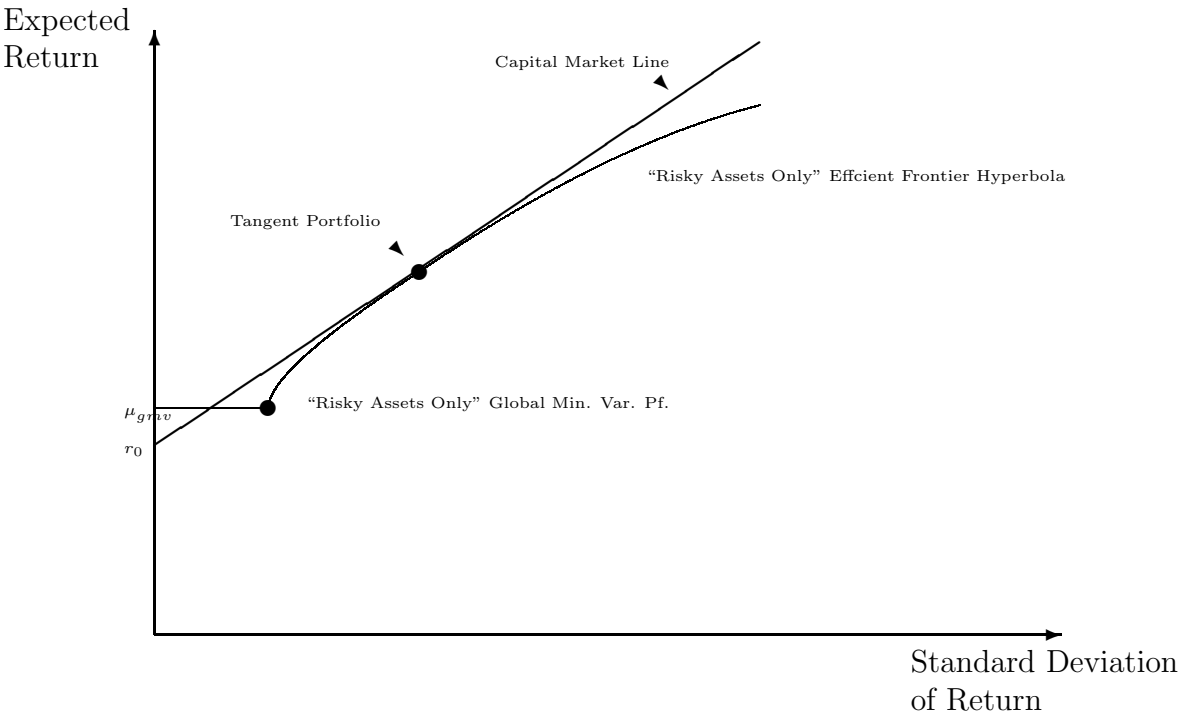


Figure 9.3: The capital market line.

and the following expression for the variance of the minimum variance portfolio with mean excess return r_P:

$$\widehat{\sigma}_P^2 = \frac{(r_P^e)^2}{(\boldsymbol{\mu}^e)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^e}. \tag{9.16}$$

So we have determined the efficient frontier. For required returns above the riskfree rate, the efficient frontier in standard deviation-mean space is a straight line passing through $(0, r_0)$ with a slope of $\sqrt{(\boldsymbol{\mu}^e)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^e}$. This line is called the capital market line.

The tangent portfolio, $\mathbf{x}$, is the minimum variance portfolio with all wealth invested in the risky assets, i.e. $\mathbf{x}_{tan}^\top \mathbf{1} = 1$. The mean excess return on the tangent portfolio is

$$r_{tan}^e = \frac{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}},$$

which may be positive or negative. It is economically plausible to assert that the riskless return is lower than the mean return of the global minimum

variance portfolio of the risky assets. In this case the situation is as illustrated in Figure 9.3, and that explains why we use the term "tangency". When $r^e_{tan} > 0$, the tangent portfolio is on the capital market line. But the tangent portfolio must also be on the "risky assets only" efficient frontier. So the straight line (the CML) and the hyperbola intersect at a point corresponding to the tangency portfolio. But clearly the CML must be above the efficient frontier hyperbola (we are minimizing variance with an extra asset). So the CML is a tangent to the hyperbola.

For any portfolio, P we define the *Sharpe-ratio* as excess return relative to standard deviation,

$$\text{Sharpe-ratio}_P = \frac{\mu_P - r_0}{\sigma_P}.$$

In the case where $r^e_{tan} > 0$, we see note from Figure 9.3 that the tangency portfolio is the "risky assets only"-portfolio with the highest Sharpe-ratio since the slope of the CML is the Sharpe-ratio of tangency portfolio. (Generally/"strictly algebraically" we should say that $\mathbf{x}_{tan}$ has maximal squared Sharpe-ratio.)

Note that a portfolio with full investment in the riskfree asset is orthogonal to any other portfolio; this means that we can prove the following result in exactly the manner as Proposition 32.

Proposition 33 *Let $\mathbf{x}_{mv}$ be a portfolio on the mean-variance frontier with rate of return r_{mv}, expected rate of return μ_{mv} and variance σ^2_{mv}. Let $\mathbf{x}_P$ be an arbitrary portfolio, and use similar notation for rates of return on these portfolios. Then the following holds:*

$$\mu_P - r_0 = \beta_{P,mv}(\mu_{mv} - r_0),$$

where

$$\beta_{P,mv} = \frac{Cov(r_P, r_{mv})}{\sigma^2_{mv}}.$$

9.2 The Capital Asset Pricing Model (CAPM)

With the machinery of portfolio optimization in place, we are ready to formulate one of the key results of modern finance theory, the CAPM-relation. Despite the clearly unrealistic assumptions on which the result is built it still provides invaluable intuition on what factors determine the price of assets in equilibrium. Note that until now, we have mainly been concerned with pricing (derivative) securities when taking prices of some basic securities as given.

Here we try to get more insight into what determines prices of securities to begin with.

We consider an economy with n risky assets and one riskless asset. Here, we let r_i denote the rate of return on the i'th risky asset and we let r_0 denote the riskless rate of return. We assume that r_0 is strictly smaller than the return of the global minimum variance portfolio.

Just as in the case of only risky assets one can show that with a riskless asset the expected return on any asset or portfolio can be expressed as a function of its beta with respect to an efficient portfolio. In particular, since the tangency portfolio is efficient we have

$$Er_i - r_0 = \beta_{i,tan}(Er_{tan} - r_0) \tag{9.17}$$

where

$$\beta_{i,tan} = \frac{Cov(r_i, r_{tan})}{\sigma^2_{tan}} \tag{9.18}$$

The critical component in deriving the CAPM is the identification of the tangency portfolio as the *market portfolio.* The market portfolio is defined as follows: Assume that the initial supply of risky asset j at time 0 has a value of P_0^j. (So P_0^j is the number of shares outstanding times the price per share.) The market portfolio of risky assets then has portfolio weights given as

$$w_j^m = \frac{P_j^0}{\sum_{i=1}^n P_i^0} \tag{9.19}$$

Note that it is quite reasonable to think of a portfolio with these weights as reflecting "the average of the stock market".

Now if all (say K) agents are mean-variance optimizers (given wealths of $W_i(0)$ to invest), we know that since there is a riskless asset they will hold a combination of the tangency portfolio and the riskless asset since two fund separation applies. Hence all agents must hold the same mix of risky assets as that of the tangency portfolio. This in turn means that in equilibrium where market clearing requires all the risky assets to be held, the market portfolio (which is a convex combination of the individual agents' portfolios) has the same mixture of assets as the tangency portfolio. Or in symbols: Let ϕ_i denote the fraction of his wealth that agent i has invested in the tangency portfolio. By summing over all agents we get

$$\begin{aligned} \text{Total value of asset } j &= \sum_{i=1}^K \phi_i W_i(0) \mathbf{x}_{tan}(j) \\ &= \mathbf{x}_{tan}(j) \times \text{Total value of all risky assets,} \end{aligned}$$

where we have used that market clearing condition that all risky assets must be held by the agents. (This is a very weak consequence of equilibrium; some would just call it an accounting identity. The main *economic assumption* is that agents are mean-variance optimizers so that two fund separation applies.) Hence we may as well write the market portfolio in equation (9.17). This is the CAPM:

$$Er_i - r_0 = \beta_{i,m}(Er_m - r_0) \tag{9.20}$$

where $\beta_{i,m}$ is defined using the market portfolio instead of the tangency portfolio. Note that the type of risk for which agents receive excess returns are those that are correlated with the market. The intuition is as follows: If an asset pays off a lot when the economy is wealthy (i.e. when the return of the market is high) that asset contributes wealth in states where the marginal utility of receiving extra wealth is small. Hence agents are not willing to pay very much for such an asset at time 0. Therefore, the asset has a high return. The opposite situation is also natural at least if one ever considered buying insurance: An asset which moves opposite the market has a high pay off in states where marginal utility of receiving extra wealth is high. Agents are willing to pay a lot for that at time 0 and therefore the asset has a low return. Indeed it is probably the case that agents are willing to accept a return on an insurance contract which is below zero. This gives the right intuition but the analogy with insurance is actually not completely accurate in that the risk one is trying to avoid by buying an insurance contract is not linked to market wide fluctuations.

Note that one could still view the result as a sort of relative pricing result in that we are pricing everything in relation to the given market portfolio. To make it more clear that there is an equilibrium type argument underlying it all, let us see how characteristics of agents help in determining the risk premium on the market portfolio. Consider the problem of agent i in the one period model. We assume that returns are multivariate normal and that the utility function is twice differentiable and concave[1]:

$$\begin{aligned} &\max_{\mathbf{w}} E(u_i(W_1^i)) \\ &\text{s.t.} W_1^i = W_0(\mathbf{w}^\top \mathbf{r} + (1 - \mathbf{w}^\top \mathbf{1})r_0) \end{aligned}$$

When forming the Lagrangian of this problem, we see that the first order condition for optimality is that for each asset j and each agent i we have

$$E\left(u_i'(W_1^i)(r_j - r_0)\right) = 0 \tag{9.21}$$

[1]This derivation follows Huang and Litzenberger: *Foundations for Financial Economics*

Remembering that $Cov(X,Y) = EXY - EXEY$ we rewrite this as

$$E\left(u_i'(W_1^i)\right) E(r_j - r_0) = -Cov(u_i'(W_1^i), r_j)$$

A nice lemma known as Stein's lemma says that for bivariate normal distribution (X,Y) we have

$$Cov(g(X), Y) = Eg'(X)Cov(X,Y)$$

and using this we have the following first order condition:

$$E\left(u_i'(W_1^i)\right) E(r_j - r_0) = -Eu_i''(W_1^i)Cov(W_1^i, r_j)$$

i.e.

$$\frac{-E\left(u_i'(W_1^i)\right) E(r_j - r_0)}{Eu_i''(W_1^i)} = Cov(W_1^i, r_j)$$

Now define the following measure of agent i's absolute risk aversion:

$$\theta_i := \frac{-Eu_i''(W_1^i)}{Eu_i'(W_1^i)}.$$

Then summing across all agents we have that

$$\begin{aligned} E(r_j - r_0) &= \frac{1}{\sum_{i=1}^K \frac{1}{\theta_i}} Cov(W_1, r_j) \\ &= \frac{1}{\sum_{i=1}^K \frac{1}{\theta_i}} W_0 Cov(r_m, r_j) \end{aligned}$$

where the total wealth at time 1 held in risky assets is $W_1 = \sum_{i=1}^K W_1^i$, W_0 is the total wealth in risky assets at time 0, and

$$r_m = \frac{W_1}{W_0} - 1$$

therefore is the return on the market portfolio. Note that this alternative representation tells us more about the risk premium as a function of the aggregate risk aversion across agents in the economy. By linearity we also get that

$$Er_m - r_0 = W_0^M Var(r_m) \frac{1}{\sum_{i=1}^K \frac{1}{\theta_i}},$$

which gives a statement as to the actual magnitude expected excess return on the market portfolio. A high θ_i corresponds to a high risk aversion and this contributes to making the risk premium larger, as expected. Note that if

one agent is very close to being risk neutral then the risk premium (holding that person's initial wealth constant) becomes close to zero. Can you explain why that makes sense?

The derivation of the CAPM when using returns is not completely clear in the sense that finding an equilibrium return does not separate out what is found exogenously and what is found endogenously. One should think of the equilibrium argument as determining the initial price of assets given assumptions on the distribution of the price of the assets at the end of the period. A sketch of how the equilibrium argument would run is as follows:

1. Given the expected value and the covariance of end of period asset prices for all assets

2. Given a utility function for each investor which depends only on mean and variance of end-of-period wealth. Assume that utility decreases as a function of variance and increases as a function of mean. Assume also sufficient differentiability

3. Let investor i have an initial fraction of the total endowment of risky asset j.

4. Assume that there is riskless lending and borrowing at a fixed rate r. Hence the interest rate is exogenous.

5. Given initial prices of all assets, agent i chooses portfolio weights on risky assets to maximize end of period utility. The difference in price between the initial endowment of risky assets and the chosen portfolio of risky assets is borrowed n/placed in the money market at the riskless rate. (In equilibrium where all assets are being held this implies zero net lending/borrowing.)

6. Compute the solution as a function of the initial prices.

7. Find a set of initial prices such that markets clear, i.e such that the sum of the agents positions in the risky assets sum up to the initial endowment of assets.

8. The prices will reflect characteristics of the agents' utility functions, just as we saw above.

9. Now it is possible to derive the CAPM relation by computing expected returns etc. using the endogenously determined initial prices. This is a purely mathematical exercise translating the formula for prices into formulas involving returns.

Hence CAPM is to be thought of as an equilibrium argument explaining asset prices.

There are of course many unrealistic assumptions underlying the CAPM. The distributional assumptions are clearly problematic. Even if basic securities like stocks were well approximated by normal distributions there is no hope that options would be well approximated due to their truncated payoffs. An answer to this problem is to go to continuous time modelling where 'local normality' holds for very broad classes of distributions but that is outside the scope of this course. Note also that a conclusion of CAPM is that all agents hold the same mixture of risky assets which casual inspection show is not the case. A final problem, originally raised by Roll (1977)[2], concerns the observability of the market portfolio and the logical equivalence between the statement that the market portfolio is efficient and the statement that the CAPM relation holds. To see that observability is a problem think for example of human capital. Economic agents face many decisions over a life time related to human capital - for example whether it is worth taking a loan to complete an education, weighing off leisure against additional work which may increase human capital etc. Many empirical studies use all traded stocks (and perhaps bonds) on an exchange as a proxy for the market portfolio but clearly this is at best an approximation. And what if the test of the CAPM relation is rejected using that portfolio? The relation At the intuitive level, the (9.17) tells us that this is equivalent to the inefficiency of the chosen portfolio. Hence one can always argue that the reason for rejection was not that the model is wrong but that the market portfolio is not chosen correctly (i.e. is not on the portfolio frontier). Therefore, it becomes very hard to truly test the model. While we are not going to elaborate on the enormous literature on testing the CAPM, note also that even at first glance it is not easy to test what is essentially a one period model. To get estimates of the fundamental parameters (variances, covariances, expected returns) one will have to assume that the model repeats itself over time, but when firms change the composition of their balance sheets they also change their betas.

Hence one needs somehow to accommodate betas which change over time and this inevitably requires some statistical compromises.

[2]R. Roll (1977): A critique of the asset pricing theory's test; Part I, Journal of Financial Economics, **4:**pp 129 - 76

9.3 Relevant, but not particularly structured, remarks on CAPM

9.3.1 Systematic and non-systematic risk

This section follows Huang and Litzenberger's Chapters 3 and 4. We have two versions of the capital asset pricing model. The most "popular" version, where we assumed the existence of a riskless asset whose return is r_0, states that the expected return on any asset satisfies

$$Er_i - r_0 = \beta_{i,m}(Er_m - r_0). \tag{9.22}$$

This version we derived in the previous section. The other version is the so-called zero-beta CAPM, which replaces the return on the riskless asset by the expected return on $m's$ zero-covariance portfolio:

$$Er_i - Er_{zm} = \beta_{i,m}(Er_m - Er_{zm}).$$

This version is proved by assuming mean-variance optimizing agents, using that two-fund separation then applies, which means that the market portfolio is on the mean-variance locus (note that we cannot talk about a tangent portfolio in the model with no riskfree asset) and using Proposition 32. Note that both relations state that excess returns (i.e. returns in addition to the riskless returns) are linear functions of β_{im}.

From now on we will work with the case in which a riskless asset exists, but it is easy to translate to the zero-beta version also. Dropping the expectations (and writing "error terms" instead) we have also seen that if the market portfolio m is efficient, the return on any portfolio (or asset) q satisfies

$$r_q = (1 - \beta_{q,m})r_f + \beta_{q,m}r_m + \epsilon_{q,m}$$

where

$$E\epsilon_{q,m} = E\epsilon_{q,m}r_m = 0.$$

Hence

$$\mathrm{Var}(r_q) = \beta_{q,m}^2 \mathrm{Var}\,(r_m) + \mathrm{Var}(\epsilon_{q,m}).$$

This decomposes the variance of the return on the portfolio q into its *systematic risk* $\beta_{qm}^2 \mathrm{Var}\,(r_m)$ and its *non-systematic* or *idiosyncratic* risk $\mathrm{Var}(\epsilon_{q,m})$. The reason behind this terminology is the following: We know that there exists a portfolio which has the same expected return as q but whose variance is $\beta_{qm}^2 \mathrm{Var}\,(r_m)$ - simply consider the portfolio which invests $1 - \beta_{qm}$ in the riskless asset and $\beta_{q,m}$, in the market portfolio. On the other hand, since this

portfolio is efficient, it is clear that we cannot obtain a lower variance if we want an expected return of Er_q. Hence this variance is a risk which is correlated with movements in the market portfolio and which is non-diversifiable, i.e. cannot be avoided if we want an expected return of Er_q. On the other hand as we have just seen the risk represented by the term VAR($\epsilon_{q,m}$) can be avoided simply by choosing a different portfolio which does a better job of diversification without changing expected return.

9.3.2 Problems in testing the CAPM

Like any model CAPM builds on simplifying assumptions. The model is popular nonetheless because of its strong conclusions. And it is interesting to try and figure out whether the simplifying assumptions on the behavior of individuals (homogeneous expectations) and on the institutional setup (no taxation, transactions costs) of trading are too unrealistic to give the model empirical relevance. What are some of the obvious problems in testing the model?

First, the model is a one period model. To produce estimates of mean returns and standard deviations, we need to observe years of price data. Can we make sure that the distribution of returns over several years remain the same[3]?

Second (and this a very important problem) what is the 'market portfolio' ? Since investments decisions of firms and individuals in real life are not restricted to stocks and bonds but include such things as real estate, education, insurance, paintings and stamp collections, we should include these assets as well, but prices on these assets are hard to get and some are not traded at all.

A person rejecting the CAPM could always be accused of not having chosen the market portfolio properly. However, note that if 'proper choice' of the market portfolio means choosing an efficient portfolio then this is mathematically equivalent to having the CAPM hold.

This point is the important element in what is sometimes referred to as Roll's critique of the CAPM. When discussing the CAPM it is important to remember which facts are mathematical properties of the portfolio frontier and which are behavioral assumptions. The key behavioral assumption of the CAPM is that the market portfolio is efficient. This assumption gives the CAPM-relation mathematically. Hence it is impossible to separate the claim 'the portfolio m is efficient' from the claim that 'CAPM holds with m acting as market portfolio'.

[3]Mulitiperiod versions exist, but they also face problems with time varying parameters.

9.3.3 Testing the efficiency of a given portfolio

Since the question of whether CAPM holds is intimately linked with the question of the efficiency of a certain portfolio it is natural to ask whether it is possible to devise a statistical test of the efficiency of a portfolio with respect to a collection of assets. If we knew expected returns and variances exactly, this would be a purely mathematical exercise. However, in practice parameters need to be estimated and the question then takes a more statistical twist: Given the properties of estimators of means and variances, can we reject at (say) a 5% level that a certain portfolio is efficient? Gibbons, Ross and Shanken (Econometrica 1989, 1121-1152) answer this question - and what follows here is a sketch of their test.

Given a portfolio m and N assets whose excess returns are recorded in T time periods. It is assumed that a sufficiently clear concept of riskless return can be defined so that we can really determine excess returns for each period. NOTE: We will change our notation in this section slightly and assume that r_p, Er_p and μ_p refer to *excess* returns, mean excess returns and estimated mean excess returns of an asset or portfolio p. Hence using this notation the CAPM with a riskless asset will read

$$Er_p = \beta_{p,m} Er_m.$$

We want to test this relation or equivalently whether m is an efficient portfolio in a market consisting of N assets. Consider the following statistical model for the excess returns of the assets given the excess return on the portfolio m :

$$r_{it} = \alpha_i + \gamma_i r_{mt} + \epsilon_{it}$$

$$i = 1, \ldots, N \text{ and } t = 1, \ldots, T$$

where r_{it} is the (random) *excess* return[4] of asset i in the t'th period, r_{mt} is the observed *excess* return on the portfolio in the t'th period, α_i, γ_i are constants and the ϵ_{it}'s are normally distributed with $Cov(\epsilon_{it}, \epsilon_{jt}) = \sigma_{ij}$ and $Cov(\epsilon_{it}, \epsilon_{is}) = 0$ for $t \neq s$. Given these data a natural statistical representation of the question of whether the portfolio m is efficient is the hypothesis that $\alpha_1 = \cdots = \alpha_N = 0$. This condition must hold for (9.22) to hold.

To test this is not difficult in principle (but there are some computational tricks involved which we will not discuss here): First compute the MLE's of the parameters. It turns out that in this model this is done merely by computing Ordinary Least Squares estimators for α, γ and the covariance

[4]Note this change to excess returns.

matrix for each period Σ. A so-called Wald test of the hypothesis $\alpha = 0$ can then be performed by considering the test statistic

$$W_0 = \widehat{\alpha} Var(\widehat{\alpha}) \widehat{\alpha}^{-1}$$

which you will learn more about in a course on econometrics. Here we simply note that the test statistic measures a distance of the estimated value of α from the origin. Normally, this type of statistics leads to an asymptotic chi squared test, but in this special model the distribution can be found explicitly and even more interesting from a finance perspective, it is shown in GRS that W_0 has the following form

$$W_0 = \frac{(T-N-1)}{N} \frac{\left(\frac{\widehat{\mu}_q^2}{\widehat{\sigma}_q^2} - \frac{\widehat{\mu}_m^2}{\widehat{\sigma}_m^2} \right)}{\left(1 + \frac{\widehat{\mu}_m^2}{\widehat{\sigma}_m^2} \right)}$$

where the symbols require a little explanation: In the minimum variance problem with a riskless asset we found that the excess return of any portfolio satisfies

$$Er_p = \beta_{pm} Er_m.$$

We refer to the quantity

$$\frac{Er_p}{\sigma(r_p)}$$

as the Sharpe ratio for portfolio p. The Sharpe ratio in words compares excess return to standard deviation. Note that using the CAPM relation we can write

$$\frac{Er_p}{\sigma(r_p)} = \frac{\sigma(r_m)\rho_{mp}}{\sigma^2(r_m)} (Er_m)$$

where ρ_{mp} is the correlation coefficient between the return of portfolios p and m. From this expression we see that the portfolio which maximizes the Sharpe ratio is (proportional) to m. Only portfolios with this Sharpe ratio are efficient. Now the test statistic W_0 compares two quantities: On one side, the maximal Sharpe ratio that can be obtained when using for parameters in the minimum variance problem the estimated covariance matrix and the estimated mean returns for the economy consisting of the N assets and the portfolio m. On the other side, the Sharpe ratio for the particular portfolio m (based on its estimated mean return and standard deviation).

Large values of W_0 will reject the hypothesis of efficiency and this corresponds to a case where the portfolio m has a very poor expected return per unit of standard deviation compared to what is obtained by using all assets.

Chapter 10

The APT model

10.1 Introduction

Although the APT stands for 'Arbitrage Pricing Theory' the model presented here is somewhat different from the arbitrage models presented earlier. The framework is in one sense closer to CAPM in that we consider a one-period model with risky assets whose distributions may be continuous. On the other hand, there is also a clear analogue with arbitrage pricing. We will present the basic idea of the model in two steps which will illustrate how the restriction on mean return arises.

10.2 Exact APT with no noise

Consider the following model for the returns $r_1, \ldots, r_N$ of N risky assets:

$$r = \mu + Bf$$

where $r = (r_1, \ldots, r_N)^\top$ is a vector of random returns, $\mu = (\mu_1, \ldots, \mu_N)^\top$, and B is an $N \times K-$matrix whose entries are real numbers, and $f = (f_1, \ldots, f_K)$ is a vector of random variables (*factors)* which satisfies

$$\begin{aligned} Ef_i &= 0, \qquad i = 1, \ldots, K, \\ Cov(f) &= \Phi, \qquad \Phi \text{ positive definite.} \end{aligned}$$

Note that this means that $Er_i = \mu_i$, $i = 1, \ldots, N$. We will assume that $N > K$, and you should think of the number of assets N as being much larger than the number of factors K. The model then seeks to capture the idea that returns on assets are correlated through a common dependence on a (small) number of factors. The goal is to use the assumption of such a

common dependence to say something about the vector of mean returns μ. Assume that there is also a riskless asset with return r_0. When we talk about a portfolio, w, we mean a vector in $\mathbb{R}^N$ where the ith coordinate measures the relative share of total wealth invested in the ith risky asset, and the rest in the riskfree asset. (So the term 'investment strategy given by w' would probably be better). Hence the coordinates need not sum to 1, and the expected rate of return on w is

$$\begin{aligned} E((1 - w^\top \mathbf{1})r_0 + w^\top r) &= r_0 + E(w^\top (r - r_0\mathbf{1})) \\ &= r_0 + w^\top(\mu - r_0\mathbf{1}), \end{aligned} \tag{10.1}$$

where as usual $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^N$.

Since $N > K$ it is possible to find a portfolio $w \in \mathbb{R}^N$ of risky assets which is orthogonal to the column space of B. This we will write as $w \in (B)^\perp$. Now the mean return is $r_0 + w^\top(\mu - \mathbf{1}r_0)$ and by using the "covariance matrix algebra rules" in the first part of Chapter 9 we see that the variance of the return on this portfolio is given as

$$\begin{aligned} V(w^\top r) &= Cov(w^\top Bf, w^\top Bf) \\ &= w^\top B\Phi B^\top w = 0. \end{aligned}$$

A reasonable no arbitrage condition to impose is that a portfolio consisting only of risky assets which has zero variance should earn the same return as the riskless asset. Hence the following implication should hold in an arbitrage free market:

$$w^\top \mathbf{1} = 1,\ w \in (B)^\perp : w^\top \mu = r_0 \Longleftrightarrow w^\top(\mu - r_0\mathbf{1}) = 0.$$

By scaling we see that

$$w^\top \mathbf{1} \neq 0,\ w \in (B)^\perp : w^\top(\mu - r_0\mathbf{1}) = 0.$$

By using the same "arbitrage reasoning" on (10.1) we get that

$$w^\top \mathbf{1} = 0,\ \ w \in (B)^\perp : w^\top \mu = 0 \Longleftrightarrow w^\top(\mu - r_0\mathbf{1}) = 0.$$

From these two statements we see that any vector which is orthogonal to the columns of B is also orthogonal to the vector $(\mu - r_0\mathbf{1})$, and this implies[1]

[1] If you prefer a mathematical statement, we are merely using the fact that

$$(B)^\perp \subset (\mu - r_0\mathbf{1})^\perp \Longrightarrow (\mu - r_0\mathbf{1}) \subset (B).$$

that $\mu - r_0\mathbf{1}$ is in the column span of B. In other words, there exists $\lambda = (\lambda_1, \ldots, \lambda_K)$ such that

$$\mu - r_0\mathbf{1} = B\lambda. \tag{10.2}$$

The vector $\lambda = (\lambda_1, \ldots, \lambda_K)$ is called the vector of *factor risk premia* and what the relation tells us is that the excess return is obtained by multiplying the *factor loadings* with the factor risk premia. This type of conclusion is of course very similar to the conclusion of CAPM, in which there is 'one factor' (return on the market portfolio), β_{im} plays the role of the factor loading and $Er_m - r_0$ is the factor risk premium.

10.3 Introducing noise

We continue the intuition building by considering a modification of the model above. Some of the reasoning here is heuristic - it will be made completely rigorous below.

Assume that

$$r = \mu + Bf + \epsilon$$

where r, μ, B and f are as above and $\epsilon = (\epsilon_1, \ldots, \epsilon_N)$ is a vector of random variables (*noise terms* or *idiosyncratic risks)* satisfying

$$\begin{aligned} E\epsilon_i &= 0, \qquad i = 1, \ldots, N, \\ \mathrm{Cov}(\epsilon_i, f_j) &= 0, \qquad i = 1, \ldots, N, \quad j = 1, \ldots, K, \\ \mathrm{Cov}(\epsilon) &= \sigma^2 I^N, \qquad I^N \text{ is the } N \times N \text{ identity matrix.} \end{aligned}$$

Clearly, this is a more realistic model since the returns are not completely decided by the common factors but 'company specific' deviations captured by the noise terms affect the returns also. However if the variance in the noise term is not too large then we can almost eliminate the variance arising from the noise term through diversification.

Since $N > K$ it is possible to find portfolios of risky assets $v_1, \ldots, v_{N-K}$ which are orthogonal and lie in $(B)^\perp$. Let a be the maximal absolute value of the individual portfolio weights. Now consider the portfolio

$$v = \frac{1}{N-K}(v_1 + \cdots + v_{N-K}).$$

The variance of the return of this portfolio is

$$\begin{aligned} \mathrm{V}(v^\top r) &= \mathrm{V}(v^\top Bf + v^\top \epsilon) \\ &= 0 + \frac{1}{(N-K)^2}\,\mathrm{V}((v_1 + \cdots + v_{N-K})^\top \epsilon) \end{aligned}$$

$$\begin{aligned} &\leq \frac{1}{(N-K)^2}(N-K)a^2\sigma^2 \\ &= \frac{a^2\sigma^2}{N-K}, \end{aligned}$$

where the inequality follows from the orthogonality of the v_i's (and the definition of a).

If we think of N as very large, this variance is very close to 0, and - entering into heuristic mode - therefore the expected return of this portfolio ought to be close to that of the riskless asset:

$$Ev^\top r = v^\top \mu \approx r_0. \tag{10.3}$$

By a slight modification of this argument it is possible to construct portfolios with good diversification which span $(B)^\perp$ and for each portfolio we derive a relation of the type (10.3). This then would lead us to expect that there exists factor risk premia such that

$$\mu - r_0\mathbf{1} \approx B\lambda.$$

The precise theorem will be given below.

10.4 Factor structure in a model with infinitely many assets

In this section we present a rigorous version of the APT.

Given is a riskless asset with return r_0 and an infinite number of risky assets with random returns $(r_1, r_2, \ldots)$. We will use the following notation repeatedly: If $x = (x_1, x_2, \ldots)$ is an infinite sequence of scalars or random variables, then x^N is the column vector consisting of the first N elements of this sequence. Hence $r^N = (r_1, \ldots, r_N)^\top$.

Definition 41 *The returns $(r_1, r_2, \ldots)$ are said to have an approximate factor structure with factors $(f_1, \ldots, f_K)$ if for all N*

$$r^N = \mu^N + B^N f + \epsilon^N$$

where B^N is the N first rows of a matrix B with infinitely many rows and K columns, where B^N has rank K for N large,

$$\begin{aligned} E\epsilon_i &= 0, \qquad i = 1, 2, \ldots, \\ \mathrm{Cov}(\epsilon_i, f_j) &= 0, \qquad i = 1, 2, \ldots, \quad j = 1, \ldots, K, \\ \mathrm{Cov}(f) &= I^K \qquad \textit{(the } K \times K \textit{ identity matrix)} \\ \mathrm{Cov}(\epsilon^N) &= \Omega^N \qquad (\Omega^N \textit{ is positive definite)} \end{aligned}$$

and where the eigenvalues of Ω^N *are bounded uniformly in* N *by a constant* $\overline{\lambda}$.

In other words, the same K factors are governing the returns on an infinite collection of securities except for noise terms captured by ϵ which however are uniformly of small variance. The simplest case would be where the elements of ϵ are independent[2] and have variance less than or equal to σ^2 in which case $\overline{\lambda} = \sigma^2$. Although our definition is slightly more general you can think of each element of ϵ as affecting only a finite number of returns and factors as affecting infinitely many of the returns.

The assumption that the covariance matrix of the factors is the identity may seem very restrictive. Note however, that if we have a structure of the form

$$r = \mu + Bf + \epsilon$$

which satisfies all the requirements of the definition of an approximate factor structure with the only exception being that

$$\mathrm{Cov}(f) = \Phi \qquad (\Phi \text{ is a positive definite, } K \times K\text{-matrix})$$

then using the representation $\Phi = CC^\top$ for some invertible $K \times K$ matrix we may choose g such that $Cg = f$. Then we have

$$r = \mu + BCg + \epsilon$$

and then this will be an approximate factor structure with g as factors and BC as factor loadings. To verify this note that

$$\mathrm{Cov}(g) = C^{-1}CC^\top(C^{-1})^\top = I^K.$$

Hence in one sense nothing is lost by assuming the particular structure of f. We may represent the same distribution of returns in this way as if we allow a general positive definite matrix to be the covariance matrix of the factors. However, from a statistical viewpoint the fact that different choices of parameters may produce the same distributions is a cause for alarm. This means that we must be careful in saying which parameters can be identified when estimating the model: Certainly, no observations can distinguish between parameters which produce the same distribution for the returns. We will not go further into these problems and to discussions of what restrictions can be imposed on parameters to ensure identification. We now need to introduce a modified notion of arbitrage:

[2]In this case the returns are said to have a *strict factor structure.*

Definition 42 *An asymptotic arbitrage opportunity is a sequence of portfolios* (w^N)*, where* $w^N \in \mathbb{R}^N$*, in the risky assets which satisfies*

$$\lim_{N\to\infty} E(w^N \cdot r^N) \;=\; \infty$$

and

$$\lim_{N\to\infty} \mathrm{V}(w^N \cdot r^N) = 0.$$

The requirement that expected return goes to infinity (and not just some constant greater than the riskless return) may seem too strong, but in the models we consider this will not make any difference.

The theorem we want to show, which was first stated by Ross and later proved in the way presented here by Huberman, is the following:

Theorem 34 *Given a riskless asset with return* $r_0 > -1$ *and an infinite number of risky assets with random returns* $(r_1, r_2, ...)$*. Assume that the returns have an approximate factor structure. If there is no asymptotic arbitrage then there exists a vector of factor risk premia* $(\lambda_1, \ldots, \lambda_K)$ *such that we have*

$$\sum_{i=1}^{\infty} (\mu_i - r_0 - \lambda_1 b_{i1} - \cdots - \lambda_K b_{iK})^2 < \infty. \tag{10.4}$$

This requires a few remarks: The content of the theorem is that the expected excess returns of the risky assets are in a sense close to satisfying the exact APT-relation (10.2): The sum of the squared deviations from the exact relationship is finite. Note that (unfortunately) this does not tell us much about the deviation of a particular asset. Indeed the mean return of an asset may show significant deviation from (10.2). This fact is crucial in understanding the discussion of whether the APT is a testable model.

The proof must somehow use the same arbitrage argument as in the case with exact APT above by getting rid of the noise terms through diversification. Although this sounds easy, we discover once again that 'the devil is in the details'. To do the proof we will need the following two technical lemmas:

Lemma 35 *Let* Ω *be a symmetric positive definite* $N \times N$ *matrix and let* $\overline{\lambda}$ *be its largest eigenvalue. Then*

$$w^\top \Omega w \leq \overline{\lambda} \left\| w \right\|^2.$$

Proof Let $(v_1, \ldots, v_N)$ be an orthonormal set of eigenvectors of Ω and $(\lambda_1, \ldots, \lambda_N)$ the corresponding eigenvalues. There exist $\alpha_1, \ldots, \alpha_N$ such that $w = \sum_1^N \alpha_i v_i$ and hence

$$\begin{aligned} w^\top \Omega w &= \sum_1^N \alpha_i^2 \lambda_i v_i^\top v_i \\ &= \sum_1^N \alpha_i^2 \lambda_i \\ &\leq \overline{\lambda} \sum_1^N \alpha_i^2 = \overline{\lambda} \left\| w \right\|^2 . \quad \blacksquare \end{aligned}$$

Lemma 36 *Let X be a compact set. Let $(K_i)_{i \in I}$ be a family of closed subsets of X which satisfy the finite intersection property*

$$\bigcap_{i \in I_0} K_i \neq \emptyset \qquad \text{for all finite subsets } I_0 \subset I.$$

Then the intersection of all sets is in fact non-empty. i.e.

$$\bigcap_{i \in I} K_i \neq \emptyset.$$

Proof If $\bigcap_{i \in I} K_i = \emptyset$, then $\bigcup_{i \in I} K_i^c$ is an open covering of X. Since X is compact the open cover contains a finite subcover $\bigcup_{i \in I_0} K_i^c$, but then we apparently have a finite set I_0 for which $\bigcap_{i \in I_0} K_i = \emptyset$, and this violates the assumption of the theorem. $\blacksquare$

The proof is in two stages. First we prove the following:

Proposition 37 *Given a riskless asset with return $r_0 > -1$ and an infinite number of risky assets with random returns $(r_1, r_2, ...)$. Assume that the returns have an approximate factor structure. If there is no asymptotic arbitrage then there exists a sequence of factor risk premia vectors (λ^N), $\lambda^N \in \mathbb{R}^K$, and a constant A such that for all N*

$$\sum_{i=1}^{\infty} (\mu_i - r_0 - \lambda_1^N b_{i1} - \cdots - \lambda_K^N b_{iK})^2 \leq A. \tag{10.5}$$

Proof Let us WLOG assume B^Nhas rank K for all N. Consider for each N the regression of the expected excess returns on the columns of B^N, i.e. the $\lambda^N \in \mathbb{R}^K$ that solves

$$\min_{\lambda}(\mu^N - r_0\mathbf{1} - B^N\lambda^N)^\top(\mu^N - r_0\mathbf{1} - B^N\lambda^N) = \min ||c^N||^2$$

where the *residuals* are defined by

$$c^N = \mu^N - r_0\mathbf{1}^N - B^N\gamma^N$$

By (matrix)-differentiating we get the first order conditions

$$(B^N)^\top c^N = 0$$

But the $K \times K$ matrix $(B^N)^\top B^N$ is invertible (by our rank K assumption), so the unique solution is

$$\lambda^N = ((B^N)^\top B^N)^{-1}(B^N)^\top(\mu^N - r_0\mathbf{1}).$$

We also note from the first order condition that the residuals c^N are orthogonal to the columns of B^N. To reach a contradiction, assume that there is no sequence of factor risk premia for which (10.5) holds. Then we must have $||c^N|| \to \infty$ (since $||c^N||^2$ *is* the left hand side of (10.5) with summation to N). Now consider the sequence of portfolios given by

$$w^N = \left\|c^N\right\|^{-\frac{3}{2}} c^N.$$

The expected excess return is given by

$$\begin{aligned} E(w^N \cdot (r^N - r_0\mathbf{1}^N)) &= E(w^N \cdot (\mu^N - r_0\mathbf{1}^N + B^N f + \epsilon^N)) \\ &= E(w^N \cdot (B^N\gamma^N + c^N + B^N f + \epsilon^N)) \\ &= \left\|c^N\right\|^{-\frac{3}{2}} c^N \cdot c^N \\ &= \left\|c^N\right\|^{\frac{1}{2}} \to \infty \text{ as } N \to \infty. \end{aligned}$$

where in the third equality we used the fact that c^N is orthogonal to the columns of B and that both factors and noise terms have expectation 0. The variance of the return on the sequence of portfolios is given by

$$\begin{aligned} \mathrm{V}(w^N \cdot (r^N - r_0\mathbf{1}^N)) &= \mathrm{V}(w^N \cdot (\mu^N - r_0\mathbf{1}^N + B^N f + \epsilon^N)) \\ &= \mathrm{V}(w^N \cdot \epsilon^N) \\ &= \left\|c^N\right\|^{-3} (c^N)^\top \Omega^N c^N \\ &\leq \left\|c^N\right\|^{-3} \overline{\lambda} \left\|c^N\right\|^2 \to 0 \text{ as } N \to \infty \end{aligned}$$

where we have used Lemma 35 and the same orthogonality relations as in the expected return calculations. Clearly, we have constructed an asymptotic arbitrage opportunity and we conclude that there exists a constant A and a sequence of factor risk premia such that

$$\sum_{i=1}^{\infty}(\mu_i - r_0 - \lambda_1^N b_{i1} - \cdots - \lambda_K^N b_{iK})^2 \leq A. \quad \blacksquare$$

Now we are ready to finish.

Proof of Theorem 34 Let A be as in the proposition above. Consider the sequence of sets (H^N) where

$$H^N = \left\{ \lambda \in \mathbb{R}^K : \sum_{i=1}^{N}(\mu_i - r_0 - \lambda_1 b_{i1} - \cdots - \lambda_K b_{iK})^2 \leq A \right\}.$$

By the preceding proposition, each H^N is non-empty and clearly $H^{N+1} \subset H^N$. Define the functions $f^N : \mathbb{R}^K \mapsto \mathbb{R}$ by

$$f^N(\lambda) = (\mu - r_0\mathbf{1} - B\lambda)^\top(\mu - r_0\mathbf{1} - B\lambda) = ||\mu - r_0\mathbf{1}||^2 + (\mu - r_0\mathbf{1})^\top B\lambda + \lambda^\top B^\top B\lambda,$$

where some of the N-superscripts have been dropped for the ease of notation. Then f^N is a convex function (because $B^\top B$ is always positive semidefinite), and we see that

$$H^N = \left\{ \lambda \in \mathbb{R}^K : f(\lambda) \leq A \right\}$$

is a closed convex set. Now pick an N so large that B has rank K. To show that H^N is then compact, it suffices (by convexity) to show that for all nonzero $\lambda \in H^N$ there exists a scaling factor (a real number) a such that $a\lambda \notin H^N$. But since B has full rank, there is no nonzero vector (in $\mathbb{R}^K$) that is orthogonal to all of B's (N) rows. Hence for an arbitrary nonzero $\lambda \in H^N$ we have that $||B\lambda|| \neq 0$ and

$$f^N(a\lambda) = ||\mu - r_0\mathbf{1}||^2 + a(\mu - r_0\mathbf{1})^\top B\lambda + a^2||B\lambda||^2,$$

so by choosing a large enough a we go outside H^N, so H^N is not compact. Then we may use Lemma 36 to conclude that

$$\bigcap_{N=1}^{\infty} H^N \neq \emptyset.$$

Any element $\lambda = (\lambda_1, \ldots, \lambda_K)^\top$ of this non-empty intersection will satisfy 10.4. $\blacksquare$

Chapter 11

On financial decisions of the firm

11.1 Introduction

One may think of decisions of firms as divided into two categories: *real* decisions and *financial* decisions. The real decisions focus on which projects the firms should undertake, the financial decisions deal with how the firm should raise money to undertake the desired projects. The area of *corporate finance* tries to explain the financial decisions of firms.

This chapter gives a very short introduction to the most basic issues in this area. The goal is to understand a couple of famous *irrelevance propositions* set forth by Modigliani and Miller stating conditions under which the firms financing decisions are in fact of no consequence. The conditions are in fact very restrictive but very useful since any discussion on optimality and rationality of financing decisions must start by relaxing one or several of these conditions.

We will consider only two types of securities: bonds and stocks. In reality there are many other types of securities (convertible bonds, callable bonds, warrants,...) and an important area of research (*security design)* seeks to explain why the different types of financing even exist. But we will have enough to do just learning the basic terminology and the reader will certainly see how to include more types of securities into the analysis.

Finally, it should be noted that a completely rigorous way of analyzing the firm's financing decisions requires general equilibrium theory - especially a setup with incomplete markets - but such a rigorous analysis will take far more time than we have in this introductory course.

11.2 'Undoing' the firm's financial decisions

At the heart of the irrelevance propositions are the investors ability to 'undo' the firm's financial decision: If a firm changes the payoff profile of its debt and equity, the investor can under restrictive assumptions change his portfolio and have an unchanged payoff of his investments. We illustrate all this in a one-period, finite state space model.

Given two dates 0 and 1 and a finite state space with S states. Assume that markets are complete and arbitrage free. Let p_s denote the price of an Arrow-Debreu security for state s, i.e. a security which pays 1 if the state at date 1 is s and 0 in all other states. Assume that an investment policy has been chosen by the firm which costs I_0 to initiate at time 0 and which delivers a state contingent payoff at time 1 given by the vector (i.e. random variable) $x = (x_1, \ldots, x_S)$.

The firm at date 0 may choose to finance its investment by issuing debt maturing at date 1 with face value D, and by issuing shares of stocks (equity). Assuming no bankruptcy costs, the payoff at the final date to equity and debt is given by the random variables

$$\begin{aligned} E_1 &= \max(x - D, 0) \\ B_1 &= \min(x, D) \end{aligned}$$

respectively. If we assume that there are N shares of stocks, the payoff to each stock is given by $S_1 = \frac{1}{N} E_1$. Note that the entire cash flow to the firm is distributed between debt and equity holders. If we define the value of the firm at time 0 as the value at time 0 of the cash flows generated at time 1 minus the investment I_0, it is clear that the value of the firm at time 0 is independent of the level of D. This statement is often presented as Modigliani-Miller theorem but as we have set it up here (and as it is often presented) it is not really a proposition but an assumption: The value of the firm is by assumption unaffected by D since by assumption the payoff on the investment is unaffected by the choice of D. As we shall see below, this changes for example when there are bankruptcy costs or taxes.

Consider two possible financing choices: One in which the firm chooses to be an all equity (*unlevered)* firm and have $D = 0,$ and one in which the firm chooses a level of $D > 0$ (a *levered* firm). We let superscript U denote quantities related to the unlevered firm and let superscript L refer to the case of a levered firm. By assumption $V^L = V^U$ since the assumption of leverage only results in a different distribution of the 'pie' consisting of the firm's cash flows, not a change in the pie's size. Now consider an agent who in his optimal portfolio in equilibrium wants to hold a position of one stock in the

unlevered firm. The payoff at date 1 of this security is given by

$$S_1^U = \frac{1}{N} E_1^U = \frac{1}{N} x.$$

If the firm decides to become a levered firm, the payoff of one stock in the firm becomes

$$S_1^L = \frac{1}{N} \max(x - D, 0)$$

which is clearly different from the unlevered case. However note the following: Holding $\frac{1}{N}$ shares in the levered firm *and* the fraction $\frac{1}{N}$ of the firms debt, produces a payoff equal to

$$\frac{1}{N} \max(x - D, 0) + \frac{1}{N} \min(x, D) = \frac{1}{N} x.$$

From this we see that even if the firm changes from an unlevered to a levered firm, the investor can adapt to his preferred payoff by changing his portfolio (something he can always do in a complete market). Similarly, we note from the algebra above that it is possible to create a position in the levered firm's stock by holding one share of unlevered stock and selling the fraction $\frac{1}{N}$ of the firm's debt. In other words, the investor is able to undo the firm's financial decision. In general equilibrium models this implies, that if there is an equilibrium in which the firm chooses no leverage, then there is also an equilibrium in which the firm chooses leverage and the investors choose portfolios to offset the change in the firm's financing decision. This means that the firm's capital structure remains unexplained in this case and more structure must be added to understand how a level of debt may be optimal in some sense.

It is important to note that something like complete markets is required and this is very restrictive. In real world terms, to imitate a levered stock in the firm, the investor must be able to borrow at the same conditions as the firm (highly unrealistic) and furthermore have the debt contract structured in such a way, that it imitates the payoffs of the firm when it is in bankruptcy.

We now consider another financing decision at time 0, namely the dividend decision. We consider for simplicity a firm which is all equity financed. To give this a somewhat more realistic setup, imagine that we are in fact considering the last period of a firm's life and that it carries with it an 'endowment' of cash W_0 from previous periods, which you can also think of as 'earnings' from previous activity. Also, imagine that the firm has N shares of stock outstanding initially. The value of the firm at time 0 is given by the value of the cash flows that the firm delivers to shareholders:

$$V_0 = Div_0 - \Delta E_0 + \sum_s p_s x_s$$

where Div_0 is the amount of dividends paid at time 0 to the shareholders and ΔE_0 is the amount of new shares issued (*repurchased* if negative) at time 0. It must be the case that

$$W_0 + \Delta E_0 = I_0 + Div_0$$

i.e. the initial wealth plus money raised by issuing new equity is used either for investment or dividend payout. If the firm's investment decision has been fixed at I_0 and W_0 is given, then $Div_0 - \Delta E_0 = W_0 - I_0$ is fixed, and substituting this into the equation for firm value tells us that firm value is independent of dividends when the investment decision is given. The dividend payment can be financed with issuing stocks. This result is also sometimes referred to as the Modigliani-Miller theorem.

But you might think that if the firm issues new stock to pay for a dividend payout, it dilutes the value of the old stocks and possibly causes a loss to the old shareholders. In the world with Arrow-Debreu prices this will not happen:

Consider a decision to issue new stocks to finance a dividend payment of Div_0. Assume for simplicity that $I_0 = W_0$. The number of stocks issued to raise Div_0 amount is given by M where

$$Div_0 = \frac{M}{N+M} \sum_s p_s x_s.$$

The total number of stocks outstanding after this operation is $M + N$ and the value that the old stockholders are left with is the sum of the dividend and the diluted value of the stocks, i.e.

$$\begin{aligned} & Div_0 + \frac{N}{N+M} \sum_s p_s x_s \\ = \; & \frac{M}{N+M} \sum_s p_s x_s + \frac{N}{N+M} \sum_s p_s x_s \\ = \; & \sum_s p_s x_s \end{aligned}$$

which is precisely the value before the equity financed dividend payout. This means that the agent who depends in his optimal portfolio choice on no dividends can undo the firm's decision to pay a dividend by taking the dividend and investing it in the firms equity. Similarly, if a dividend is desired at time 0 but the firm does not provide one, the investor can achieve it by selling the appropriate fraction of his stock position. The key observation is that as long as the value of the shareholders position is unchanged by the dividend

policy, the investor can use complete financial markets to design the desired cash flow.

A critical assumption for dividends to have no effect on the value of the firm and on the shareholder's wealth is that income from dividend payouts and income from share repurchases are taxed equally - something which is not true in many countries. If there is a lower tax on share repurchases it would be optimal for investors to receive no dividends and have any difference between W_0 and I_0 paid out by a share repurchase by the firm. Historically, one has observed dividend payouts even when there is lower taxes on share repurchases. Modigliani-Miller then tells us that something must be going on in the real world which is not captured by our model. The most important real world feature which is not captured by our model is asymmetric information. Our model assumes that everybody agrees on what the cash flows of the firm will be in each state in the future. In reality, there will almost always be insiders (managers and perhaps shareholders) who know more about the firm's prospects than outsiders (potential buyers of stocks, debtholders) and both the dividend policy and the leverage may then be used to signal to the outside world what the prospects of the firm really are. Changing the outsider's perception of the firm may then change the value of the firm.

11.3 Tax shield

If we change the model a little bit and assume that there are corporate taxes but that equity and debt financing are treated differently in the tax code then the capital structure becomes important. Change the model by assuming that the cash flows at time 1 are taxed at a rate of τ_c but that interest payments on debt can be deducted from taxable income. Let rD denote the part of the debt repayment which is regarded as 'interest'. The after tax cash flow of an unlevered firm at time 1 is given by

$$V_1^U = (1 - \tau_c)x$$

whereas the after tax cash flow of the levered firm (assuming full deduction of interest in all states) is given by

$$V_1^L = rD + (1 - \tau_c)(x - rD).$$

The difference in the cash flows is therefore

$$V_1^L - V_1^U = \tau_c rD$$

which means that the levered firm gets an increased value of

$$V_0^L - V_0^U = \tau_c r D \sum_s p_s.$$

As D increases, so does the value of this tax shield. Hence, in this setup financing the firm's operations with debt only would be optimal. But of course, it would be hard to convince tax authorities that a 100% debt financing was not actually a 100% equity financing! On the other hand, it should at least be the case that a significant fraction of debt was used for financing when there is a tax shield.

11.4 Bankruptcy costs

However, using a very high level of debt also increases the probability of bankruptcy. And it would add realism to our model if we assumed that when bankruptcy occurs, lawyers and accountants receive a significant fraction of the value left in the firm. This means that the total remaining value of the firm is no longer distributed to debtholders, and the debtholders will therefore have an interest in making sure that the level of debt issued by a firm is kept low enough to reduce the risk of bankruptcy to an acceptable level.

The trade-off between gains from leverage resulting from a tax shield and losses due to the increased likelihood of bankruptcy gives a first shot at defining an optimal capital structure. This is done in one of the exercises.

11.5 Financing positive NPV projects

We have seen earlier that in a world of certainty, one should only start a project if it has positive NPV. When uncertainty enters into our models the NPV criterion is still interesting but we need of course to define an appropriate concept of NPV. Both the arbitrage-free pricing models and the CAPM models gave us ways of defining present values of uncertain income streams.

We consider a one-period, finite state space model in which there is a complete, arbitrage-free market. Denote state prices (Arrow-Debreu prices) by $p = (p_1, \ldots, p_S)$. Assume that a firm initially (because of previous activity, say) has a net cash flow at time 1 given by the vector $x = (x_1 , \ldots, x_S)$. The firm is financed partly by equity and partly by debt, and firm value, equity

value and debt value at time 0 are therefore given as

$$\begin{aligned} V_0 &= \sum_s p_s x_s \\ E_0 &= \sum_s p_s (x_s - D)^+ \\ B_0 &= \sum_s p_s \min(x_s, D). \end{aligned}$$

We want to consider some issues of financing new investment projects in this very simple setup. We first note that projects with identical net present values may have very different effects on debt and equity. Indeed, a positive NPV project may have a negative effect on one of the two. This means that the ability to renegotiate the debt contract may be critical for the possibility of carrying out a positive NPV project.

Consider the following setup in which three projects a, b, c are given, all assumed to cost 1$ at time 0 to initiate, and with no possibility of scaling. Also shown is the cash flow x which requires no initial investment and the corresponding values of debt and equity when the debt has a face value of 30 :

	p	a	b	c	x	E_1	B_1
state 1	0.4	2	17	2	20	0	20
state 2	0.3	-20	0	4	40	10	30
state 3	0.3	30	-10	6	60	30	30
PV	-	3.8	3.8	3.8	38	12	26

Hence there are three projects all of which have an NPV of 2.8. Therefore, the increase in overall firm value will be 2.8. But how should the projects be financed? Throughout this chapter, we assume that all agents involved know and agree on all payouts and state prices. This is an important situation to analyze to develop a terminology and to get the 'competitive' situation straight first.

One possibility is to let the existing shareholders finance it out of their own pockets, i.e. pay the one dollar to initiate a project and do nothing about the terms of the debt: Here is what the value of equity looks like in the three cases at time 0:

	E_0^{new}	$E_0^{new} - E_0^{old}$	B_0^{new}	$B_0^{new} - B_0^{old}$
$x + a$	18	6	23.8	-2.2
$x + b$	11.8	-0.2	30	4
$x + c$	15	3	26.8	0.8

Clearly, only projects a and c will be attractive to the existing shareholders. Project a is however not something the bondholders would want carried through since it actually redistributes wealth over to the shareholders. What if a project is instead financed by issuing new stocks to other buyers? Let us check when this will be attractive to the old shareholders: This we can handle theoretically without actually working out the numbers:

To raise one dollar by issuing new equity, the new shareholders must acquire m shares, where m satisfies

$$\frac{m}{n+m}E_0^{new} = 1$$

and where n is the existing number of outstanding shares and E_0^{new} is the value of the equity after a new project has been carried out. Note that in this way new shareholders are by definition given a return on their investment consistent with the state prices. The old shareholders will be happy about the project as long as

$$\frac{n}{n+m}E_0^{new} - E_0^{old} > 0$$

i.e. as long as they have a capital gain on their shares. But this is equivalent to requiring that

$$\begin{aligned}\left(1 - \frac{m}{n+m}\right) E_0^{new} - E_0^{old} &> 0 \text{ i.e.} \\ E_0^{new} - E_0^{old} &> 1\end{aligned}$$

where we have used the definition of m to get to the last inequality. Note that this requirement is precisely the same as the one stating that in the case of financing by the old shareholders, the project should cost less to initiate than the capital gain. Note the similarity with the dividend irrelevance argument. In the argument we have just given, the shareholders decide whether to get a capital gain of $E_0^{new} - E_0^{old}$ and have a negative dividend of 1\$, whereas in the other case, the capital gain is $\frac{n}{n+m}E_0^{new} - E_0^{old}$ but there is no dividend.

Now consider debt financing. There are many ways one could imagine this happening: One way is to let the debtholders finance the projects by having so much added to face value D that the present value of debt increases by 1. This requires three very different face values:

	new face value	E_0^{new}	B_0^{new}
a	40.67	14.8	27
b	27	14.8	27
c	30,33	14.8	27

An interesting special case is the following: Assume that existing bondholders are not willing to do any renegotiation of the debt terms. One could

imagine for example, that the bondholders consisted of a large group of individuals who cannot easily be assembled to negotiate a new deal. Now, if project b were the only available project, then the shareholders would not enter into this project since the benefits of the project would go to the bondholders exclusively. This is the so-called *debt overhang* problem where it is impossible to finance a positive NPV project by issuing debt which is junior to the existing debt. To be able to carry through with project b, the shareholders would have to talk the debtholders into reducing the face value of the debt.

In general, it is easy to see that if a project has positive NPV there exists a way of financing the project which will benefit both debt holders and equity holders (can you show this?).

A special case which one often sees mentioned in textbooks is the case where the new project is of the same 'risk class' as the firm before entering into the project. This is true of project c. Such a project can always be financed by keeping the same debt-equity ratio after the financing as before. This is also left to the reader to show.

Chapter 12

Efficient Capital Markets

At the intuitive level, the efficient market hypothesis (EMH) states that it is impossible to "beat the market": The return you earn by investing in financial assets is proportional to the risk you assume. If certain assets had high returns compared to their risks, investors - who are constantly searching for and analyzing information about companies, commodities, economic indicators, etc. - would rush to buy these assets, pushing up prices until returns were "proportional" to the risk. In this way information about future returns of financial assets are quickly incorporated into prices which in other words fully and instantaneously reflect all available relevant information.

A first attempt to make this intuitive statement more rigorous is to elaborate a little bit on the definition of "relevant information" and to interpret 'reflecting information' as an inability to earn excessive returns. This leads us to the famous *degrees of efficiency:*

1. *Weak-form efficiency.* No investor can earn excess returns developing trading rules based on historical price or return information. In other words, the information in past prices or returns is not useful or relevant in achieving excess returns.

2. *Semistrong-form efficiency.* No investor can earn excess returns from trading rules based on any publicly available information. Examples of publicly available information are annual reports of companies, investment advisory data such as "Heard on the Street" in the Wall Street Journal, or ticker tape information.

3. *Strong-form efficiency.* No investor can earn excess returns using any information, whether publicly available or not.

Of course, we could define efficiency with respect to any information set I_t: No trader can earn excess returns between time t and $t+1$ given the information I_t.

Once we start thinking about this definition, however, we note that it is

still well short of being rigorous. What do we mean by "risk" and by return being "proportional to risk"? And if indeed prices (in some sense) reflect all available information why are banks and other financial institutions paying analysts to find information, and traders to 'look for arbitrage' in the market? In this note we will analyze the problem of defining market efficiency as follows:

First, we consider the problem of defining excess return. The only 'model free' definition of this concept would be an arbitrage opportunity. Other definitions are invariably linked to a particular model for security returns. If we are unable to find arbitrage opportunities can we then ever reject the hypothesis of 'market efficiency'?

Second, we consider some statistical properties of prices which are often thought of as being consequences of EMH. We try to give some rigorous definitions of these properties and (briefly) discuss if they are in any way necessarily linked to market efficiency.

Third, we look at anomalies and a very interesting challenge to market efficiency which attempts to show that stock prices are 'too volatile' in the sense that they vary much more than accounted for by changes in economic fundamentals.

Finally we mention some attempts to make the definition of market efficiency completely rigorous in a general equilibrium context.

12.1 Excess returns

First, we try to make "excess returns" a little more precise: As a first attempt we could say that for any security P it must be the case that

$$E\left[R_{t+1} \mid I_t\right] = 1 + r_t \tag{12.1}$$

where r_t is the riskless rate of return at time t and

$$R_{t+1} = \frac{P_{t+1}}{P_t}.$$

But having heard of CAPM and risk aversion we immediately object to the use of r_t: Assets with high risk (as measured e.g. through their β's) should (and do!) earn more than the riskless rate. So we restate (12.1) as

$$E\left[R_{t+1}|I_t\right] = 1 + r_t^P \tag{12.2}$$

where r_t^P is some return which is suitable for asset P. Note that we may rewrite (12.2) as

$$\frac{1}{1 + r_t^P} E\left[P_{t+1} \mid I_t\right] = P_t,$$

which states that the properly discounted price is a martingale. If markets are arbitrage free, such a discount factor will exist, so the requirement (12.2) is really nothing more than a statement of no arbitrage. In practice the problem is to find a good model for r_t^P - indeed we spent a lot of time looking at CAPM which tried to do just that. But if we have a model for r_t^P and we see violations of (12.2), it would seem more natural to reject our model for r_t^P than to reject market efficiency. Assume for instance that we use CAPM to model r_t^P. The literature often refers to a rejection of (12.2) as a rejection of "a joint test of CAPM and market efficiency". Given the very strong assumptions needed to derive CAPM, this is a somewhat strange statement (although it is of course logically OK). I would argue that unless there is a truly compelling and extremely realistic model (in terms of assumptions describing markets accurately) for r_t^P, we should consider rejections of (12.2) as rejections of our choice of asset pricing model and not worry about the market efficiency implications. Some proponents of efficient markets would say that it is precisely the belief in efficient markets which causes us to reject our model and look further for a rational explanation of stock price returns, and this is of course a valid point. But can we ever reach a situation in which all conceivable models for rational behavior have been rejected and we have no choice but to reject the EMH?

The closest to this situation would be a case where the model for r_t^P follows from a no arbitrage condition. If, for example, we considered a portfolio of a written call, a put (both European with same exercise price and date) and a stock (the underlying security for both options), we know from put-call-parity that its return should equal the return on a zero-coupon bond. Violations of this relation which are large enough that traders may take advantage of it should clearly not persist in well functioning security markets. In other words, no arbitrage is a necessary condition for markets to be efficient. And only this provides a test of (12.2) which is not linked to a particular choice of asset pricing model. Another way of stating this is that we can use (12.2) to test for violations of relative pricing relations which we derive from an assumption of no arbitrage. Violations would seem to indicate an inefficiency of markets.

But to use (12.2) as a definition of efficiency is complicated when we are looking at "absolute" quantities like stock prices. We will never be able to reject efficiency from this alone - critics will always (justly) be able to argue that our choice of asset pricing model is wrong. A famous paper by DeBondt and Thaler[1] [1986] illustrates the problem well. They did the following:

- Record the returns of a very large number of stocks over a period of

[1]See Journal of Finance, July 1986, pp. 793-807.

Figure 12.1: The cumulative excess return (compared to the market) for the "winner" and "loser" portfolios.

(say) 36 months.

- At the end of the period, form one portfolio of "winners" i.e. pick (say) the 35 best performing stocks of the 36 month period.
- Form another portfolio of "losers" - i.e. pick (say) the 35 worst performing (but still alive!) stocks of the 36 month period.
- Compare the returns to that of the market.

The results are illustrated in Figure 12.1, which shows cumulative excess returns compared to the market for the two portfolios. We see that the loser portfolio has a (quite large) positive return over the market, whereas "the winners" are performing badly. Now we could be tempted to see this as a clear violation of "weak-form efficiency" because, intuitively, we are beating the market using only information of past prices. But are we defining "excess return" correctly then? We have seen in the exercises (in a discrete binomial version) that the beta of an option on a stock may be related to the beta of the stock. The equivalent of this in a Black-Scholes model where we model stocks as options on an underlying firm value gives us that

$$\beta_S = N(d_1)\frac{V}{S}\beta_V.$$

A "loser" stock will typically have experienced an increase in leverage. This in turn implies an increase in β_S and therefore the increased return on a "loser" follows directly from option pricing theory and CAPM. In other words, a more careful selection of asset pricing model at least gave us a possible explanation for the results of DeBondt and Thaler. Subsequent research (see for example Chopra, Lakonishok and Ritter[2] [1992]) seems to indicate that correcting for leverage effects does not destroy the result. But the controversy is far from over and the key issue is whether more advanced asset pricing models can be shown to explain the larger returns on losing stocks. Note, however, that the controversy illustrates the sense in which (12.2) is a "hard-to-test" definition of efficiency. We will have to reject it for all conceivable (and reasonable) asset pricing models before an inefficiency is established. This seems to be a tough task, especially considering results of Summers[3] [1986] which indicate that many test traditionally employed do not have very large power (i.e. ability to reject a false hypothesis).

12.2 Martingales, random walks and independent increments

As a motivation for this section consider the footnote in Brealey and Myers:

"When economists speak of stock prices as following a random walk, they are being a little imprecise. A statistician reserves the term *random walk* to describe a series that has a constant expected change each period and a constant degree of variability. But market efficiency does not imply that expected risks and expected returns cannot shift over time."

Clearly, we need to be more precise even about the statistician's definition if we are to discuss the notions precisely. The purpose of this section is to give precise content to such concepts as random walk, independent increments and martingales.

Definition 43 *The stochastic process $X = (X_1, X_2, \ldots)$ is a random walk if it has the form*

$$X_t = \sum_{i=1}^{t} \epsilon_i$$

where $\epsilon_1, \epsilon_2, \ldots$ is a sequence of independent, identically distributed random variables.

[2]Journal of Financial Economics, 1992, pp. 235-268

[3]Journal of Finance, July 1986, pp.591-601

We will normally assume that ϵ_i has finite variance (and hence finite expectation as well). If $E\epsilon_i = 0$ we say that the random walk is symmetric.

A weaker concept is that of a process with independent increments:

Definition 44 *The stochastic process $X = (X_1, X_2, \ldots)$ is a process with independent increments if for all t, $\Delta X_t \equiv X_t - X_{t-1}$ is independent of $(X_1, X_2, \ldots, X_{t-1})$.*

Note that a random walk is a process with independent increments, but a process with independent increments may have the distribution of the increment change with time. Weaker yet is the notion of orthogonal increments where we only require increments to be uncorrelated:

Definition 45 *The stochastic process $X = (X_1, X_2, \ldots)$ is a process with orthogonal increments if for all t , $EX_t^2 < \infty$ and for all t, u*

$$\mathrm{Cov}(\Delta X_t, \Delta X_u) = 0.$$

Finally a somewhat different (but as we have seen equally important) concept

Definition 46 *The stochastic process $X = (X_1, X_2, \ldots)$ is a martingale with respect to the filtration $(\mathcal{F}_t)$ if for all t , $E\,|X_t| < \infty$ and*

$$E(X_t\,|\mathcal{F}_{t-1}) = X_{t-1}.$$

If no filtration is specified it is understood that $\mathcal{F}_{t-1} = \sigma(X_1, X_2, \ldots, X_{t-1})$.

A symmetric random walk is a martingale and so is a process with mean zero independent increments. A martingale with finite second moments has orthogonal increments, but it need not have independent increments: Think for example of the following ARCH process:

$$X_t = \epsilon_t \sigma_t$$

where $\epsilon_1, \epsilon_2, \ldots$ is a sequence of i.i.d. $N(0,1)$ random variables and

$$\sigma_t^2 = \alpha_0 + \alpha_1 x_{t-1}^2 \qquad \alpha_0, \alpha_1 > 0$$

where x_{t-1} is the observed value of X_{t-1}. The abbreviation ARCH stands for Autoregressive Conditional Heteroskedasticity which means (briefly stated) that the conditional variance of X_t given the process up to time $t-1$ depends on the process up to time $t-1$.

It is clear from observation of market data, that stock prices have a drift and they are therefore not martingales. However it could be that by an appropriate choice of discount factor, as in (12.2), we get the martingale property. In fact, we know that this will be possible in an arbitrage free market (and for all types of financial securities), but then we are back to the situation discussed above in which the specification of the pricing relation becomes essential.

The independent increment property is often heard as a necessary condition for efficient markets, but research over the last decade has pointed to ARCH effects in stock prices, exchange rates and other financial securities.[4]

Another serious problem with all the properties listed above is that it is perfectly possible to construct sensible general equilibrium models (which in particular are arbitrage free) in which prices and returns are serially correlated (for example because aggregate consumption is serially correlated). Therefore, it seems unnatural to claim that random walks, independent increments and martingale properties of prices are in any way logical consequences of a definition of efficient markets. The most compelling consequence of an efficiency assumption of no arbitrage is that there exists an equivalent measure under which securities are martingales but this does not in general say much about how processes behave in the real world. Indeed, even in the case of futures contracts which would be very natural candidates to being martingales, we have seen that in models where agents are risk-averse, the equivalent martingale measure will in general be different from the empirically measure and hence the martingale property of futures prices is by no means a necessary condition for efficiency.

Also, note that bonds for example have predetermined 'final' value and therefore it is not at all clear that any of the above mentioned properties of price processes are even relevant for this class of securities.

12.3 Anomalies

Another way to challenge the EMH is to look for strange patterns in security prices which seem impossible to explain with any pricing model. Some such anomalies are *week-end effects* and *year-end effects:*

Evidence found in French (1980)[5] seems to suggest that returns on Mondays were significantly negative, compared to returns on other trading days. Although the effect is difficult to exploit due to transactions costs, it is still unclear why the effect exists. Possible explanations have tried to look at

[4]For an collection of papers see Engle: *ARCH.*Oxford University Press, 1995.

[5]Journal of Financial Economics, 1980, pp. 55-69.

whether there is a tendency for firms to release bad news on Friday afternoons after trading closes. But if this were the case traders should learn this an adjust prices accordingly earlier. The week-end effect is hard to explain - after all it does seem to suggest that traders who are going to trade anyway should try to sell on Fridays and buy on Mondays.

Another effect documented in several studies is the year-end effect which shows that stocks have a tendency to fall in December and rise in January. An obvious reason for such behavior could be tax considerations but it is still unclear if this explanation is sufficient or whether there is actually an effect which trading rules can take advantage of.

A famous anomaly is the *closed-end mutual fund discount.* A closed-end fund is a mutual fund which holds publicly traded assets, but which only issues a fixed number of shares and where shareholders can only sell their shares in the market. Since the assets are so easy to determine, so should the share price be. However, throughout a long period of time mutual funds seemed to sell at a discount: The values of the funds' assets seemed larger than the market value of shares. Small discounts could be explained by tax-considerations and illiquidity of markets, but no theory could explain a pattern (shown in Figure 12.2) like the one observed for the Tricontinental Corporation[6] during 1960-1986.

It seems that finally in 1986 prices adjusted, but that prices should "instantaneously" have reflected fundamental information seems implausible.

So a newer version of EMH stated in Malkiel [1990][7] could be that "pricing irregularities may well exist and even persist for periods of time, but the financial laws of gravity will eventually take hold and true value will come out." Certainly a weakening of our original version of efficiency.

12.4 Excess volatility.

One of the most interesting challenges to EMH argues that there is too much volatility in stock markets - more than can be explained by *any* sensible asset pricing theory. If indeed markets reflect all relevant information it should be the case that price movements and arrival of new information were somehow in harmony. Trading alone ought not to generate volatility. Several pieces of evidence suggest that trading índeed creates volatility:

1. October 19, 1987. The Dow Jones Industrial Average fell by more than 22% - a much bigger drop than any previous one-day movement. Yet extensive surveys among investment managers seem to suggest that no important

[6] See Lee, Shleifer and Thaler, Journal of Finance 1991, pp.75-109.

[7] *A Random Walk Down Wall Street.*

Figure 12.2: Percentage discount (= 100* (net asset value per share - share price of the fund)) at which Tricontinental Corporation was traded 1960-86.

news related to stock prices arrived that day.

2. In the second half of 1986 the stock exchange in New York closed for a series of Wednesdays to catch up on paperwork. Volatility between Tuesday's close and Thursday's opening of trade was significantly smaller than when the exchange was open on Wednesdays. As Thaler [1993][8] puts it, traders react to each others as well as to news.

3. A famous study on "Orange juice and weather" by Roll [1984][9] suggests that surprises in weather forecasts for the Florida area, where 98% of oranges used for orange juice are traded, are too small to explain variations in the futures price (whose main concern is the likelihood of a freeze)

But perhaps the most controversial attack on EMH is the one put forward by Shiller, and Leroy and Porter. Their line of analysis is that what determines stock prices must ultimately be some combination of dividends and earnings of the firm. Since stock price is an expected, discounted value of future quantities, the price should fluctuate less than the quantities themselves.

To understand their line of reasoning, consider the following example: Assume that the quantity x_t (which could be dividends) determines the stock

[8]R. Thaler (ed.) *Advances in Behavioral Finance,* Russell Sage Foundation, NY 1993

[9]American Economic Review, December 1984, pp. 861-880.

price through a present value relation of the form:

$$Y_t = \sum_{j=0}^{\infty} \beta^j X_t^e(j),$$

where β is some discount factor, $\beta \in]0,1[$, and

$$X_t^e(j) = E[X_{t+j}|I_t],$$

where I_t can - for simplicity - be thought of as the information generated by the process X up to time t. If we assume that X_t follows an autoregressive process of order one, i.e.

$$X_t - c = \phi(X_{t-1} - c) + \varepsilon_t$$

where $c > 0, |\phi| < 1$ and ε_t's are independent normally distributed random variables, then we can calculate the *coefficient of dispersion* of X_t,

$$CD(X_t) = \frac{\sigma(X_t)}{E(X_t)}$$

and the similar quantity for Y_t. We are interested in $\frac{CD(Y_t)}{CD(X_t)}$. Using a fact from time series analysis, which says that X_t be represented in the form

$$X_t - c = \sum_{j=0}^{\infty} \varphi^j(\varepsilon_{t-j})$$

we find

$$EX_t = c \quad VX_t = \frac{1}{1-\varphi^2}$$

and

$$EY_t = \frac{c}{1-\beta} \quad VY_t = \frac{VX_t}{(1-\beta\varphi)^2}$$

and hence

$$\frac{CD(Y_t)}{CD(X_t)} = \frac{1-\beta}{1-\beta\varphi} < 1$$

by our choice of parameters. In other words, the coefficient of dispersion for actual prices is less than that of X_t (dividends, earnings). What Shiller, and Leroy and Porter show is that this result holds for a large class of processes for X *and* that the relationship is severely violated for observed data.

Figure 12.3: Variability of a detrended S&P index (full curve) compared to the variability of dividends (dotted line).

A useful bound also derived in these authors' work is the variance bound on the perfect foresight price

$$P_t^* = \sum_{k=0}^{\infty} X_{t+k} \prod_{j=0}^{k} \gamma_{t+j}$$

(where γ_{t+j} is the discount factor between time j and $j+1$ as recorded at time t) which "knows" all the paid dividends, and the actual price

$$P_t = E_t P_t^*$$

Shiller argues that the bound $V(P_t) \leq V(P_t^*)$ is grossly violated in practice, as illustrated in Figure 12.3.

Marsh and Merton [1986] criticize the work of Shiller and argue that with non-stationary dividend policies, the variance bound relations cannot hold and should in fact be reversed. Future versions of this chapter will discuss this controversy!

12.5 Informationally efficient markets are impossible

The title of this section is the title of a paper by Grossman and Stiglitz[10] (1980) Their first paragraph in that work sums up an essential problem with the assumption of informational efficiency:

"If competitive equilibrium is defined as a situation in which prices are such that all arbitrage profits are eliminated, is it possible that a competitive economy can always be in equilibrium? Clearly not, for then those who arbitrage make no (private) return from their (privately) costly activity. Hence the assumption that all markets, including that for information, are always in equilibrium and always perfectly arbitraged are inconsistent when arbitrage is costly." Grossman and Stiglitz go on to build a model in which gathering (costly) information is part of an equilibrium among informed and uninformed traders.Without going into details of their model, we note here that in equilibrium, prices reflect some but not all of the information. There is a fraction of traders who are informed and spend money to gather information, and a fraction of traders who are uninformed but try to learn as much as possible by observing prices (which they know reflect partly the knowledge that the informed traders have). If some of the informed traders give up information gathering, prices will reflect less information and there will be an incentive for non-informed traders to become informed because the costs of information gathering will more than compensated for by gains from trade. If some uninformed traders in the equilibrium situation decide to become informed prices will reflect more information and there will in fact be an incentive for some of the informed to give up information gathering and 'free-ride' on the information gathered by others which is reflected in prices. Clearly, a situation in which all information is reflected in prices is impossible.

This paper has a number of interesting results that we will not go into here. Suffice it to say that it contains one of the most interesting attempts to give a formal definition of efficiency which does not have the inherent logical problems of the original definitions.

There are numerous other more rigorous attempts to define efficiency. A typical line of approach is to model a situation with asymmetric information and ask whether the equilibrium price (and possibly the portfolio holdings of individuals) would be altered if all of the information held by traders was given to all traders simultaneously. If prices and allocations would not change in this situation, then it would be fair to say that prices reflected all

[10] American Economic Review, vol 70, pp. 393-408.

information.

These models will be discussed in a future version of the notes.