Finance 1 and Beyond

David Lando
Rolf Poulsen

October 10, 2023

Contents

Chapter 1

We hold these truths not to be self-evident: Aim and scope

It is a truth universally acknowledged that *there is no such thing as a free lunch.* When applied to financial markets – under the label “no arbitrage” – this simple, prudent principle turns out to have quite profound consequences. As we shall see.

This manuscript covers the basic pillars of quantitative finance:

1. Analysis of deterministic cash-flows; Chapter 2.
2. Mean-variance analysis and the capital asset pricing model (CAPM); Chapter 3.
3. Construction of and valuation – particularly of options – in stochastic multi-period models; Chapters 4-6.

Additionally, we present – in a far from exhaustive way – extensions of the basic material: Chapter 7 looks at the continuous-time Black-Scholes model and formula, Chapter 8 analyses stochastic interest rate models.

Our aim is to be mathematically precise without abandoning neither the economic intuition — such intuition is hard work, not just hand-waving — nor the reason d’etre of quantitative finance which is to do well-founded calculations with sensible inputs – and hopefully outputs. We are men and women of letters *and* of numbers.

The notes are not littered with references to books and research papers; self-contained and timeless are positive terms to decribe the intended format. But we must mention two text-books that cover roughly the same material and from which we have learned a lot: John Hull’s ubiquitous *Options, Futures and Other Derivative Securities* and David Luenberger’s lesser known *Investment Science.*

We are deeply endebted to Simon Ellersgaard Nielsen for his help with the preparation of this manuscript.

Chapter 2

No uncertain terms: Analyzing deterministic payment streams

In this chapter we consider a simple setup where there is no uncertainty; payment streams are deterministic. A non-exhaustive list of reasons for this is:

1. The financial concepts and mathematical techniques introduced will be extremely useful later. For this reason we start somewhat abstractly.

2. The terminology of bond markets is conveniently introduced in this setting. Even if there were uncertainty in our model, bonds would be characterized by having payments whose size at any date are known in advance.

3. The classical techniques such as the net present value (NPV) rule of capital budgeting and bond portfolio immunization are easily understood in this framework – including their non-triviality and limitations.

The list above also describes how the material is ordered.

2.1 Abstract financial markets, absence of arbitrage, and the fundamental theorems

We will make heavy use of vectors and matrices; throughout the notes and in particular in this subsection. By $v^\top$ we denote the transpose of the vector v and vectors without the transpose sign are always thought of as column vectors. We use the following conventions for positivity of a vector $v \in \mathbb{R}^N$:

- $v \geq 0$ ("v is non-negative") means that all of v's coordinates are non-negative. ie. $\forall i$: $v_i \geq 0$.

- $v > 0$ ("v is positive") means that $v \geq 0$ and that at least one coordinate is strictly positive, ie. $\forall i$: $v_i \geq 0$ and $\exists i$: $v_i > 0$, or differently that $v \geq 0$ and $v \neq 0$.

- $v \gg 0$ ("v is strictly positive") means that every coordinate is strictly positive, $\forall i$: $v_i > 0$. We will sometimes write as $v \in \mathbb{R}^N_{++}$. This saves a bit of space when we want to indicate both strict positivity and the dimension of v.

We now consider a model for a financial market (sometimes also called a security market or price system; individual components are then referred to as securities) with $T + 1$ dates: $0, 1, \ldots, T$ and no uncertainty.

Definition 1. *A* financial 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 $(c_{i,1}, \ldots, c_{i,T})$ at dates $1, \ldots, T$. Negative components are interpreted as amounts that the owner of the security has to pay. These N payment streams can be bought and sold generally combined; we use the concept of a portfolio to handle this.

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* $\pi \cdot \theta = \pi^\top \theta = \theta^\top \pi$.

The coordinate θ_i is interpreted as the number of units of security i that we buy at time 0. To see why the rest of the definition is then financially reasonable, let us write out the time t payment of the portfolio using the definition of transposition and matrix multiplication:

$$(C^\top \theta)_t = \sum_{i=1}^N C_{i,t} \theta_i = \sum_{i=1}^N \theta_i \pi_i.$$

So the total payment is how much one unit of a specific security pays multiplied how many units of that security we hold, summed over all securities. It is bookkeeping via matrix-notation. Be aware that payment matrices are generally not quadratic and very rarely symmetric, so it is important to remember transpositions; generally $C^\top \theta \neq C\theta$, even in the quadratic case ($N = T$). 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. Short positions are not just a convenient mathematical abstraction. For instance when you borrow money to buy a home, you take a short position in bonds.

Before using (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 some equilibrium model in which the agents use the market to transfer wealth between periods, we clearly want a payment stream of $(1, \ldots, 1)$ to have a lower price than that of $(2, \ldots, 2)$. 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*

$$\begin{pmatrix} -\pi \cdot \theta \\ C^\top \theta \end{pmatrix} > 0.$$

In proofs we may have to treat the case "$\pi \cdot \theta = 0$, $C^\top \theta > 0$" (called weak or type 1 arbitrage) slightly separately from the case "$\pi \cdot \theta < 0$, $C^\top \theta \geq 0$" (strong or type 2 arbitrage). In most – but not all – models, one type can be turned into the other. In models with uncertainty, more sub-categories of arbitrage can come into play. It is important to understand how attractive a portfolio an arbitrage is. It is not just a favorable bet or a sensible investment, it is a money machine, it is a "free lunch", it is the finance version of alchemy. The next example illustrates.

Example 1. Consider the following *odds* that a number of internet bookmakers put on the 2004 African Nation's Cup match between Burkina Faso and Mali.

	Burkina Faso - Mali		
Bookmaker	1	X	2
	(B F win)	(draw)	(Mali win)
`Aebet`	5.50	3.10	1.61
`Bet-at-home.com`	3.65	3.20	1.75
`EasyBets`	4.20	3.30	1.73
`Expekt`	4.05	3.15	1.85
`InterWetten`	3.50	2.80	2.00
`MrBookmaker`	4.60	3.05	1.73

These are so-called decimal odds; betting \$1 and winning gives you \$1 $\cdot$ odds back; betting on the wrong outcome means you lose your \$1 stake. Now imagine that we pick the best odds for each outcome and bet \$ $1/5.5 = 0.1818$ on Burkina Faso, \$$1/3.3 = 0.3030$ on a draw, and \$$1/2.0 = 0.5$ on Mali. The total cost of this is \$0.9848. Irrespective of what happens we win \$1. This makes money out of thin air.
How does this relate to our definitions of market, portfolios and arbitrage? First, mentally change "future date" to "possible future outcome". In that framework portfolio vectors contain the amounts we bet (on different outcomes and with different bookmakers) and each bookmaker offers three securities; the price vector consists purely of ones and the payment-matrix is diagonal with the bookmaker's odds along the diagonal. Taking this model literally, any two different odds on the same outcome constitute an arbitrage. But that would involve a short position, and in practice we can't force bookmakers to bet with us. (So-called betting exchanges do exist, but that is a different and longer story.) However, if we assume that a risk-free asset exists and (not that unreasonably) that we can take short positions in it (i.e. borrow money), then the situation above is one where we can (unless the interest rate is prohibitively high) construct an arbitrage with only long positions against the bookmakers. ■

The example above notwithstanding, a prudent financial assumption is that markets do not contain arbitrages.

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

If arbitrages do exist, then we would love to find them. The way to do that, however, is to study closely the consequences of absence of arbitrage. If they are violated, then there must be arbitrage, and our means of analysis give us constructive ways to find it/them. It turns out that there is a simple characterization of arbitrage-free markets.

Theorem 1. *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 $\pi = Cd$.*

We will refer to results of this type as *first fundamental theorems.* In the context of the models in this chapter the vector d will be referred to as a vector of discount factors, or simply a discount vector – for reasons that will become clear before too long.

Proof. Let us first prove the "if" part, which is an easy implication. So assume that d is a discount vector, and suppose contrarily that θ is an arbitrage. If θ is a type 2 arbitrage, we have $\pi^\top \theta < 0$. Because $\pi = Cd$, we also have that $\pi^\top \theta = d^\top C^\top \theta$. But as $d >> 0$ and $C^\top \theta \geq 0$, the strict negativity is a contradiction. If θ is type 1 arbitrage, we know that $C\theta > 0$, and since $d >> 0$ similar reasoning again leads to a contradiction. Hence, no arbitrages can exist.
The "only if part" is more difficult and makes use of concepts and results from convex analysis; Appendix B is a crash course in that subject. Let us define the matrix

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

and the two sets

$$C_1 = \{y \in \mathbb{R}^{T+1} | \exists z \in \mathbb{R}^N \text{ such that } y = A^\top z\}$$

and

$$C_2 = \{y \in \mathbb{R}^{T+1} | \sum_{t=1}^{T+1} y_t = 1, \; y_t \geq 0 \text{ for all } t\}.$$

Note that C_1 is a closed convex set and C_2 is a compact convex set. If there is no arbitrage then the two sets are disjoint. Hence by a strongly separating hyperplane theorem, Theorem 13 in Appendix B, there exists a vector $x \in \mathbb{R}^{T+1}$ such that

$$\inf_{y \in C_2} x^\top y > \sup_{y \in C_1} x^\top y.$$

Because C_1 is in fact a linear space, the supremum on the right hand side must be 0; otherwise we would get a contradiction with the left hand side being finite. This means that $x^\top y = 0$ for all y in C_1, i.e. $x^\top A^\top z = 0$ for all $z \in \mathbb{R}^N$, which can only hold if $Ax = 0$. As C_2 contains the standard basis, we also conclude that $x >> 0$. Finally, we can then safely put $d_i = \frac{x_{i+1}}{x_1}$ for $i = 1, \ldots, T$ and by rewriting $Ax = 0$, we see that this d is indeed a discount vector. ■

Arguably, using an abstract separating hyperplane theorem in the proof is overkill. It may be also proven using so-called theorems of the alternative encountered when studying duality for linear programming problems; something that we will elaborate on in Section 2.7. There are several reasons why we have chosen the approach above. First, the most commonly encountered versions of theorems of the alternative (such as Farkas' lemma) and duality do not give us strictness of inequalities. Second, the proof above demonstrates that one implication is easy and one is hard. Third, the proof is structurally very similar to what we will encounter later in more complicated models. And finally it keeps the presentation self-contained – up to the separation theorem being proved in the appendix.

The first fundamental theorem tells us that the search for arbitrages comes down analyzing the solution space to $\pi = Cd$; a problem of linear algebra. If C is an invertible quadratic matrix (meaning that there are as many securities as there are time-points, and that the securities are linearly independent), a unique discount vector candidate is found as $C^{-1}\pi$. If $N > T$ and we can find and invertible T-by$T$ sub-matrix of C by picking out rows, there is also a unique candidate solution, but more conditions to check (the remaining $N - T$ equations). Checking for absence of arbitrage in the where $N < T$ (ie. there are more payment dates than assets) requires more thought – we will discuss it at the end of this section.

Two other financially important concepts are replication and market completeness.

Definition 5. *We say that a payment vector $y \in \mathbb{R}^T$ can replicated if there exists a $\theta \in \mathbb{R}^N$ such that $C^\top \theta = y$. We say that the security market is complete if* every *$y \in \mathbb{R}^T$ can be replicated.*

In financial terms completeness means that an investor can use the existing market to generate any payment stream she likes. Note that while the investor can decide freely on the payment stream, she cannot decide what it costs to relicate it.

In linear algebra terms completeness means that the rows of C span $\mathbb{R}^T$, something that can certainly only happen if $N \geq T$. Mathematically pleasingly, the next theorem shows that completeness is very closely related to uniqueness of discount vectors.

Theorem 2. *Assume that (π, C) is arbitrage-free. Then the market is complete if and only if there is a unique vector of discount factors.*

A result of this nature is commonly referred to as *second fundamental theorem.* Notice that while the result links completeness to uniqueness of discount vectors, completeness (or lack hereof) is solely dependent on C (not π – unlike absence of arbitrage), speficically whether or not its rank is T.

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 then $\mathbb{R}^T$ is spanned by the columns of $C^\top$, ie. the rows of C of which there are N. This means that C has T linearly independent rows, and from basic linear algebra (look around where *rank* is defined) it also has T linearly independent columns, which is to say that all the columns are independent. They therefore form a basis for a T-dimensional linear subspace of $\mathbb{R}^N$ (remember we must have $N \geq T$ to have completeness), ie. any vector in this subspace has unique representation in terms of the basis-vectors. Put differently, the equation $Cx = y$ has at most one solution. And in case where $y = \pi$, we know there is one by absence of arbitrage. For the other direction assume that the model is incomplete. Then the columns of C are linearly dependent, and that means that there exists a vector $\widetilde{d} \neq 0$ such that $0 = C\widetilde{d}$. Since $d \gg 0$, we may choose $\epsilon > 0$ such that $d + \epsilon\widetilde{d} \gg 0$. Clearly, this produces a vector of discount factors different from d. ■

The first fundamental theorem says that in any arbitrage-free market we may write

$$\pi_i = \sum_{t=1}^{T} C_{i,t} d_t \text{ for all } i \text{ and for some discount vector } d.$$

In words, today's price of a bond is the sum of its discounted future payments. This means:

- We are only allowed to add or compare payments occuring at different dates if we properly discount them. Receiving \$1000 in 10 years does not have the same value to us today as receiving \$1000 in 10 years.

- The same discount vector must be used for all bonds. So receiving \$1000 in 5 years has the same value irrespective of which bond it comes from. If it seems strange to the reader that anyone would treat payments differently depending on whence they came, then this section has fulfilled an important purpose. (And: It can happen easier than one would think.)

A point that can challenge the intuition is that the discount vector of which we write is not necessarily unique. In (arbitrage-free) incomplete markets there are many sensible or consistent ways to discount future payments. However, the second fundamental theorem tells us that if there are enough bonds, the discount factor is unique. (And vice versa.)

Example 2. Consider this small financial market

$$\pi = \begin{pmatrix} 101 \\ 101 \\ 99 \\ 105 \end{pmatrix} \qquad C = \begin{pmatrix} 53 & 53 & 0 \\ 35 & 35 & 35 \\ 2 & 102 & 0 \\ 4 & 4 & 104 \end{pmatrix}.$$

The two first securities, whose payments are constant where they are not 0, are called (respectively) 2- and 3-year annuities; more on such later.

Looking at the payments matrix, we notice that security 2's payment stream can be constructed from those of securites 1 and 4. Specifically, the portfolio $\theta^\top = \left(\frac{875}{1378}, 0, 0, \frac{35}{104}\right)$ has the price

$$\pi^\top \theta = \frac{875}{1378} \cdot 101 + \frac{35}{104} \cdot 105 \approx 99.4693$$

and the payment vector

$$C^\top \theta = \frac{875}{1378} \begin{pmatrix} 53 \\ 53 \\ 0 \end{pmatrix} + \frac{35}{104} \begin{pmatrix} 4 \\ 4 \\ 104 \end{pmatrix} = \begin{pmatrix} 35 \\ 35 \\ 35 \end{pmatrix},$$

ie. θ replicates the payments of the 3-year annuity but at a lower price. The portfolio

$$\theta_{\text{arb2}}^\top = \left(\frac{875}{1378}, -1, 0, \frac{35}{104}\right)$$

is therefore an arbitrage: It has the price

$$\pi^\top \theta_{\text{arb2}} = \frac{875}{1378} \cdot 101 - 1 \cdot 101 + \frac{35}{104} \cdot 105 \approx -1.5307 < 0$$

and the payment vector

$$C^\top \theta_{\text{arb2}} = \frac{875}{1378} \begin{pmatrix} 53 \\ 53 \\ 0 \end{pmatrix} - \begin{pmatrix} 35 \\ 35 \\ 35 \end{pmatrix} + \frac{35}{104} \begin{pmatrix} 4 \\ 4 \\ 104 \end{pmatrix} = 0.$$

If we remove the 3-year annuity, we get the market

$$\pi = \begin{pmatrix} 101 \\ 99 \\ 105 \end{pmatrix} \qquad C = \begin{pmatrix} 53 & 53 & 0 \\ 2 & 102 & 0 \\ 4 & 4 & 104 \end{pmatrix},$$

which is arbitrage-free and complete, as $Cd = \pi$ has the unique and strictly positive solution

$$d = \begin{pmatrix} 0.9537736 \\ 0.9518868 \\ 0.9363208 \end{pmatrix}.$$

Were we to use this vector of discount factors to price the 3-year annuity, the full market would be arbitrage-free.

The even smaller market

$$\widetilde{\pi} = \begin{pmatrix} 101 \\ 105 \end{pmatrix} \qquad \widetilde{C} = \begin{pmatrix} 53 & 53 & 0 \\ 4 & 4 & 104 \end{pmatrix}$$

is arbitrage-free but incomplete. The relevant equations are $d_1 + d_2 = 101/53$ og $d_3 = (105 - 4 * 101/53)/104 = 0.9363208$. So $d_1 = d_2 = 101/106$ is one positive solution (hence no arbitrage), while $d_1 = 51/53$, $d_2 = 50/53$ is another (hence incompleteness). Note that d_3 is the same for all discount vectors. This is because there *does* exist a solution to $\widetilde{C}^\top \theta = (0, 0, 1)^\top$, or: Because the 3-year zero coupon bond (a term that we will define shortly) can be replicated. ■

A model with $T > N$ is necessarily incomplete. Our go-to guess would be that it is also arbitrage-free. We can try to verify that via qualified guesses; attach some values to $T - N$ of the d-coordinates, solve for the remaining ones, and hope that things fit. This, however, is not a *fail-safe* method; we may guess poorly or there may actually be arbitrage in the model. For instance, that happens if we divide π_2 by 14 in the incomplete model in the example above. A slow but safe way is to check if (for all $i = 1, \ldots, T$) the linear programming problems

$$\max_{d \in \mathbb{R}^T} d_i \text{ subject to } Cd = \pi, d \geq 0$$

have finite, strictly positive optimal values. If "yes", then the model is arbitrage-free, if "no", then there is arbitrage.

2.2 Zero coupon bonds and the term structure

In this section we will — unless we very clearly say otherwise – consider models (π, C) ithat are arbitrage-free and complete. We 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 an invertible $T \times T$ matrix.

Definition 6. *The payment stream of 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 1. *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 $\pi^\top \theta_t = (Cd)^\top \theta_t = d^\top C^\top \theta_t = d^\top e_t = d_t$. ■

From the definition and uniqueness of d it makes sense to talk about *the* value of a stream of payments c as the sum $\sum_{t=1}^T c_t d_t$. In other words, the value of a stream of payments is obtained by discounting back the individual components. Notice that the dicount factors used to find values are the same across payment streams, but can (will!) vary with payment date, t. Also, there is nothing neither in our definition of d, nor in the theorems about it, that prevents $d_t > d_s$ for $t > s$, or in financial terms negative interest rates. It just rarely, but not never, happens in practice.

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

Definition 7. *The short 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*

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

where $d_0 = 1$ *by convention.*

The interpretation of the short rate ("short" here means "over the next short time-period", not "short" as in "having sold") is 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, less obviously, 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 $t+1$.

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 $\frac{d_t}{d_{t+1}} = 1 + f(0,t)$. If it were possible to agree on any other rate than $f(0,t)$ (for a loan/deposit between t and $t+1$), there would be arbitrage.

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

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

Because

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

we may 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 curve) at date 0 is given by* $(y(0,1),\ldots,y(0,T))$.

If we have any one of the vector of zero coupon yields, the vector of one-period forward rates, or 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 rates depicts yields of zero coupon bonds. We do however also speak of yields on securities with general positive payment steams:

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*

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

Note that the discount factors implied by this definition, $d_i = \frac{1}{((1+y)^i}$, (a) have a quite particular payment time-dependence, (b) are security specific.

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

$$\left(1+\frac{r_m}{m}\right)^{mn}. \tag{2.1}$$

The standard example: If you borrow \$1in 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 (2.1) (and then drop the m-subscript) and and let m tend to infinity, then a basic result from calculus gives us that

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

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 of length 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 corresponds 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 $m = 1$).And then a final piece of advice: Whenever you do calculations be careful always to plug in interest rates as decimal numbers, not as percentages. There is a large difference between $e^{0.12}$ and e^{12}, much larger than what can be recovered by dividing the end result by 100. ■

2.3 Real-world bond markets: Annuities and bullets

Typically, zero-coupon bonds 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 in (so-called) fixed income markets are annuities, serial loans, and bullet bonds (in Danish; "stående lån"). In literature relating to the American market, "bond" is usually understood to mean "bullet bond with 2 yearly payments". Further, "bills" are term short bonds, annuities explicitly referred to as such, and serial loans rare. 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 or notional 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 $c_1, \dots, c_\tau$ as satisfying the following difference equation:

$$p_t = (1+R)p_{t-1} - c_t \qquad t = 1, \dots, \tau, \tag{2.2}$$

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 dates* 1 *and τ, and whose principal evolves according to Equation (2.2).*

With constant payments we can use (2.2) repeatedly to write the remaining principal at time t as

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

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

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

so by using the well-known formula $\sum_{i=0}^{n-1} x^i = (x^n-1)/(x-1)$ for the summation of a geometric series, we get

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

Note that the size of the payment is homogeneous (of degree 1) in the principal, so it's usually enough to look at the $F = 1$. (This rather trivial observation can in fact be extremely useful in a dynamic context.) 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:

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

In the definition of an annuity, the size of the payments is implicitly defined. The definitions of bullets and serials are more direct.

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 loan or 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	$\frac{F}{\tau} + R\left(F - \frac{t-1}{\tau}F\right)$	$R\left(F - \frac{t-1}{\tau}F\right)$	$\frac{F}{\tau}$

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

[1] In Danish: Et stående lån

$\alpha_{3\rceil 6} = 2.6730$ 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 4 % coupon rate and yearly coupon payments that matures on January 1 2010. Around February 1 2005 you could read the following on the CSX homepage or on the financial pages of decent newspapers

Bond type	Current date	Maturity date	Price	Yield
4% bullet	February 1 2005	January 1 2010	104.02	3.10 %

Let us see how the yield was calculated. First, we need to set up the cash-flow stream that results from buying the bond. The first cash-flow, π in the sense of Definition 8 would take place today. (Actually it wouldn't, even these days trades take a couple of days to be in effect; *valør* in Danish. We don't care here.) And how large is it? By convention, and reasonably so, the buyer has to pay the price (104.02; this is called the *clean price*) plus compensate the seller of the bond for the accrued interest over the period from January 1 to February 1, ie. for 1 month, which we take to mean 1/12 of a year. (This is not as trivial as it seems. In practice there are a lot of finer - and extremely boring - points about how days are counted and fractions calculated. Suffice it to say that mostly actual days are used in Denmark.) By definition the buyer has to pay accrued interest of "coupon × year-fraction", ie. $4 \times 1/12 = 0.333$, so the total payment (called the *dirty price*) is $\pi = 104.35$. So now we can write down the cash-flows and verify the yield calculation:

Date	t_k	Cash-flow (c_k)	$d_k = (1 + 0.0310)^{-t_k}$	PV= $d_k * c_k$
Feb. 1 2005	0	- 104.35	1	
Jan. 1 2006	$\frac{11}{12}$	4	0.9724	3.890
Jan. 1 2007	$1\frac{11}{12}$	4	0.9432	3.772
Jan. 1 2008	$2\frac{11}{12}$	4	0.9148	3.660
Jan. 1 2009	$3\frac{11}{12}$	4	0.8873	3.549
Jan. 1 2010	$4\frac{11}{12}$	104	0.8606	89.505
				SUM = 104.38

(The match, 104.35 vs. 104.38 isn't perfect. But to 3 significant digits 0.0310 *is* the best solution, and anything else can be attributed to out rough approach to exact dates.)

Example 6 (Finding the yield curve)**.** In early February you could find prices 4%-coupon rate bullet bonds with a range of different maturities (all maturities fall on January firsts):

Maturity year	2006	2007	2008	2009	2010
Clean price	101.46	102.69	103.43	103.88	104.02
Maturity year	2011	2012	2013	2014	2015
Clean price	103.80	103.50	103.12	102.45	102.08

These bonds (with names like `4%10DsINKx`) are used for the construction of private home-owners variable/floating rate loans such as "FlexLån". (Hey! How does the interest rate get floating? Well, it does if you (completely) refinance your 30-year loan every year or every 5 years with shorter maturity bonds.) In many practical contexts these are not the right bonds to use; yield curves "should" be inferred from government bonds. (Of course this statement makes no sense within our modelling framework.)

Dirty prices, these play the role of π, are found as in Example 5, and the (10 by 10) C-matrix has the form

$$C_{i,j} = \begin{cases} 4 & \text{if } j < i \\ 104 & \text{if } j = i \\ 0 & \text{if } j > i \end{cases}$$

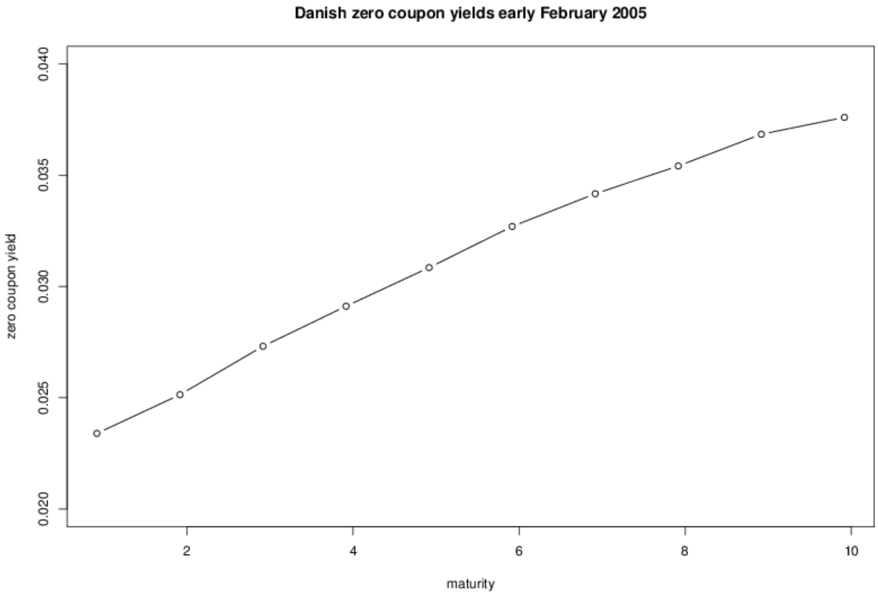


Figure 2.1: The term structure of interest rates in Denmark, February 2005. The o's are the points we have actually calculated, the rest is just linear interpolation.

The system $Cd = \pi$ has the positive ($\sim$ no arbitrage) unique ($\sim$ completeness) solution

$$d = (0.9788, 0.9530, 0.9234, 0.8922, 0.8593, 0.8241, 0.7895, 0.7555, 0.7200, 0.6888)^\top.$$

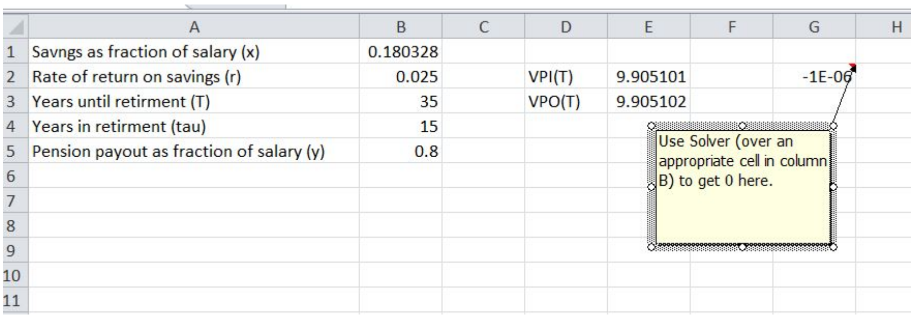
and that corresponds to these (yearly compounded) zero coupon yields:

Maturity	0.92	1.92	2.92	3.92	4.92	5.92	6.92	7.92	8.92	9.92
ZC yield in %	2.37	2.55	2.77	2.95	3.13	3.32	3.48	3.61	3.75	3.83

as depicted in Figure 2.1. Estimating yield curves (also known determining discount factors) is a very important, though not particularly glamorous, task in the financial sector. Two things that make it challenging are (1) there are more relevant payments dates than there are bonds, (2) following the 2007-8-9 financial crisis/turmoil credit/default/bankruptcy risk can be/is being ignored less.

2.4 Financial planning and the net present value rule

In this section we look at some (toy-size) financial planning problems (also known as capital budgeting problems) and make good use of the insights from the previous sections. We will assume throughout that we have a complete security market as defined in the



	A	B	C	D	E	F	G	H
1	Savngs as fraction of salary (x)	0.180328						
2	Rate of return on savings (r)	0.025		VPI(T)	9.905101		-1E-06	
3	Years until retirment (T)	35		VPO(T)	9.905102			
4	Years in retirment (tau)	15						
5	Pension payout as fraction of salary (y)	0.8						
6								
7								
8								
9								
10								
11								

Use Solver (over an appropriate cell in column B) to get 0 here.

Figure 2.2: Pension savings calculations done in Excel. File: https://tinyurl.com/2ajx2r5w

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 $y(0,i)$ when it causes no confusion.

Example 7. (Saving for retirement) Annuity type calculations are very useful for pension savings calculations. Suppose that a newly (i.e. at time 0) graduated person saves the fraction x of his or her salary (assumed fixed) every year for retirement, and that the pension savings have a yearly (deterministic) rate of return of r. We assume that payments are made at times 1, 2, $\ldots$, T where the person retires. The pension is then paid out in yearly installments of y (which can be interpreted as a fraction of the pre-retirement salary) at times $T+1$, $T+2$, $\ldots$, $T+\tau$. Using the geometric summation formula $\sum_{j=0}^{n-1} z^j = \frac{z^n-1}{z-1}$ we can find the value at time T of the money that has been paid in to be

$$VPI(T) := x + x(1+r) + \ldots + x(1+r)^{T-1} = x\frac{(1+r)^T-1}{r}.$$

Similarly, the time-T value of the pension payouts is

$$VPO(T) := \frac{y}{1+r} + \ldots + \frac{y}{(1+r)^\tau} = y\frac{1-(1+r)^{-\tau}}{r}.$$

Various strategies for saving for retirement can then be analyzed by studying $VPI(T) = VPO(T)$. Or more concretely, by fixing all-but-one input parameters and solving (possibly numerically) for the remaining one. A spreadsheet is very well-suited for this. Figure 2.2 shows a numerical example. The 18% savings rate is typical for Denmark. The 2.5% rate of return (which we should think of as being in excess of inflation, that we ignore here) is on the conservative side. Or rather it should be, but is actually above what most pension companies will promise you these days. ■

Similarly to yield, we can define the so-called internal rate of return (IRR) for any cash flow stream, 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 (with* $y > -1$*) 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. For securities whose future payments alternate in sign 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 (NPV):

Definition 15. *Given a term structure* $(y(0,1), \ldots, y(0,T))$*, the PV (present value) and NPV (net present value) of a payment stream* $(c_1, \ldots, c_T)$ *with price* c_0 *are defined as*

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

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; this is the NPV rule or criterion. We will see why this is shortly, Proposition 2. We will see examples of other, seemingly reasonable, criteria that are generally inconsistent with the NPV criterion. 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.

Proposition 2. *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, with price d_1) 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. ■

The interpretation of the proposition is the following: Never accept a project with negative NPV because 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 financial market. One could call this an arbitrage opportunity; buy the project and sell the corresponding future cash flow in the financial market generating a profit at time 0 with no future obligations. However, we insist on relating the term arbitrage to the financial 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', which 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. 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 are larger than yields on short-maturity bonds. (A precise explanation of why this is a stylized fact requires the introduction of interest rate models with uncertainty – and even in that case it is tricky.)

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

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

This gives us what is known a Gordon's growth formula: With an interest rate of r, the value, V of an infinite stream of payments growing at the rate g is

$$V = \frac{c}{r-g}.$$

The formula shows that if r and g are close, then small changes in the interest rate or growth rate will have a large effect on the current value of the payments stream.

Example 8. (Some seemingly sensible rules that are inconsistent with the NPV rule) The internal rate of return is defined without referring to the underlying term structure. It 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.

2.4.1 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 financial 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. 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. In practice, the maximization over feasible "artificial" projects may not be easy – or at least require some sleight of hand.

Example 9. *(Adapted from Luenberger (1997))* A company need a certain type of machine to produce *widgets*. The machine costs \$10,000 (say, to paid at time 0) to buy and its yearly maintenance costs grow linearly; \$2,000 in year 1 (say, paid at time 1), 3,000 in year 2, and so on. At any time the company can buy a new machine (suppose the old one has 0 scrap value). The new machine has (even in nominal terms) the same price and cost profile. The yield curve is flat at 10%. How often should the machine be changed? Let's analyze: Changing every year gives the cash flow stream (-10,-2, 0, ...) + (0,-10,-2, 0, ...) + (0, 0,-10,-2, 0 ...) + ... The present value of this (up to change of sign and division by 1,000) must (as everything starts over at time 1) solve

$$PV = 10 + 2/1.1 + PV/1.1 \Rightarrow PV = 130.$$

(Pedants should verify that the infinite sum we work with here are sufficiently convergent for such manipulation to be allowed.) If instead we change every k years and denote the present value of the all payments by $PV_{k;total}$ then

$$PV_{k;total} = PV_{k;1\ cycle} + \left(\frac{1}{1.1}\right)^k PV_{k;total},$$

where $PV_{k;1\ cycle} = 10 + \sum_{j=1}^{k} \frac{j+1}{1.1^j}$. The task is to choose k, such that the total PV is minimized (yes, minimized; we changed the sign). This is done numerically:

k	$PV_{k;total}$
1	130.00
2	82.38
3	69.58
4	65.36
5	64.48
6	65.20
7	66.76
8	68.79
9	71.09
10	73.53

So changing after five years is optimal. ■

The moral of this section is simple: Given a complete financial market, investors who are offered projects should 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. 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 but involve risk and uncertainty. 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 versions of the NPV rule that work for uncertain cash flows as well.

2.5 Duration and convexity

2.5.1 Duration with a flat term structure.

We now introduce the notions of duration and convexity which are often used in practical bond risk management and asset/liability management. It must be stressed that because we work in a setting without “proper” modelling of risk/uncertainty/randomness some care (more than is often seen) is needed when applying the ideas in more general settings.

Consider an arbitrage-free and complete financial market where the discount function $d = (d_1, \dots d_T)$ has the form

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

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.

We are about to investigate changes in present values as a function of changes in r. 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.

Definition 16. *Let c be a non-negative payment stream. The Macaulay duration $D(c; r)$ of c is given by*

$$\begin{aligned} D(c;r) &= \left(-\frac{\partial}{\partial r} PV(c;r)\right) \frac{1+r}{PV(c;r)} \qquad (2.3) \\ &= \frac{1}{PV(c;r)} \sum_{t=1}^{T} t \frac{c_t}{(1+r)^t} \end{aligned}$$

The Macaulay duration is the classical one; several other forms of duration have been proposed in the literature. Note that rather than saying the duration measure is based on a flat term structure, we could refer to it as being based on the yield of the payment stream. By defining

$$w_t = \frac{c_t}{(1+r)^t} \frac{1}{PV(c;r)}, \qquad (2.4)$$

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

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

This shows that duration has a dual interpretation. On the one hand it is (by our definition) a price sensitivity (a sign changed elasticity, to be precise) to interest rate changes. On the other hand it is (as it turns out from the math above) a value-weighted average of payment dates – which is the reason for the use of the name duration. Note 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.

Definition 17. *The convexity of c is given by*

$$K(c;r) = \sum_{t=1}^{T} t^2 \, w_t. \tag{2.5}$$

where w_t is given by (2.4).

Let us try to interpret D and K by computing the first and second derivativesof $PV(c;r)$ 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.

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

Taylor expanding the numerator to the second order 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.

Example 10. For the bullet bond in Example 5 the present value of the payment stream is 104.35 and $y = 0.0310$, so therefore the Macaulay duration is

$$\frac{\sum_{k=1}^{4} t_k c_k (1+y)^{-t_k}}{PV} = \frac{475.43}{104.35} = 4.556$$

while the convexity is

$$\frac{\sum_{k=1}^{4} t_k^2 c_k (1+y)^{-t_k}}{PV} = \frac{2266.35}{104.35} = 21.72,$$

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

Yield	$\triangle$yield	Exact rel. (%)	First order	Second order
		PV-change	approximation	approximation
0.021	-0.010	4.57	4.42	4.54
0.026	-0.005	2.27	2.21	2.24
0.031	0	0	0	0
0.036	0.005	-2.15	-2.21	-2.18
0.041	0.010	-4.27	-4.42	- 4.30

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). ■

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 maturity guarantees the owner an annual return of r between time 0 and time t. This is not true of a bond with maturity t that 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, i.e.

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

Consider a change in r that occurs an instant after time 0. How would such a change affect $FV(c;r,H)$? There are two effects with opposite directions that 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; a price risk. As the interest rate falls 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. A 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 a world of payment certainty.

2.5.2 Relaxing the assumption of a flat yield curve

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:

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

which gives us

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

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 we get

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

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

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

The multiplicative case is by far the most common one, it goes by the name Fisher-Weil duration,

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

Given the value-weighted average of payment dates interpretation of (Macaulay) duration, this is exactly what we would conjecture a duration measure based on a non-flat term structure to look like. But by going through this analysis, we maintained the connection to yield curve shifts.

A slightly different path to Fisher-Weil duration is this: Let us write the discount function as

$$d(t;x) = \exp(-t \times (z_0(t) + x)),$$

where we suppose $x = 0$ gives us today's zero coupon yield curve. The variable x then creates parallel (additive) shifts in the continuously compounded zero coupon rates. We have $PV(c,x) = \sum_t c_t d(t;x)$ and from this

$$-\frac{1}{PV(c,x)}\frac{\partial PV(c,x)}{\partial x}|_{x=0} = D_{FW}(c)$$

So Fisher-Weil duration is sensitivity to additive shifts in continuously compounded rates.

Example 11 (Macaulay vs. Fisher-Weil)**.** 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

So not much difference. Similarly, the Fisher-Weil duration of the bullet bond from Examples 5, 6 and 10 is 4.552, whereas its Macaulay duration was 4.556. ■

2.6 Two examples to mess with your head

2.6.1 The barbell, or why immunization can't be the whole story

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 dollars worth of 10 year bullet bonds with time to maturity 10 years and a coupon rate of 7 percent, $R = 0.07$. Assume that the term structure is flat at $r = 0,07$. The revenue (of 100 million dollars) 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 (Fisher-Weil) 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.

What this example shows is that our fundamentally deterministic framework is not good enough to deal with uncertainty, with changes; from seemingly sensible assumptions we get out paradoxical results. (After treating arbitrage-free multi-period stochastic model, we will show in Section 28 that we can't have only flat yield curves in an arbitrage-free model.)

2.6.2 Riding the yield curve

First some notation. Slightly cumbersome, but we need it. For a calender date t and a time to maturity τ, let $y(t, \tau)$ be the continuously compounded zero-coupon rate for time to maturity τ, i.e. the (annualized) rate of return we get by investing (at time t) in zero-coupon bonds maturing at time $t + \tau$ and holding them until they mature. The mapping

$$\tau \mapsto y(t, \tau)$$

we call the (time-t zero-coupon) yield curve. In terms of zero-coupon bond prices ($P(t, T)$, first argument current time t, second argument the maturity date T) we have

$$P(t, T) = e^{-(T-t)y(t,T-t)}.$$

Suppose we want to invest at time 0, look one year ahead, and have at our disposal zero-coupon bonds for all possible maturity dates. The rate of return we get from investing in a maturity date T zero-coupon bond is

$$i_1(T) := \frac{P(t, T) - P(0, T)}{P(0, T)} = \frac{P(1, T)}{P(0, T)} - 1.$$

(The notation ":=" means "equal to by definition", with the term nearest the : being defined.) A natural question is: Can we maximize this rate of return by choosing an appropriate T. The short answer is no, not without being able to look into the future. We do not know until time 1 what $P(1, T)$ turns out to be. So we need to make further modelling assumptions. A natural first step is the hypothesis,

$$\boxed{H_0\text{: The time 1 yield curve will be the same as the time 0 yield curve.}}$$

At time 1, a maturity date T zero-coupon bond has time *to* maturity $T - 1$, so under the H_0-hypothesis we have

$$P(1, T) = e^{-(T-1)y(0,T-1)} = P(0, T - 1).$$

Thus the rate of return becomes known. It is in fact our old friend the forward rate,

$$i_1(T) = \frac{P(0, T - 1)}{P(0, T)} - 1 = f(0, T - 1).$$

So to maximize we should invest according to the highest forward rate. (Or more precisely: Find the time to maturity for the maximal forward rate and then invest in

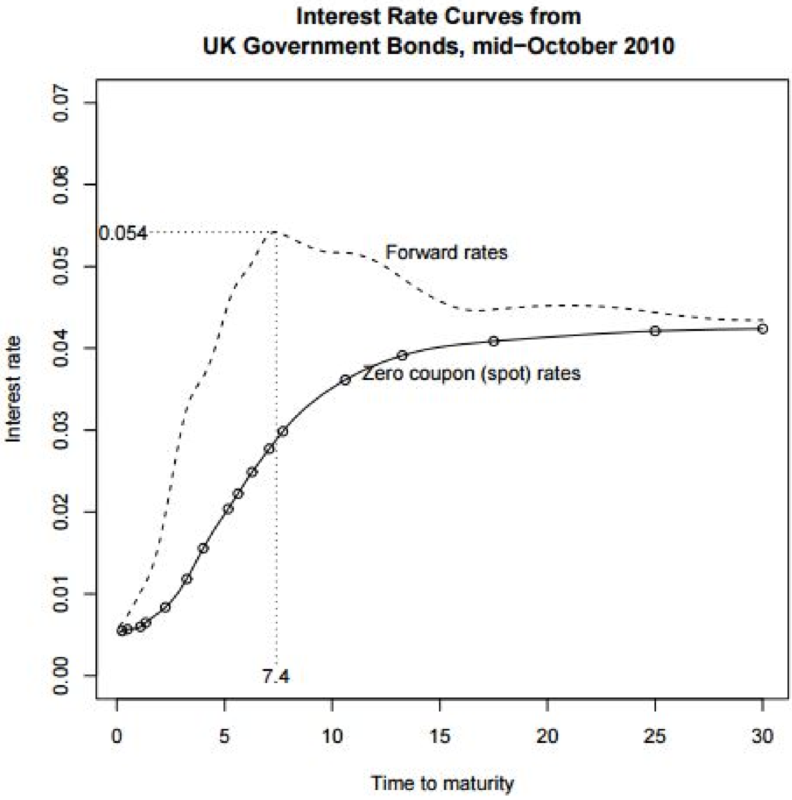


Figure 2.3: UK zero-coupon yield and forward rate curves from mid-October 2010.

zero-coupon bonds with 1 year more to maturity than that.) And that forward rate will then be our return. This strategy is called “riding the yield curve”. Note that it can be carried out whether H_0 holds or not — but only in the former case are we sure what our return will be. In words, this strategy says that to maximize investment returns, go not where the yield curve is at its highest, but where it is at its steepest. If the yield curve is “truly curved” then this can have surprising effects.

A good example is provided by the UK yield curve from mid-October 2010. The zero-coupon and (1-year-ahead) forward curves are shown in Figure 2.3. The circles are (more or less) observed points on the zero-coupon curve. The smooth curve was fitted through them with an interpolation technique called cubic spline. The smooth curve was then used to calculate forward rates. (Notice how the forward rate curve is considerable less smooth than the zero-coupon rate curve. How to deal with this is the focus of numerous research articles.) We see that the forward rate curve attains its maximum around 7.5 years, and that the maximal value (5.4% continuously compounded) is considerably above any zero-coupon rate; that curve has its maximum at 4.2% for 30-year maturities.Could this trick be repeated over and over for, say, 30 years £1 would grow to $e^{30 \cdot 0.054} = 5.08$, while investing in maturity-date 30 zero coupon bonds and holding pays back only 3.57. That difference should make any pension fund manager sit up and take notice.

Magic? Alchemy? Or: Is yield curve riding really the free lunch that is seems to be? Of course not. First, it's risky. We only get the return we think if hypothesis H_0 holds, i.e. if the yield curve does not move. And that is a big if. And the longer the maturity of the bond we have invested in, the greater the sensitivity. Second, we may reverse the question and ask: How should the yield curve move for all 1-year returns to be the same? It turns out that if future zero-coupon spot rates are realized at the current forward rates then no gains be achieved by short term riding or rolling. (It should be added: "for appropriately matched times to maturity and years-ahead". To make the statement precise we would need three time indices; we will spare the reader.) This then leads to the counter-argument called the unbiased expectations hypothesis,

H_1: Forward rates are expected future zero-coupon (spot) rates.

So which hypothesis is it then? Well, the truth (if such a thing exists at all) is somewhere in between. First, these are technical arguments (to do with expectations of non-linear functions, Jensen's Inequality, and absence of arbitrage) against the unbiased expectations hypothesis. But the main reason is risk-aversion: If all bonds give the same expected return, then why invest in risky ones at all? Thus prices of long-term bonds will be "low" and one is rewarded for taking the riskier positions — such as riding the yield curve. But, arguably, riding the yield curve gives "double exposure": It's risk and to the extend that it's not, movements will go against you!

2.7 Arbitrage pricing and linear programming

2.7.1 The first fundamental theorem via duality

Consider a linear programming (or optimization) problem on standard form:

$$\max_{x\in\mathbb{R}^n} c^\top x \text{ subject to } Ax \leq b,\ x \geq 0, \qquad \text{(P)}$$

where A in an m-by-n matrix, $b \in \mathbb{R}^m$, and $c \in \mathbb{R}^n$. We call this the primal problem, indicated by (P). To this there is a so-called dual problem. This is Bizarro World where things become their opposite; constraints become variables and vice versa, max becomes min, matrices get transposed and (some) inequalities get reversed. More precisely, the dual, (D), to (P) is

$$\min_{y\in\mathbb{R}^m} b^\top y \text{ subject to } A^\top y \geq c,\ y \geq 0. \qquad \text{(D)}$$

The (strong) duality theorem says if either (P) or (D) has a solution (so there are no issues with infeasibility or unboundedness) then so has the other and the optimal values are equal. (There is also a weak duality version that deals with infeasible and/or unbounded problems.)

With some sleight of hand primal-dual versions can be formulated for problems on non-standard form. This mostly involves dealing with equality (rather than inequality) constraints and free (rather than non-negative) variables through the introduction of artificial variables. The connections are given in the table in Table 2.1. The duality theorem still applies to all these cases/combinations.

	Minimization Problem		Maximization Problem	
Variable	≥ 0	$\Longleftrightarrow$	$\leq$	Constraint
	≤ 0	$\Longleftrightarrow$	$\geq$	
	Unrestricted	$\Longleftrightarrow$	$=$	
Constraint	$\geq$	$\Longleftrightarrow$	≥ 0	Variable
	$\leq$	$\Longleftrightarrow$	≤ 0	
	$=$	$\Longleftrightarrow$	Unrestricted	

Table 2.1: Relations between primal and dual problems on non-standard forms. The table is inspired by Bazaraa, Jarvis and Sherali (1990), “Linear Programming and Network Flows”, Wiley.

We can use duality to prove Theorem 1. Or almost. In the arguments in the following we are “playing fast and loose” with inequality strictness. It can be fixed, but it’s not completely trivial.

Suppose we have a financial market in vector/matrix form, (π, C). Let us consider this problem

$$\max_{d \in \mathbb{R}^T} 0^\top d \text{ subject to } Cd = \pi,\ d \geq 0. \qquad \text{(P')}$$

This corresponds to finding a vector of discount factors. And clearly, the optimal value in (P') is 0. The dual of (P') is (from the table)

$$\min_{\theta \in \mathbb{R}^N} \pi^\top \theta \text{ subject to } C^\top \theta \geq 0,\ \theta \text{ free.} \qquad \text{(D')}$$

If there exists a vector of discount factors, then (P') has a solution, and by the duality theorem so has (D'), but since the theorem also tells that the optimal values are equal — here 0 — then there cannot be an arbitrage, since that would be a θ that satisfies the constraints in (D') but for which the optimal value is strictly less than 0. Going in the other direction, let us assume that (π, C) is arbitrage-free. Clearly (D') is feasible (take $\theta = 0$) and the objective function is bounded below by 0 on the set of feasible θs (otherwise there would be arbitrage). Hence a (trivial) solution to the minimization problem (D') exists. By the duality theorem (P') then has a solution, in particular it is not infeasible, and therefore a discount vector exists.

2.7.2 Arbitrage-free price intervals in incomplete models

We consider again a financial market (π, C) that we assume to be arbitrage-free, but possibly (grossly) incomplete. Now introduce a new contract that has $x \in \mathbb{R}^T$ as its payment vector, and whose time 0 price we denote by π_x. In an incomplete model there is (typically) not a unique arbitrage-free price of this new claim, but we will now show that there is considerable and computable structure on the set of arbitrage-free π_x-values.

Let us look at the linear programming problem ((U-P) for upper-primal)

$$\min_{\theta \in \mathbb{R}^N} \ \theta^\top \pi \ \text{ subject to } C^\top \theta \geq x,\ \ \theta \text{ free.} \qquad \text{(U-P)}$$

We assume that (U-P) has a finite solution (boundedness comes from absence of arbitrage; infeasibility would make things uninteresting) which we denote by $\theta_{*,\mathrm{U}}$ and let $\pi_{*,\mathrm{U}}$ be the associated optimal value. Solving (U-P) can be interpreted as super-replicating x as cheaply as possible.

We claim that $\pi_{*,\mathrm{U}}$ is a minimal upper bound on the arbitrage-free price of x. The "upper" part is easy. Suppose that $\pi_x > \pi_{*,\mathrm{U}}$. Then we buy $\theta_{*,\mathrm{U}}$ and sell the x-claim. That is an arbitrage. Minimality is a little more involved. First, you might think that we were to show that for any $\pi_x < \pi_{*,\mathrm{U}}$, the x-extended financial model is arbitrage-free. But we can't do that, because it's not true; it might involve a violation of the lower bound that we will derive shortly. What we can show is that if $\pi_x < \pi_{*,\mathrm{U}}$ then there can be no arbitrage that involves a short position — that we can without loss of generality take to be -1 — in the x-contract. To do this assume that $\pi_x < \pi_{*,\mathrm{U}}$ and suppose (on the contrary) that $(\theta_A, -1)$ is an arbitrage. From the definition of arbitrage this means that $C^\top \theta_A - x \geq 0$, i.e. that θ_A satisfies the constraints in (U-P). Moreover, we have $0 \geq \theta_A^\top \pi - \pi_x > \theta_A^\top \pi - \pi_{*,\mathrm{U}}$, where the first inequality follows from the definition of an arbitrage and the second (the strict one) from the price assumption. But that contradicts the assumed optimality of $\pi_{*,\mathrm{U}}$, hence there can't be a short x-arbitrage.

By exactly similar reasoning solving

$$\max_{\theta \in \mathbb{R}^N} \theta^\top \pi \text{ subject to } C^\top \theta \leq x, \quad \theta \text{ free.} \qquad \text{(L-P)}$$

gives us a maximal lower bound, $\pi_{*,\mathrm{L}}$, on the arbitrage-free price of x.
Combining the results gives us an interval of arbitrage-free prices. A natural mathematical question: Is the interval open or closed or ...? If the x-claim can be replicated, then the upper and lower bounds collapse into a single point, which is then the unique arbitrage-free price; technically a closed set. If the x-claim cannot be replicated (which is the typical situation where we would apply this whole approach), then at optimum (for both optimization problems) at least one constraint inequality is sharp. Suppose now that the x-claim trades at exactly $\pi_{*,\mathrm{U}}$. Then an arbitrage is constructed by buying $\theta_{*,\mathrm{U}}$ and selling the x-claim. And vice versa if the x-claim trades at $\pi_{*,\mathrm{L}}$. Hence if the x-claim cannot be replicated, then the x-extended model is arbitrage-free if and only if

$$\boxed{\pi_x \ \in \ ]\pi_{*,\mathrm{L}}, \pi_{*,\mathrm{U}}[.}$$

From an application point of view there is good news and bad news. The good news is that the upper and lower bounds are eminently computable; there are excellent and wide-spread numerical methods for solving linear optimization problems. The bad news is that the arbitrage-free price interval is typically too wide to be of practical use.

Using the duality table again we see that the dual of (U-P) is

$$\max_{d \in \mathbb{R}^T} d^\top x \text{ subject to } Cd = \pi, \quad d \geq 0. \qquad \text{(U-D)}$$

This (via the duality theorem) tells us that we can find the upper no arbitrage-bound by finding the discount vector that maximizes the value of the newly introduced claim. A similar result holds in more advanced models (continuous time and space) where we might have the analogous object to d (the martingale measures) parameterized in some way. For specific claims we can then analyze how affected they are by incompleteness by optimizing over the parameters, possibly restricted to what we think are reasonable values.

Chapter 3

Risky business: Mean-variance optimal portfolios and the capital asset pricing model

Facts we need

First, the reader should familiarize herself with the Appendix A about probability theory, in particular the concepts of expectation (denoted by E), variance (Var), and covariance (Cov).cSecond, we need some few facts about matrices. (A very useful reference for mathematical results in the large classcimprecisely defined as "well-known" is Berck & Sydsæter (1992), "Economists' Mathematical Manual", Springer.)

- When $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ then

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top \mathbf{V}\mathbf{x}) = (\mathbf{V} + \mathbf{V}^\top)\mathbf{x}$$

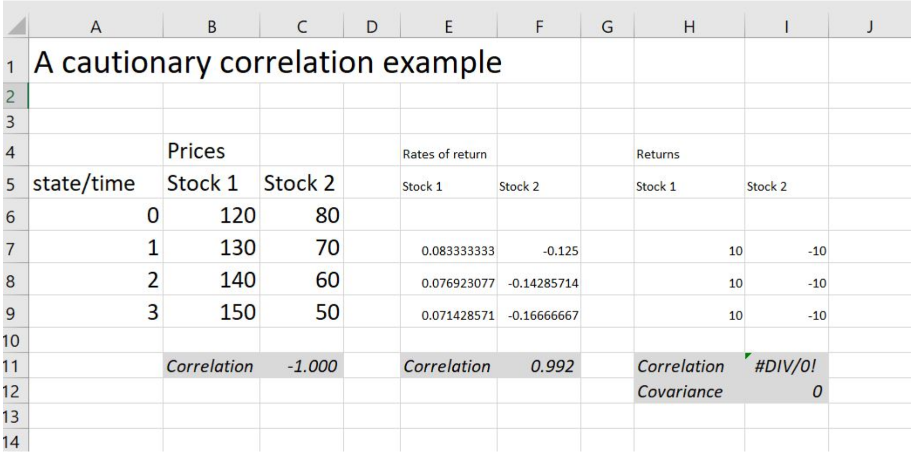
- A matrix $\mathbf{V} \in \mathbb{R}^{n \times n}$ is said to be *positive definite* if $\mathbf{z}^\top \mathbf{V}\mathbf{z} > 0$ for all $\mathbf{z} \neq \mathbf{0}$. If $\mathbf{V}$ is positive definite then $\mathbf{V}^{-1}$ exists and is also positive definite.

- Multiplying (appropriately) partitioned matrices is just like multiplying 2×2-matrices.

- Covariance is bilinear. Or more specifically: 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.



	A	B	C	D	E	F	G	H	I	J
1	A cautionary correlation example									
2										
3										
4		Prices			Rates of return			Returns		
5	state/time	Stock 1	Stock 2		Stock 1	Stock 2		Stock 1	Stock 2	
6	0	120	80							
7	1	130	70		0.083333333	-0.125		10	-10	
8	2	140	60		0.076923077	-0.14285714		10	-10	
9	3	150	50		0.071428571	-0.16666667		10	-10	
10										
11		Correlation	-1.000		Correlation	0.992		Correlation	#DIV/0!	
12								Covariance	0	
13										
14										

Figure 3.1: Uwe Wystup's cautionary correlation example. File: https://tinyurl.com/fe2dc8sn

Example 12. (Inspired by Uwe Wystup.) Ever so often we need to be careful and precise about prices, returns, and rates of return, as well as to whether we mean correlation or covariance. To see why, consider the two-stock example in Figure 12. Someone asks: What is the correlation between the stocks? At that level of specificity the answer can be almost anything. The correlation between prices is -1 (stock 1 goes up-up, stock 2 goes down-down; cell C111), but if we look at rates of return – which is what we use in the mean-variance problem formulation –the correlation is +0.992 (F11), i.e. almost perfectly positive (both rates of return decrease over time because of the numerators; stock 1's become less positive, stock 2's become more negative). To make matters even more confusing: The correlation between returns (price changes) is not well-defined (the standard deviations in the denominator are 0), while the covariance between returns is 0. ■

Basic definitions and 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 $\mathbf{w} = (w_1, \dots w_n)^\top$, 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 coordinates 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

$$\begin{aligned} W_1 &= \sum_{i=1}^{n} a_i S_{i,1} = \sum_{i=1}^{n} a_i (S_{i,1} - S_{i,0}) + \sum_{i=1}^{n} a_i S_{i,0} \\ &= W_0 \left(1 + \sum_{i=1}^{n} \frac{S_{i,0} a_i}{W_0} \frac{S_{i,1} - S_{i,0}}{S_{i,0}} \right) \\ &= W_0 (1 + \mathbf{w}^\top r), \end{aligned} \tag{3.1}$$

so if we know the relative portfolio weights and the realized rates of return, we know terminal wealth. We also see that

$$E(W_1) = W_0 (1 + \mathbf{w}^\top E(r)),$$

where E denoted expectation (or mean) and

$$\mathrm{Var}(W_1) = W_0^2 \mathrm{Cov}(\mathbf{w}^\top r, \mathbf{w}^\top r) = W_0^2 \mathbf{w}^\top \underbrace{\mathrm{Cov}(r)}_{n \times n} \mathbf{w}.$$

In this chapter we will look at how agents should 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 she derives utility 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,

$$\begin{aligned} u(W_1) &= u(E(W_1)) + u'(E(W_1))(W_1 - E(W_1)) \\ &\quad + \frac{1}{2}u''(E(W_1))(W_1 - E(W_1))^2 + R_3, \end{aligned}$$

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 conditions this means that the expected terminal wealth can be written as

$$E(u(W_1)) = u(E(W_1)) + \frac{1}{2}u''(E(W_1))\mathrm{Var}(W_1) + E(R_3),$$

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 (3.1), 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.

3.1 Mathematics of minimum variance portfolios

3.1.1 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

$$\begin{aligned} E(r) &= \mu, && (3.2) \\ \operatorname{Cov}(r) &= \mathbf{\Sigma}, && (3.3) \end{aligned}$$

where $\mathbf{\Sigma}$ is positive definite (hence invertible) and not all coordinates of μ are equal. As a covariance matrix $\mathbf{\Sigma}$ 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:

$$\min_{\mathbf{w}} \frac{1}{2} \underbrace{\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}}_{:=\sigma_P^2} \quad \text{subject to} \quad \mathbf{w}^\top \mu = \mu_P$$

$$\mathbf{w}^\top \mathbf{1} = 1$$

Analysis of such a problem is called mean/variance analysis, or Markowitz analysis after Harry Markowitz who studied the problem in the 40'ies and 50'ies. (He won the Nobel prize in 1990 together with William Sharpe and Merton Miller both of whom we'll meet later.)

Our assumptions on μ and $\mathbf{\Sigma}$ ensure that a unique finite solution exits for any value of μ_P. 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 portfolio's rate return ($\mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}$; the "1/2" is just there for convenience) is minimized given that we want a specific expected rate of return (μ_P; "P is for portfolio").

To solve the problem we set up the Lagrange-function with multipliers

$$\mathcal{L}(\mathbf{w}, \lambda_1, \lambda_2) = \frac{1}{2} \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} - \lambda_1 (\mathbf{w}^\top \mu - \mu_P) - \lambda_2 (\mathbf{w}^\top \mathbf{1} - 1).$$

The first-order conditions for optimality are

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{\Sigma} \mathbf{w} - \lambda_1 \mu - \lambda_2 \mathbf{1} &= 0, && (3.4) \\ \mathbf{w}^\top \mu - \mu_P &= 0, && (3.5) \\ \mathbf{w}^\top \mathbf{1} - 1 &= 0. && (3.6) \end{aligned}$$

Usually we might say "and these are linear equations that can easily be solved", but working on them algebraically leads to a deeper understanding and intuition about the model. Invertibility of $\mathbf{\Sigma}$ gives that we can write (3.4) as (check for yourself)

$$\mathbf{w} = \mathbf{\Sigma}^{-1} [\mu \;\; \mathbf{1}] \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}, \tag{3.7}$$

and (3.5)-(3.6) as

$$[\mu\ \ \mathbf{1}]^\top \mathbf{w} = \begin{bmatrix} \mu_P \\ 1 \end{bmatrix}. \tag{3.8}$$

Multiplying both sides of (3.7) by $[\mu\ \ \mathbf{1}]^\top$ and using (3.8) gives

$$\begin{bmatrix} \mu_P \\ 1 \end{bmatrix} = [\mu\ \mathbf{1}]^\top \mathbf{w} = \underbrace{[\mu\ \mathbf{1}]^\top \boldsymbol{\Sigma}^{-1}[\mu\ \mathbf{1}]}_{=:\mathbf{A}} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}. \tag{3.9}$$

Using the multiplication rules for partitioned matrices we see that

$$\mathbf{A} = \begin{bmatrix} \mu^\top \boldsymbol{\Sigma}^{-1}\mu & \mu^\top \boldsymbol{\Sigma}^{-1}\mathbf{1} \\ \mu^\top \boldsymbol{\Sigma}^{-1}\mathbf{1} & \mathbf{1}^\top \boldsymbol{\Sigma}^{-1}\mathbf{1} \end{bmatrix} =: \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

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

$$\mathbf{y} = [\mu\ \mathbf{1}] \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = z_1\mu + z_2\mathbf{1} \neq \mathbf{0},$$

because the coordinates of μ 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 $\boldsymbol{\Sigma}^{-1}$ is positive definite (because $\boldsymbol{\Sigma}$ is). In other words, $\mathbf{A}$ is positive definite. Hence we can solve (3.9) for the λ's,

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \mathbf{A}^{-1} \begin{bmatrix} \mu_P \\ 1 \end{bmatrix},$$

and insert this into (3.7) in order to determine the optimal portfolio weights

$$\widehat{\mathbf{w}} = \boldsymbol{\Sigma}^{-1}[\mu\ \mathbf{1}]\mathbf{A}^{-1} \begin{bmatrix} \mu_P \\ 1 \end{bmatrix}. \tag{3.10}$$

The portfolio $\widehat{\mathbf{w}}$ is called the minimum variance portfolio for a given mean μ_P. (We usually can't be bothered to say the correct full phrase: "minimum variance on rate of return for a given mean rate on return μ_P ".) The minimal portfolio return variance is

$$\begin{aligned} \widehat{\sigma}_P^2 &= \widehat{\mathbf{w}}^\top \boldsymbol{\Sigma} \widehat{\mathbf{w}} \\ &= [\mu_P\ \ 1]\mathbf{A}^{-1}[\mu\ \ \mathbf{1}]^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}[\mu\ \ \mathbf{1}]\mathbf{A}^{-1}[\mu_P\ \ 1]^\top \\ &= [\mu_P\ \ 1]\mathbf{A}^{-1} \underbrace{\left([\mu\ \ \mathbf{1}]^\top \boldsymbol{\Sigma}^{-1}[\mu\ \ \mathbf{1}]\right)}_{=\mathbf{A}\ \text{by def.}} \mathbf{A}^{-1}[\mu_P\ \ 1]^\top \\ &= [\mu_P\ \ 1]\mathbf{A}^{-1} \begin{bmatrix} \mu_P \\ 1 \end{bmatrix}, \end{aligned}$$

where symmetry (of $\mathbf{\Sigma}$ and $\mathbf{A}$ and their inverses) was used to obtain the second line. But since

$$\mathbf{A}^{-1} = \frac{1}{ac - b^2} \begin{bmatrix} c & -b \\ -b & a \end{bmatrix},$$

we have

$$\widehat{\sigma}_P^2 = \frac{a - 2b\mu_P + c\mu_P^2}{ac - b^2}. \tag{3.11}$$

In (3.11) 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 *minimum variance portfolio frontier* or *locus*.

Note that we have not just solved one "minimize variance" problem, but a whole bunch of them, namely one for each conceivable expected rate of return.

In mean-standard deviation-space the relation is expressed as a hyperbola. Figure 3.2 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 3.2 (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*.

Figure 3.2 also shows the *global minimum variance portfolio*, the portfolio with the smallest possible variance for any given mean return. Its mean, μ_G, is found by minimizing (3.11) with respect to μ_P, and is $\mu_{gmv} = \frac{b}{c}$. By substituting this in the general $\widehat{\sigma}^2$-expression we obtain

$$\widehat{\sigma}_{gmv}^2 = \frac{a - 2b\mu_{gmv} + c\mu_{gmv}^2}{ac - b^2} = \frac{a - 2b(b/c) + c(b/c)^2}{ac - b^2} = \frac{1}{c},$$

while the general formula for portfolio weights gives us

$$\widehat{\mathbf{w}}_{gmv} = \frac{1}{c}\mathbf{\Sigma}^{-1}\mathbf{1}.$$

Example 13. Consider the case with 3 assets (referred to as A, B, and C) and

$$\mu = \begin{bmatrix} 0.1 \\ 0.12 \\ 0.15 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} 0.25 & 0.10 & -0.10 \\ 0.10 & 0.36 & -0.30 \\ -0.10 & -0.30 & 0.49 \end{bmatrix}.$$

The all-important $\mathbf{A}$-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

$$\widehat{\sigma}_P^2 = 4.22918 - 65.3031\mu_P + 255.097\mu_P^2.$$

The locus is illustrated in Figure 3.3 in both in (variance, expected return)-space and (standard deviation, expected return)-space.

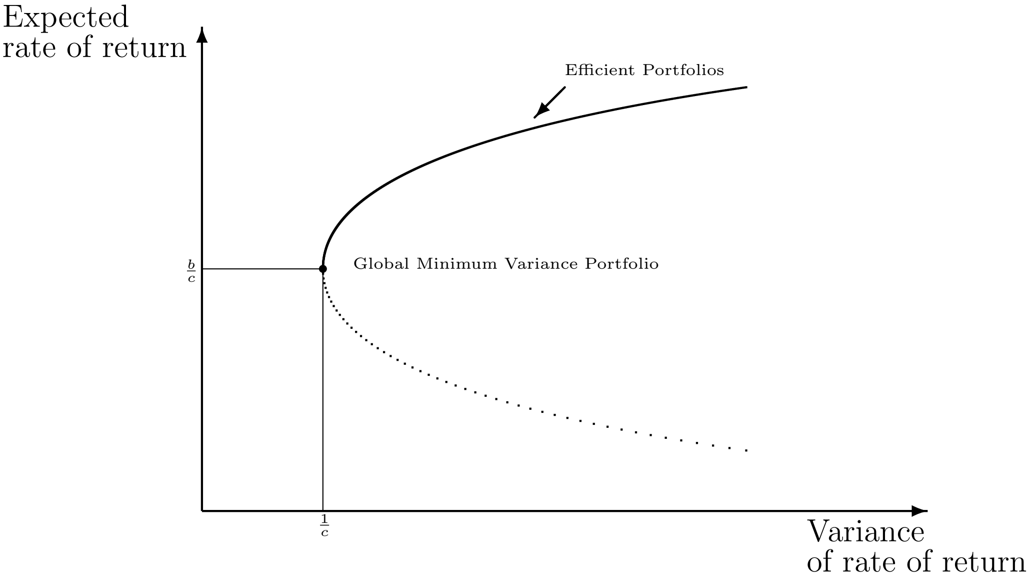


Figure 3.2: The minimum variance portfolio frontier.

An important property of the set of minimum variance portfolios is the so-called two-fund separation. This means that the minimum variance portfolio frontier can be generated by any two distinct minimum variance portfolios.

Proposition 3. *Let $\mathbf{x}_a$ and $\mathbf{x}_b$ be two minimum variance portfolios with mean returns μ_a and μ_b, $\mu_a \neq \mu_b$. 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 μ_c denote the mean return on a given minimum variance portfolio $\mathbf{x}_c$. Now choose α such that $\mu_c = \alpha\mu_a + (1-\alpha)\mu_b$, that is $\alpha = (\mu_c - \mu_b)/(\mu_a - \mu_b)$ (which is well-defined because $\mu_a \neq \mu_b$). But since $\mathbf{x}_c$ is a minimum variance portfolio we know that (3.10) holds, so

$$\begin{aligned}
\mathbf{x}_c &= \mathbf{\Sigma}^{-1}[\mu \ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} \mu_c \\ 1 \end{bmatrix} \\
&= \mathbf{\Sigma}^{-1}[\mu \ \ \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} \alpha\mu_a + (1-\alpha)\mu_b \\ \alpha + (1-\alpha) \end{bmatrix} \\
&= \alpha\mathbf{x}_a + (1-\alpha)\mathbf{x}_b,
\end{aligned}$$

where the third line is obtained because $\mathbf{x}_a$ and $\mathbf{x}_b$ also fulfill (3.10). 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

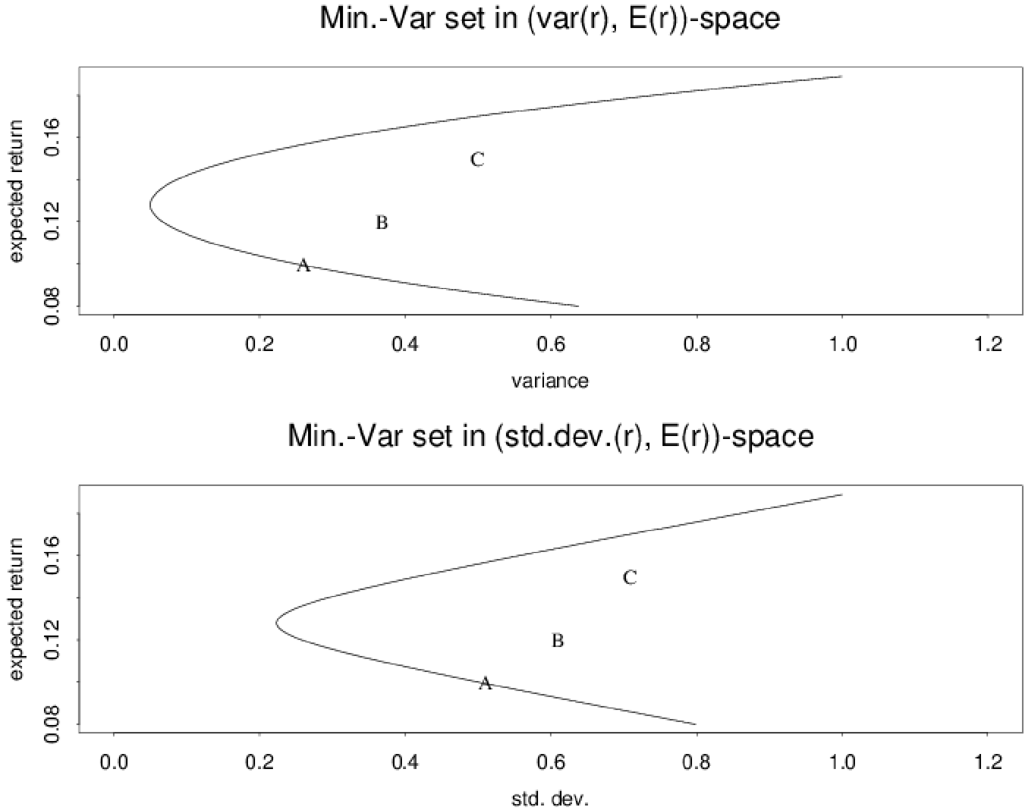


Figure 3.3: The minimum variance frontiers and individual assets

with expected return $\alpha\mu_a + (1-\alpha)\mu_b$. 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.

$$\mathbf{x}_{zP}^\top \mathbf{\Sigma} \mathbf{x}_P = 0. \tag{3.12}$$

Often $\mathbf{x}_{zP}$ is called $\mathbf{x}_P$'s 0-β portfolio (we'll see why later).

Proposition 4. *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 μ_P, its orthogonal one has mean*

$$\mu_{zP} = \frac{a - b\mu_P}{b - c\mu_P}.$$

Proof. First note that μ_{zP} is well-defined for any portfolio except the global minimum variance portfolio. By (3.10) we know how to find the minimum variance portfolios with means μ_P and $\mu_{zP} = (a - b\mu_P)/(b - c\mu_P)$. This leads to

$$\begin{aligned}
\mathbf{x}_{zP}^\top \mathbf{\Sigma} \mathbf{x}_P &= [\mu_{zP} \;\; 1]\mathbf{A}^{-1}[\mu \;\; \mathbf{1}]^\top \mathbf{\Sigma}^{-1}\mathbf{\Sigma}\mathbf{\Sigma}^{-1}[\mu \;\; \mathbf{1}]\mathbf{A}^{-1}[\mu_P \;\; 1]^\top \\
&= [\mu_{zP} \;\; 1]\mathbf{A}^{-1}\underbrace{\left([\mu \;\; \mathbf{1}]^\top \mathbf{\Sigma}^{-1}[\mu \;\; \mathbf{1}]\right)}_{=\mathbf{A} \text{ by def.}}\mathbf{A}^{-1}[\mu_P \;\; 1]^\top \\
&= [\mu_{zP} \;\; 1]\mathbf{A}^{-1}\begin{bmatrix} \mu_P \\ 1 \end{bmatrix} \qquad (3.13) \\
&= \begin{bmatrix} \dfrac{a - b\mu_P}{b - c\mu_P} & 1 \end{bmatrix} \frac{1}{ac - b^2} \begin{bmatrix} c & -b \\ -b & a \end{bmatrix} \begin{bmatrix} \mu_P \\ 1 \end{bmatrix} \\
&= \frac{1}{ac - b^2} \begin{bmatrix} \dfrac{a - b\mu_P}{b - c\mu_P} & 1 \end{bmatrix} \begin{bmatrix} c\mu_P - b \\ a - b\mu_P \end{bmatrix} \\
&= 0,
\end{aligned}$$

which was the desired result. □

Proposition 5. *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_{mv}^2*. 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

$$\beta_{P,mv} = \frac{Cov(r_P, r_{mv})}{\sigma_{mv}^2}.$$

*P*roof. Consider first the covariance between return on asset i and $\mathbf{x}_{mv}$. By using (3.10) we get

$$\begin{aligned}
\text{Cov}(r_i, r_{mv}) &= \mathbf{e}_i^\top \mathbf{\Sigma} \mathbf{x}_{mv} \\
&= \mathbf{e}_i^\top [\mu \;\; \mathbf{1}]\mathbf{A}^{-1}\begin{bmatrix} \mu_{mv} \\ 1 \end{bmatrix} \\
&= [\mu_i \;\; 1]\mathbf{A}^{-1}\begin{bmatrix} \mu_{mv} \\ 1 \end{bmatrix}.
\end{aligned}$$

From calculations in the proof of Proposition 4 we know that the covariance between $\mathbf{x}_{mv}$ and $\mathbf{x}_{zvp}$ is given by (3.13). We also know that it is 0. Subtracting this 0 from the above equation gives

$$\begin{aligned}
\text{Cov}(r_i, r_{mv}) &= [\mu_i - \mu_{zmv} \;\; 0]\mathbf{A}^{-1}\begin{bmatrix} \mu_{mv} \\ 1 \end{bmatrix} \\
&= (\mu_i - \mu_{zmv})\underbrace{\frac{c\mu_{mv} - b}{ac - b^2}}_{:=\gamma}, \qquad (3.14)
\end{aligned}$$

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 (3.14) we get the desired result for individual assets. But then linearity ensures that it holds for all portfolios. ■

Proposition 5 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.

The converse of Proposition 5 holds in the following sense: If there is a candidate portfolio x_C and a number μ_{zC} such that for any individual asset i we have

$$\mu_i - \mu_{zC} = \beta_{i,C}(\mu_C - \mu_{zC}), \tag{3.15}$$

with $\beta_{i,C} = \mathrm{Cov}(r_i, r_C)/\sigma_C^2$, then x_C is a minimum-variance portfolio. To see why, put $\gamma_i = \sigma_C^2(\mu_i - \mu_{zC})/(\mu_C - \mu_{zC})$ and note that we have $\gamma = \Sigma x_C$, which uniquely determines the candidate portfolio. But by Proposition 5 we know that the minimum variance portfolio with expected rate of return μ_C is (the then) one (and only) portfolio for which (3.15) holds.

3.1.2 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 μ_0. For the risky assets we let $\mu_i{}^e$ denote the *excess* rate of return over the riskfree asset, i.e. the actual rate of return less μ_0. We let μ^e denote the mean excess rate of return, and $\mathbf{\Sigma}$ the covariance matrix (which is of course unaffected). A portfolio is now a $n+1$-dimensional vector whose coordinates 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

$$\mu_P^e = \mathbf{w}^\top \mu^e$$

and the variance is

$$\sigma_P^2 = \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}.$$

Therefore the mean-variance portfolio selection problem with a riskless asset can be stated as

$$\min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^\top \mu^e = \mu_P^e.$$

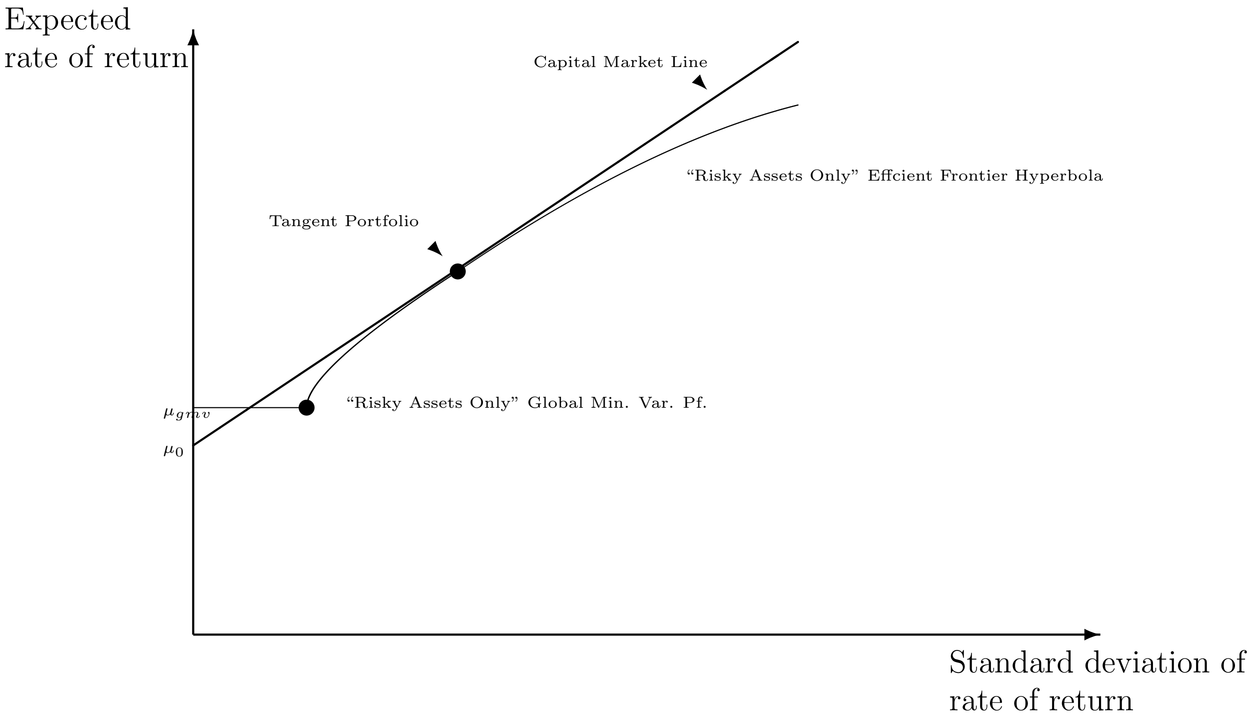


Figure 3.4: The capital market line.

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, and solve the first order conditions. This gives the optimal weights

$$\widehat{\mathbf{w}} = \frac{\mu_P^e}{(\mu^e)^\top \mathbf{\Sigma}^{-1} \mu^e} \mathbf{\Sigma}^{-1} \mu^e, \tag{3.16}$$

and the following expression for the variance of the minimum variance portfolio with mean excess return μ_P:

$$\widehat{\sigma}_P^2 = \frac{(\mu_P^e)^2}{(\mu^e)^\top \mathbf{\Sigma}^{-1} \mu^e}. \tag{3.17}$$

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 $(0, \mu_0)$ with a slope of $\sqrt{(\mu^e)^\top \mathbf{\Sigma}^{-1} \mu^e}$. This line is called the capital market line (CML). 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

$$\mu_{tan}^e = \frac{(\mu^e)^\top \mathbf{\Sigma}^{-1} \mu^e}{\mathbf{1}^\top \mathbf{\Sigma}^{-1} \mu^e},$$

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 3.4, and that explains why we use the term "tangency". When $\mu^e_{tan} > 0$, 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* (after William Sharpe) as excess return relative to standard deviation,

$$\text{Sharpe-ratio}_P = \frac{\mu_P - \mu_0}{\sigma_P}.$$

In the case where $\mu^e_{tan} > 0$, we see from Figure 3.4 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.) The observation that "Higher Sharpe-ratio is better. End of story." makes it a frequently used tool for evaluating and comparing the performance for investment funds. This may sound borderline trivial, but note that if instead we defined Sharpe-ratio with variance in the denominator, then there will be some ineffcient portfolios that have higher Sharpe-ratios that some mean-variance effcient portfolios.
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 5 (and its converse).

Proposition 6. *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:*

$$\mu_P - \mu_0 = \beta_{P,mv}(\mu_{mv} - \mu_0),$$

where

$$\beta_{P,mv} = \frac{Cov(r_P, r_{mv})}{\sigma^2_{mv}}.$$

Conversely, a portfolio for which these equations hold for all individual assets is *on the mean-variance frontier.*

3.1.3 Messing with your head: Effects of parameter uncertainty and the Black-Litterman model

Let $\mathbf{R} = (R_1, R_2, ..., R_N)^T$ be the **excess return** per unit time of N risky assets, i.e. let R_i be defined as the **rate of return** per unit time of the i^{th} asset *minus* the rate of return per unit time of the risk free asset for $i = 1, 2, ..., N$. The mean excess returns per unit time is represented by the vector $\boldsymbol{\mu} = E(\mathbf{R})$, and the covariance matrix per unit time is given by $\boldsymbol{\Sigma} = Var(\mathbf{R}) = E((\mathbf{R}-\boldsymbol{\mu})(\mathbf{R}-\boldsymbol{\mu})^T)$. Assume the excess returns to be normally distributed, i.e. $\mathbf{R} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Since the covariance matrix $\boldsymbol{\Sigma}$ by definition is positive definite, it follows that it admits a Cholesky decomposition:

$$\exists \boldsymbol{\sigma} \in \mathbb{R}^{N\times N} \text{ s.t. } \boldsymbol{\Sigma} = \boldsymbol{\sigma}\boldsymbol{\sigma}^T,$$

where $\boldsymbol{\sigma}$ is a lower triangular matrix. In particular, one may readily check that $\mathbf{R}$ thence can be written as

$$\mathbf{R} = \boldsymbol{\mu} + \boldsymbol{\sigma}\mathbf{Z},$$

where $\mathbf{Z}$ is a random vector $\Omega \mapsto \mathbb{R}^n$ with distribution $\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. More generally, the excess return vector over the time increment Δt is given by

$$\mathbf{R}_{\Delta t} = \boldsymbol{\mu}\Delta t + \boldsymbol{\sigma}\sqrt{\Delta t}\mathbf{Z},$$

in discrete analogy with geometric brownian motion.

Provided that $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known to us, we can compute the mean-variance frontier and the capital market line using standard techniques. Nevertheless, it remains unclear whether these quantities can be reliably estimated, and indeed what bearing a negative answer to this query might have on our capital allocation. To test just how stable the mean-variance frontier is for empirical estimates of the mean and covariance, we design the following experiment:

1. First, to get empirically plausible values for the estimators $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\Sigma}}$ we use daily empirical data for five risky assets based on Kenneth French's "five industry portfolios" and the "Farma-French 3-Factors" available for free at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`. Unbiased sample estimates for the components μ_i and Σ_{ij} are given by

$$\bar{\mu}_i := \frac{1}{n\Delta t}\sum_{k=1}^{n}(R_{\Delta t})_{ik}, \tag{3.18}$$

   and

$$\bar{\Sigma}_{ij} := \frac{1}{(n-1)\Delta t}\sum_{k=1}^{n}[(R_{\Delta t})_{ik} - \bar{\mu}_i\Delta t][(R_{\Delta t})_{jk} - \bar{\mu}_j\Delta t], \tag{3.19}$$

   where $\{(R_{\Delta t})_{ij}\}_{j=1}^{n}$ is a time series of n consecutive Δt excess returns for asset i.

2. For our present purposes we shall think of $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\Sigma}}$ as the "true" parameters of the market.

3. Using the equation $\mathbf{R}_{\Delta t} = \bar{\boldsymbol{\mu}}\Delta t + \bar{\boldsymbol{\sigma}}\sqrt{\Delta t}\mathbf{Z}$ where $\boldsymbol{\Sigma} = \bar{\boldsymbol{\sigma}}\bar{\boldsymbol{\sigma}}^T$ we now simulate m future evolutions of the five risky assets over a fixed temporal horizon.

4. For each future evolution in the simulated data, we re-compute sample estimates for the mean and the covariance matrix. Label these by $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ where $i = 1, 2, ..., m$. Obviously the *expected values* of the random variables $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ will be the "true" parameters $\bar{\boldsymbol{\mu}}$ and $\bar{\boldsymbol{\Sigma}}$.

5. Using the mean-variance equation

$$\hat{\sigma}_P^2(\mu_P) = \frac{a - 2b\mu_P + c\mu_P^2}{d}$$

or, equivalently,

$$\mu_p(\hat{\sigma}_P^2) = \frac{b}{c} \pm \sqrt{\frac{1}{c}\left[d\hat{\sigma}_P^2 + \frac{b^2}{c} - a\right]},$$

where $a = \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, $b = \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{1} = \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, $c = \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$, and $d = ac - b^2$ we plot the mean-variance frontier $(\hat{\sigma}_P^2, \mu_P)$ for the scenarios:

   (a) $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ (the simulated sample estimates).

   (b) $(\hat{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ (simulated sample mean, "true" covariance).

   (c) $(\bar{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ ("true" mean, simulated covariance).

6. We conjecture that the frontiers will be all over the place when we use the simulated sample mean, $\hat{\boldsymbol{\mu}}$. The simple explanation for this is that the real parameter $\bar{\boldsymbol{\mu}}$ is difficult to estimate reliably (as we have seen elsewhere, for log-returns the estimator is a telescoping sum, meaning that only the first and last data point in the stock price process end up determining the expected return). A similar problem does not pertain to the covariance.

The true mean ($\bar{\boldsymbol{\mu}}$) and covariance matrix ($\bar{\boldsymbol{\Sigma}}$) for French's five industry portfolios are exhibited in tables 3.1 and 3.2. The estimators are based on daily data points collected over the 20 year horizon July 1995 to July 2015. The associated mean-variance frontier, market portfolio, and capital market line (CML) are exhibited in the left-hand part of Figure 3.5. Notice that both the frontier and the CML are parabolic functions in (variance, mean)-space - had we plotted the corresponding curves in (standard deviation, mean)-space they would respectively transform to a hyperbola and a straight line. Furthermore, notice the trending inverse relationship between the expected return of the portfolios and their variances (marked by x in the figure): one would probably guess that taking on more risk (volatility) would be compensated by the promises of a higher expected return, but clearly this is not the case in this concrete empirical example!

Out of interest, the righthand side of figure 3.5 also exhibits the mean-variance frontier in the event we enforce the "no short-selling" restriction $\boldsymbol{\pi} \geq 0$.[1] Given the

[1] This problem must be solved numerically; e.g. using MATLAB's quadprog function. For consistency the depicted true mean-variance frontier has also been computed numerically.

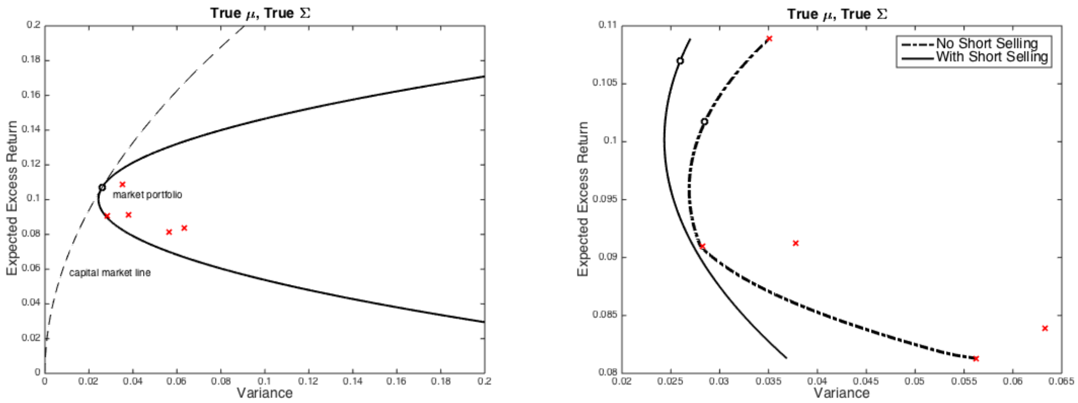


Figure 3.5: **Left:** The mean-variance frontier computed from five industrial indices, marked by x in the diagram. The white dot represents the associated market portfolio, while the dashed line connecting the origin and the market portfolio represents the capital market line. **Right:** The mean-variance frontier with and without short-selling restrictions.

linearity of portfolio returns, such a frontier can only be drawn between the data point with the lowest expected excess return and the data point with the highest expected excess return. Unsurprisingly, a ban on short-selling entails a dampening on the market portfolio from $(\hat{\sigma}_P^2, \mu_P) = (0.0259, 0.1070)$ to $(\hat{\sigma}_P^2, \mu_P) = (0.0285, 0.1017)$, and therefore more conservative Sharpe ratios for rational investors (from 0.6640 to 0.6026).

Asset 1	Asset 2	Asset 3	Asset 4	Asset 5
0.0909	0.0912	0.0839	0.1089	0.0813

Table 3.1: “True” mean excess return for the five industrial indices (French 1995-2015).

Asset 1	Asset 2	Asset 3	Asset 4	Asset 5
0.0282	0.0266	0.0311	0.0236	0.0336
0.0266	0.0378	0.0343	0.0252	0.0373
0.0311	0.0343	0.0633	0.0290	0.0448
0.0236	0.0252	0.0290	0.0351	0.0305
0.0336	0.0373	0.0448	0.0305	0.0563

Table 3.2: “True” covariance matrix for the five industrial indices (French 1995-2015).

Next we simulate 50 evolutions of the five industrial indices five years into the future. The mean-variance frontiers generated by the associated estimators $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ for $i = 1, 2, ..., 50$ are exhibited in figure 3.6. Immediately we notice that frontiers which utilise the sample estimator $\hat{\boldsymbol{\mu}}$ are *highly* scattered with respect to the “true” mean-variance frontier, irrespective of whether we use the “true” or the sample covariance. On the other hand, if we use the “true” drift $\bar{\boldsymbol{\mu}}$ the associated frontiers collapse to something resembling the "true" mean-variance frontier. The implication is clear: mis-specifications of the expected mean return of the risky assets will invariably have a devastating impact on the way rational investors think they should invest vis-a-vis how they ought to invest given full information about the governing dynamics. This form of model mis-specification can significantly curb their welfare gains. The problem, of course, is that the "true" drift $\bar{\boldsymbol{\mu}}$ is notoriously unreliable to estimate (recall that for log returns only the first and final data points in the time series go into the estimation). This point is highlighted in the lefthand part of figure 3.7 where we plot the sample parameter $\hat{\boldsymbol{\mu}}$ across the different simulations. Clearly, the simulated drift estimators oscillate wildly around their "true" counterparts. Given the quadratic nature of the covariance estimator, an analogous problem does not prevail here: there is no significant information/welfare loss affecting rational investors in deploying $\hat{\boldsymbol{\Sigma}}$ over $\bar{\boldsymbol{\Sigma}}$.

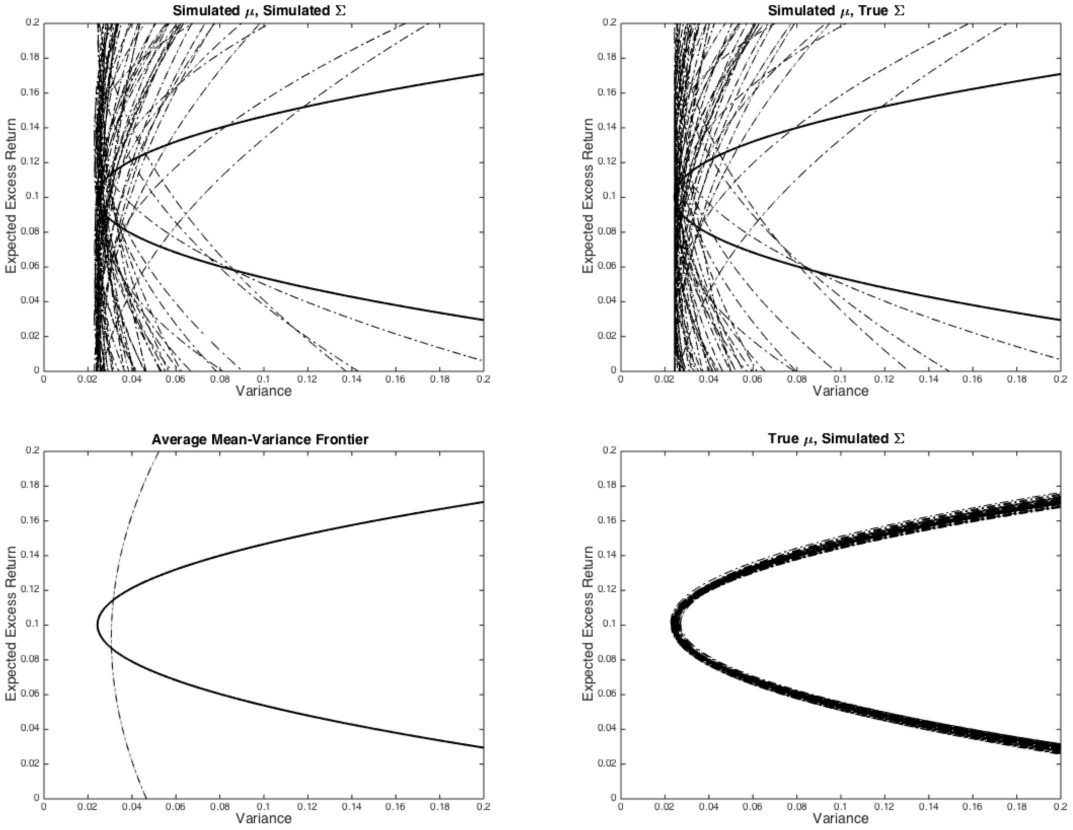


Figure 3.6: **Top left:** The mean-variance frontiers for $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ (the simulated sample estimates). **Top right:** The mean variance-frontiers for $(\hat{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ (simulated sample mean, "true" covariance). **Bottom left:** The average mean-variance frontier for the simulated sample estimates. Specifically, the dashed line represents the horizontal average of the dashed lines in the top left figure. **Bottom right:** The mean variance-frontiers for $(\bar{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ ("true" mean, simulated covariance).

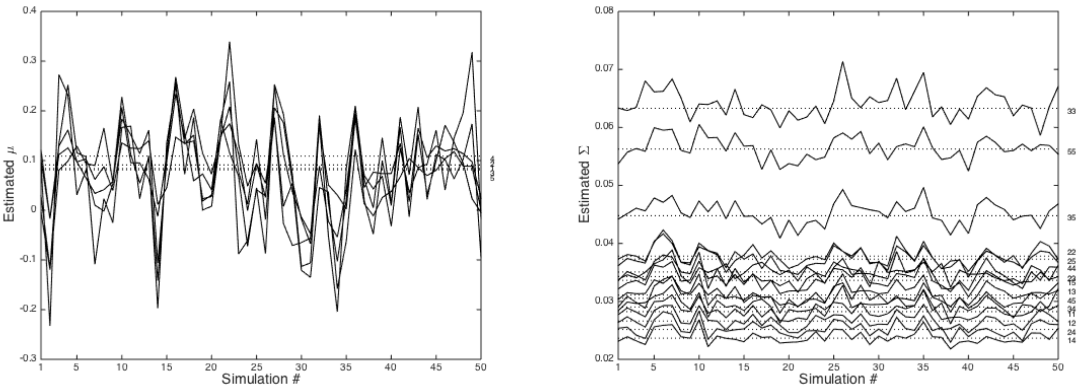


Figure 3.7: **Left:** The stability of the estimator $\hat{\boldsymbol{\mu}}$ across different simulations. The dotted lines represent the "true" means. Clearly, the estimators are highly erratic. **Right:** The stability of the estimator $\hat{\boldsymbol{\Sigma}}$ across different simulations. The dotted lines represent the "true" covariances. Clearly, the estimators are stable.

The Black-Litterman model

Consider an investor solving

$$\max_w w^\top \mu - \frac{\gamma}{2} w^\top \Omega w.$$

Here γ can be interpreted as a (relative) risk-aversion parameter and reasonable values are between 1 and 5. The optimal portfolio weights are

$$\widehat{w} = (\gamma\Omega)^{-1}\mu. \tag{3.20}$$

It is not unreasonable to think of Ω as known. Less so for μ. Imagine now that we reverse the question: Given a postulated risk-aversion and observed portfolio weights, w^{obs}, what should the expected returns be for these weights to be optimal? These we call the equilibrium expected returns, denote by Π, and immediately see that

$$\Pi = \gamma\Omega w^{obs}.$$

A model for noisy equilibrium expected returns is

$$\Pi = I_n \mu + e$$

where $e \sim N(0, \tau\Omega)$ and we call τ the precision.
A model for our views, V, on expected returns as well as our confidence in these, is

$$V = P\mu + v$$

where P is a $k \times n$ (known/specified) matrix and the error $v \sim N(0, \Sigma)$ is assumed independent of e.

As an example, suppose you are very certain in your expectation that stock 1 will outperform stock 3 by 2% on (yearly) average. Then $V_1 = 0.02$, $P_{1,1} = 1$, $P_{1,3} = -1$ and all elements of the first row (and column) of Σ are small.

A model specification combining equilibrium and views is

$$\begin{bmatrix} \Pi \\ V \end{bmatrix} = \begin{bmatrix} I_n \\ P \end{bmatrix} \mu + \epsilon,$$

where $\epsilon \sim N(0, A)$ with

$$A = \begin{bmatrix} \tau\Omega & 0 \\ 0 & \Sigma \end{bmatrix}.$$

After multiplying both sides by $A^{-1/2}$ (i.e. standardizing) we can view this as an ordinary regression model for the unknown expected return μ, which we can then estimate using the usual formula $\hat{\beta} = (X^\top X)^{-1} X^\top y$. Straightforward linear algebra leads to the Black-Litterman formula

$$\widehat{\mu} = [(\tau\Omega)^{-1} + P^\top \Sigma^{-1} P]^{-1} [(\tau\Omega)^{-1}\Pi + P^\top \Sigma^{-1} V]. \tag{3.21}$$

In the original Black-Litterman paper there is more than a touch of mystery about this formula. In practice we would then get portfolio weights consistent with our wievs by plugging $\widehat{\mu}$ from eqn. (3.21) back into (3.20). If we are absolutely certain about our views ($\Sigma = 0$), the formula degenerates to

$$\widehat{\mu} = \Pi + \tau\Omega P^\top (\tau P \Omega P^\top)^{-1} (V - P\Pi). \tag{3.22}$$

3.2 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 μ_i denote the rate of return on the i'th risky asset and we let $\mu_0 = r_0$ denote the riskless rate of return. We assume that μ_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

$$Er_i - \mu_0 = \beta_{i,tan}(E(r_{tan}) - \mu_0) \tag{3.23}$$

where

$$\beta_{i,tan} = \frac{Cov(r_i, r_{tan})}{\sigma_{tan}^2}. \tag{3.24}$$

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

$$w^j = \frac{P_0^j}{\sum_{i=1}^n P_0^i}. \tag{3.25}$$

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 (3.23). This is the CAPM:

$$E(r_i) - \mu_0 = \beta_{i,m}(E(r_m) - \mu_0), \tag{3.26}$$

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 the rates of return are multivariate normal and that the utility function is twice differentiable and concave:[2]

$$\begin{aligned}&\max_{\mathbf{w}} E(u_i(W_1^i))\\ &\text{s.t. } W_1^i = W_0(\mathbf{w}^\top \mathbf{r} + (1 - \mathbf{w}^\top \mathbf{1})r_0).\end{aligned}$$

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

$$E\left(u_i'(W_1^i)(r_j - r_0)\right) = 0. \tag{3.27}$$

Remembering that $\text{Cov}(X,Y) = EXY - EXEY$ we rewrite this as

$$E\left(u_i'(W_1^i)\right) E(r_j - r_0) = -\text{Cov}(u_i'(W_1^i), r_j).$$

A result known as Stein's lemma says that for bivariate normal distribution (X,Y) we have

$$\text{Cov}(g(X),Y) = Eg'(X)\text{Cov}(X,Y)$$

and using this we have the following first order condition:

$$E\left(u_i'(W_1^i)\right) E(r_j - r_0) = -Eu_i''(W_1^i)\text{Cov}(W_1^i, r_j)$$

i.e.

$$\frac{-E\left(u_i'(W_1^i)\right) E(r_j - r_0)}{Eu_i''(W_1^i)} = \text{Cov}(W_1^i, r_j).$$

Now define the following measure of agent i's absolute risk aversion:

$$\theta_i := \frac{-Eu_i''(W_1^i)}{Eu_i'(W_1^i)}.$$

[2] This derivation follows Huang and Litzenberger: *Foundations for Financial Economics.* If prices are positive, then returns are bigger than -1, so normality must be an approximation.

Summation over all agents gives us

$$\begin{aligned} E(r_j - r_0) &= \frac{1}{\sum_{i=1}^{K} \frac{1}{\theta_i}} \mathrm{Cov}(W_1, r_j) \\ &= \frac{1}{\sum_{i=1}^{K} \frac{1}{\theta_i}} W_0 \mathrm{Cov}(r_m, r_j), \end{aligned}$$

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

$$Er_m - \mu_0 = W_0^M \mathrm{Var}(r_m) \frac{1}{\sum_{i=1}^{K} \frac{1}{\theta_i}},$$

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. Let the expected value and the covariance of end-of-period asset prices for all assets be given.

2. Suppose further that we are 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 riskfree 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 or placed in the money market at the riskless rate – depending on the sign. (In equilibrium where all assets are being held impliying 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, and thus refrerred to as Roll's cirtique , 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? At the intuitive level, the relation (3.23) 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.

3.3 Remarks on CAPM in no particular order

3.3.1 The single index model

The CAPM says that for any stock or portfolio ($\sim i$) we have

$$\mathrm{E}(r_i) = r_f + \beta \times (\mathrm{E}(r_M) - r_0),$$

where $\beta = \mathrm{cov}(r_i, r_M)/\mathrm{var}(r_M)$, M indicates the market portfolio, and r_f is the risk-free rate. When viewing the expeted rate of return as a function of β, this relation is called the *security market line.*

The sequrity maket line or the CAPM-equation is often expressed in terms of random variables,

$$r_i - r_f = \alpha_i + \beta_i(r_M - r_f) + \epsilon_i, \tag{3.28}$$

where the noise-term ϵ_i is assumed to be independent of the market rate of return r^M. This is called the single-index model and α_i is called Jensen's α. To an empiricist it screams "regression" – in which case people add time-indicies and usually work with excess rates of return, $\widetilde{r}_{i,t} = r_{i,t} - r_{f,t}$.

In the single index model (3.28) we have

$$\mathrm{var}(r_i) = \beta_i^2 \mathrm{var}(r_M) + \mathrm{var}(\epsilon_i).$$

The first term is is called systematic risk, the second term unsystematic or idiosyncratic risk. The reason behind this terminology is the following: We know that there exists a portfolio that has the same expected rate of return as asset i but whose variance is $\beta_i^2 \mathrm{var}\,(r_M)$ – namely the portfolio that has $1 - \beta_i$ in the riskfree asset and β_i, in the market portfolio. On the one hand, because this portfolio is efficient, we cannot obtain a lower variance if we want an expected rate of return of $E(r_i)$. Hence this variance is a risk that is non-diversifiable, i.e. it cannot be avoided if we want an expected rate of return of $E(r_)i$. On the other hand as we have just seen the risk represented by the term $\mathrm{var}(\epsilon_i)$ can be avoided simply by choosing a different portfolio that a better job of diversification without changing expected rate of return.

Note that for the single index model we have

$$\mathrm{cov}(r_i, r_M) = \mathrm{cov}(\beta_i r_M + \epsilon_i, r_M) = \beta_i \mathrm{cov}(r_M, r_M),$$

so the CAPM-form of β_i is still valid, $\beta_i = \text{cov}(r_i, r_M)/\text{var}(r_M)$.

We also see that the single index model has the CAPM as the special case $\alpha_i = 0$ – seemingly a testable restriction. But careful: With muliple assets this must be viewed and tested as a joint hypothesis across assets and Proposition 6 tells us that it is really a restriction or condition on the market (or reference) portfolio M.

Example 14. (An empirical tesing caveat) When testing $\alpha_i = 0$ across assets i, a tempting assumption to make is that all errors terms, ϵ_i's, are uncorrelated. But let's multiply each of the equations (over i) in (3.28) with its corresponding market portfolio weight, say w_i, and sum over the is. This gives

$$\sum_i w_i r_i = \sum_i w_i \alpha_i + r_M \sum_i w_i \beta_i + \sum_i w_i \epsilon_i$$

Now, the left-hand side is just r_M. Let's call the first term on the right hand side $\bar{\alpha}$ and note that the second sum (by definition of β's and the market) is 1. From this we get that

$$\sum_i w_i \epsilon_i = -\bar{\alpha}.$$

So there is a linear combination of the random ϵ_i's that is equal to a non-random constant. Then not only can the ϵ_i's not be uncorrelated, but their covariance matrix must be degenerate. ■

3.3.2 CAPM as a pricing model

Consider a model with n risky assets with expected rate of return vector μ and invertible covariance matrix Σ, and put $\mathbf{1}^\top = (1, \ldots, 1)$. A slight but convenient reparametrization of the search for efficient portfolios is to solve

$$\max_w w^\top \mu - \frac{1}{2}\gamma w^\top \Sigma w \text{ s. t. } w^\top \mathbf{1} = 1,$$

for different values of γ, which can be interpreted as a risk-aversion parameter. The optimal portfolios are

$$\widehat{w} = \gamma^{-1}\Sigma^{-1}\left(\mu - \eta(\gamma; \mu, \Sigma)\mathbf{1}\right)$$

where $\eta(\gamma; \mu, \Sigma) = (\mathbf{1}^\top \Sigma^{-1}\mu - \gamma)/\mathbf{1}^\top \Sigma^{-1}\mathbf{1}$ can be interpreted as the expected rate of return on $\widehat{w}$'s zero-beta portfolio.

It seems intuitively reasonable that $\partial \widehat{w}_i / \partial \mu_i > 0$, meaning that if an asset's expected rate of return goes up, then so does its weight in any efficient portfolio. Assets for which this does not hold, we could call financial Giffen goods. We will now show that in the mean/variance optimization setting, there are no financial Giffen goods. To do this we look at the problem with the modified expected return vector $\mu + \alpha e_i$, where $\alpha \in \mathbb{R}$ and e_i is the i'th unit vector. The optimal portfolio in this case we can write as

$$\widehat{w}(\alpha) = \widehat{w} + \alpha h,$$

where $h = \gamma^{-1}(\Sigma^{-1}e_i - \frac{e_i^\top \Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}}\Sigma^{-1}\mathbf{1})$. Showing that $\partial \widehat{w}_i / \partial \mu_i > 0$ amounts to proving positivity of the i'th coordinate of h, which we can write as

$$e_i^\top h = \gamma^{-1}\left(e_i^\top \Sigma^{-1} e_i - \frac{(e_i^\top \Sigma^{-1}\mathbf{1})^2}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}} \right).$$

Because Σ^{-1} is strictly positive definite and symmetric, $x^\top \Sigma^{-1} y$ defines an inner product, and strict positivity of the term in parenthesis on the right hand side of the equation above follows immediately from the Cauchy-Schwartz inequality.

The inclusion of a risk-free asset is handled in the same way with η replaced by the risk-free rate of return because the risk-free asset is any portfolio's zero-beta portfolio.

With this result a newly introduced $(n+1)$'st asset (or "project") will be in positive demand (or: "attractive") precisely if there is strict inequality in the CAPM-like expression

$$\mu_{n+1} - r > \frac{\mathrm{cov}(r_{n+1}, r_M)}{\mathrm{var}(r_M)}(\mu_M - r), \tag{3.29}$$

where M denotes the market (or tangent) portfolio, and rs with subscripts are (stochastic) rates of returns. We know from Proposition 6 that w is mean-variance efficient precisely if for any individual asset i we have

$$\mu_i - r = \frac{\mathrm{cov}(r_i, r_w)}{\mathrm{var}(r_w)}(\mu_w - r).$$

For the portfolio $(w_M^\top, 0)^\top$ the n first necessary equations hold because the market portfolio is efficient in the old economy, and we see that the new asset is in 0-demand if equality holds in (3.29). Now the absence of Giffen tells us that if there is strict inequality as stated, the $(n+1)$'st asset has strictly positive weight in the new market portfolio.
This gives us a theoretically well-founded way to evaluate projects in the way illustrated by the following example.

Example 15. Consider a setting where

$$\mathrm{E}(r_M) = 0.07, \;\; \sigma_M := \sqrt{\mathrm{var}(r_M)} = 0.15 \text{ and } r_0 = 0.04.$$

Suppose we are given the opportunity (at time 0) to invest in a project that pays at time 1 the stochastic amount X_1 about which we know that

$$\mathrm{E}(X_1) = 1000, \;\; \sigma_X := \sqrt{\mathrm{var}(X_1)} = 400, \text{ and } \rho_{M,X} := \mathrm{corr}(r_M, X_1) = 0.5.$$

An interpretation could be: The project is an oil field and our future income depends partly on how much oil is there, partly on the price at which we can sell it in the market. Only the latter is correlated with the (general) financial market, the oil is either there or it isn't. So what is the CAPM-critical price at time 0, say X_0? Written out in detail, the CAPM says

$$\frac{\mathrm{E}(X_1) - X_0}{X_0} = r_0 + \frac{\mathrm{cov}(\frac{X_1 - X_0}{X_0}, r_M)}{\sigma_M^2}(\mathrm{E}(r_M) - r_0), \tag{3.30}$$

which we have to solve for the critical time 0-price X_0^*. To do this let us look at

$$\begin{aligned}
\mathrm{cov}\left(\frac{X_1 - X_0}{X_0}, r_M\right) &= \frac{1}{X_0}\mathrm{cov}(X_1, r_M) \\
&= \frac{1}{X_0}\sigma_X \sigma_M \rho_{M,X} \\
&= \frac{400 \cdot 0.15 \cdot 0.5}{X_0} = \frac{30}{X_0}.
\end{aligned}$$

We now rearrange equation (3.30) to get

$$\underbrace{\mathrm{E}(X_1)}_{1000} = X_0 \underbrace{(1 + r_0)}_{1.04} + \underbrace{\frac{30}{\sigma_M^2}(\mathrm{E}(r_M) - r_0)}_{40}$$

which leads to

$$X_0^* = \frac{960}{1.04} = 923.1.$$

So if the project cost less that 923.1, it is attractive, otherwise not. The equilibrium expected rate of return on the project, the left hand side of (3.30) evaluated at X_0^* is 0.0832, and the equilibrium beta of the project is $\beta_X = 30/(X_0^* \sigma_M^2) = 1.44$. Note the quantitative caveat: beta (1.44 here) is not correlation (0.5 here). ■

3.3.3 Messing with your head: How security market lines actually look

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?

Second (and this a very important problem), what is the market portfolio? Since investment 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'.

According to the capital asset pricing model, CAPM, the expected excess return of a risky asset is proportional to the expected excess return of the market portfolio. The constant of proportionality (the so-called beta) is given by the covariance between the (rate of) return of that risky asset and the (rate of) return on the market portfolio *scaled* such that beta of the market portfolio itself is unity:

$$\beta_i = \frac{Cov(r_i, r_m)}{Var(r_m)}.$$

Thus, we may construe the β of an asset as encoding its susceptibility to systemic risk: the higher the beta the higher the co-movement with the market portfolio (and accordingly also the more prone the asset will be to plummet insofar as the market crashes).

In a perfect market, if we were to plot the excess return of all risky assets against their beta, they would form a straight line in the $(\beta, E[r_i] - r_f)$ diagram, with a slope of $E[r_m] - r_f$ and an ordinate intercept of zero. This is the so-called security market line (SML). Any asset falling above it is said to have positive (Jensen) alpha and is under-valued with respect to its level of uncertainty; conversely, any asset falling below the SML is said to have negative alpha and is over-valued.

To test the empirical verisimilitude of this model we have in Figure 3.8 plotted the excess return of 49 industrial portfolios against their beta. The data for the associated linear regression model

$$\text{excess return}_i = \text{slope} \cdot \text{beta}_i + \text{intercept} + \text{error term}_i,$$

is exhibited in table 3.3. Immediately we notice that the slope of the empirical sequrity market line is much flatter than the one predicted by the theoretical CAPM (gradient 0.0128 versus 0.0808). This has the somewhat surprising implication that taking on more systemic risk (higher β) is not "sufficiently compensated" in terms of added expected excess return. Low β stocks therefore seem more appetising.

	Estimate	SE	t-Stat	p-value
Intercept:	0.080798	6.3174e-10	1.279e+08	3.3455e-114
Slope:	0.012853	6.7344e-10	1.9086e+07	8.2528e-102

Table 3.3: Linear regression data of 49 industrial indices (French 1995-2015).

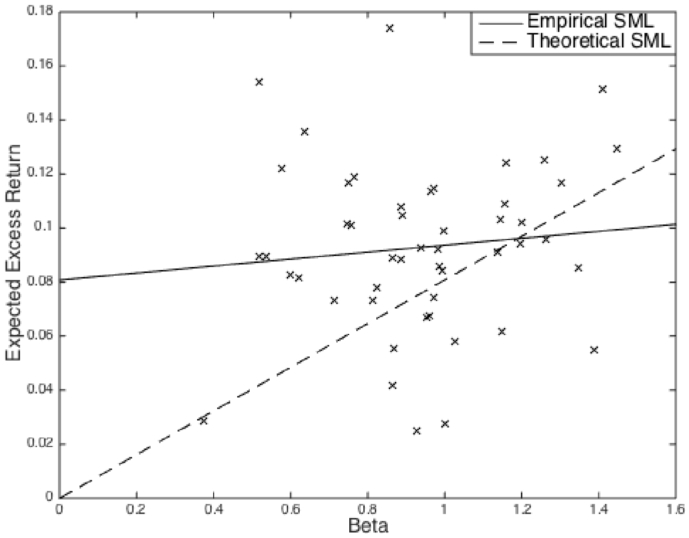


Figure 3.8: The empirical security market line computed on behalf of 49 industrial portfolios (French 1995-2015). On the same graph the theoretical "CAPM" counterpart has been plotted.

Chapter 4

No such thing as a free lunch: Arbitrage pricing in one-period models

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.

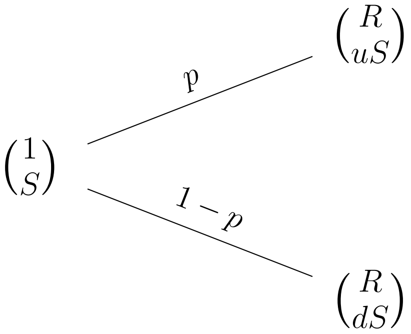
4.1 An appetizer

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. (It is tradition in probability to indicate states, possible outcomes by ω.) At time $t = 0$ nothing is known about the 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.[1]
- A money market account or bank 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*[2] on the stock with exercise (or strike) price K and expiry (sometimes called maturity, although this is primarily used for bonds) 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 replication 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

[1] We use *stock* as a generic term for a risky asset whose stochastic price behaviour we take as given. Words such as “share” or “equity” are largely synonymous.

[2] The term “European” is used for historical reasons; it has no particular meaning today, and it is doubtful if it ever had.

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

$$\begin{aligned}\frac{(C_u - C_d)}{S\,(u-d)}S + \frac{1}{R}\frac{uC_d - dC_u}{(u-d)} \cdot 1 &= \frac{R\,(C_u - C_d)}{R\,(u-d)} + \frac{1}{R}\frac{uC_d - dC_u}{(u-d)} \\ &= \frac{1}{R}\left[\frac{R-d}{u-d}C_u + \frac{u-R}{u-d}C_d\right].\end{aligned}$$

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.

Remark 1. *Notice that the hedge position in the stock (a) is positive, whilst the hedge position in the bank (b) is negative (i.e. we borrow money). This is quite apparent from the condition $d < R < u$: there are three different scenarios which arise: (I) $C_u > 0$ and $C_d > 0$, (II) $C_u > 0$ and $C_d = 0$, and the trivial case (III) $C_u = C_d = 0$. Clearly, nobody would bother hedging (III). However, (I) and (II) are readily shown to have $a > 0$ and $b < 0$.*

We are now able to draw the following highly significant conclusions about the valuation of risky securities:

- The physical probability p plays **NO** role in the expression for C_0. The fair price of the option is

$$\textbf{NOT:}\quad C_0 = E[C_1(\omega)] = pC_u + (1-p)C_d,$$

as one would perhaps initially conjecture.

- Rather, the fair price is given by the expression

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

where $q = \frac{R-d}{u-d}$ and $1-q = \frac{u-R}{u-d}$.

- q and $1-q$ formally satisfy the requirements of being *probability weights* (i.e. $\{q, 1-q\} \in (0,1)$ and $q + (1-q) = 1$) as per the criterion $d < R < u$. To see this, consider weight q: since $R > d$ and $u > d$ it immediately follows that $q > 0$. On the other hand, suppose $q \geq 1$: then $R - d \geq u - d \Leftrightarrow R \geq u$ which contradicts $u > R$. So $q < 1$.

- Hence we may equivalently write the valuation formula as the Q-expectation of the discounted payoff

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

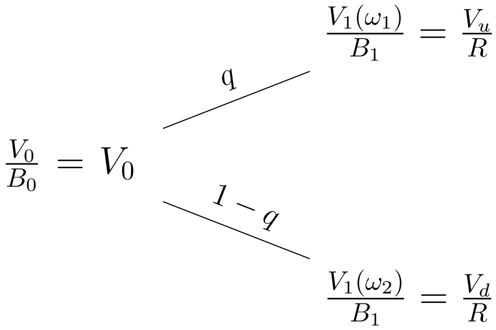
where Q is defined such that $Q(\omega_1) = q$, $Q(\omega_2) = 1-q$.

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

$$V_0 = E^Q\left[\frac{V_1(\omega)}{R}\right] = q\frac{V_u}{R} + (1-q)\frac{V_d}{R}, \tag{4.1}$$

where $V_u = V_1(\omega_1)$ and $V_d = V_1(\omega_2)$ and q is as given above.

More prosaically we might restate the implication of (4.1) as follows: suppose we use the money market account B as our *numeraire asset* - i.e. suppose we measure the value of the security V in terms of how many units of B it corresponds to. Then the fair price of V at time zero [measured in units of the initial money market account (=1)] is related to the terminal price of V [measured in units of the terminal money market account (now $= R$)] *as though* the up state occurs with a probability of q and the down state occurs with a probability of $1 - q$. Since the numeraire asset B is manifestly deterministic and therefore void of any financial risk, it is customary to refer to Q as the *risk neutral measure*.


$\frac{V_1(\omega_1)}{B_1} = \frac{V_u}{R}$
q
$\frac{V_0}{B_0} = V_0$
1 − q
$\frac{V_1(\omega_2)}{B_1} = \frac{V_d}{R}$

Now this exposition is bound to raise some questions: what (if anything) is so special about the money market account? Couldn't we have priced the security in terms of some other numeraire asset (say, the stock price process S)? The short answer is that there is nothing per se which singles out the money market account as the preferred numeraire. In fact, provided that we perform an *equivalent change of probability measure*, we can gracefully move to whichever numeraire asset tickles our fancy.[3] To see how this works out, consider changing the measure in (4.1) from Q to some $Q' := \xi^{-1}Q$ where ξ is a non-negative random variable

$$\begin{aligned} V_0 = E^Q[R^{-1}V_1(\omega)] &= R^{-1}\left(Q(\omega_1)V_1(\omega_1) + Q(\omega_2)V_1(\omega_2)\right) \\ &= R^{-1}\xi\left(Q'(\omega_1)V_1(\omega_1) + Q'(\omega_2)V_1(\omega_2)\right) \\ &= E^{Q'}[R^{-1}V_1(\omega)\xi]. \end{aligned}$$

In particular, suppose we specify ξ such that

$$\xi = \frac{RS}{S_1(\omega)},$$

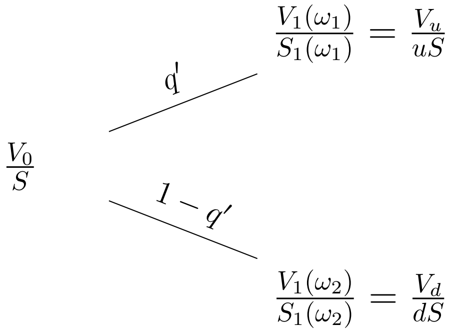
then

$$V_0 = SE^{Q'}\left[\frac{V_1(\omega)}{S_1(\omega)}\right] = q'\frac{V_u}{u} + (1 - q')\frac{V_d}{d}, \tag{4.2}$$

[3] Two probability measures Q and Q' are said to be equivalent provided that they agree on which events have probability zero. We donate this property by $Q \sim Q'$.

where $q' = Q'(\omega_1) = \xi^{-1}Q(\omega_1) = (RS/(uS))^{-1}q = uq/R$. Make sure that you see that (4.2) is consistent with formula (4.1).

We may thus repeat the conclusion above in an analogous manner: suppose we use the stock price S as our *numeraire asset* - i.e. suppose we measure the value of the security V in terms of how many units of S it corresponds to. Then the fair price of V at time zero [measured in units of initial stock (=S)] is related to the terminal price of V [measured in units of terminal stock (= uS or = dS)] *as though* the up state occurs with a probability of q' and the down state occurs with a probability of $1 - q'$.


$\frac{V_1(\omega_1)}{S_1(\omega_1)} = \frac{V_u}{uS}$
q'
$\frac{V_0}{S}$
1 − q'
$\frac{V_1(\omega_2)}{S_1(\omega_2)} = \frac{V_d}{dS}$

Whilst this numeraire-invariance (up to a change a measure) of the valuation formula for risky securities is a neat theoretical result, one must inevitably wonder whether the result carries any practical implications. What could possibly warrant a preference for formula (4.2) over (4.1)? In discrete time "not too much" is generally the answer. However, upon moving to continuous time finance, an apt choice of numeraire can have a profound impact on our quest for a closed form option pricing formula: a seemingly impenetrable valuation exercise under one probability measure, may decompose into a few lines of routine calculations under another. Indeed the risk neutral measure is not always to be preferred.

Example 16. Suppose we try to value a security which pays out whatever is the stock price according to the framework above. Obviously, for our model to be consistent with the absence of arbitrage, the time zero value should be equal to S. Let's check this. From (4.1):

$$\begin{aligned} V_0 = E^Q\left[\frac{S_1(\omega)}{R}\right] &= \left(\frac{R-d}{u-d}\right)\frac{1}{R}(uS) + \left(\frac{u-R}{u-d}\right)\frac{1}{R}(dS) \\ &= \frac{1}{(u-d)R}(RuS - duS + udS - RdS) \\ &= S. \end{aligned}$$

Equivalently, if we use equation (4.2) we find

$$V_0 = SE^{Q'}\left[\frac{S_1(\omega)}{S_1(\omega)}\right] = S(q' + (1-q')) = S,$$

as desired.

Example 17. Finally, let us show that the call option price C_0 is increasing in the interest rate R. This is quite apparent upon remembering that

$$C_0 = aS + b,$$

where

$$a = \frac{C_u - C_d}{S(u-d)}, \quad \text{and} \quad b = \frac{1}{R}\frac{uC_d - dC_u}{(u-d)}.$$

It follows that as R increases R^{-1} decreases whence b decreases. But (as a bit of algebra shows; this relies explicitly on the call-option's payoff structure), $b < 0$ so C_0 increases. Simply put: a call option increases with the interest rate because borrowing becomes more expensive.

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 financial market or a security price system consists of 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*

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

or

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

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

Remark 2. *Our conventions when using inequalities on a vector in $\mathbb{R}^k$ are the same as described in Chapter 2.*

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 2.

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*

$$\pi = D\psi$$

Sometimes ψ is called a state-price density, or its elements are referred to as Arrow-Debreu-prices and the term Arrow-Debreu attached to the elementary securities. Clearly, we have already proved the following in Chapter 2:

Proposition 7. *A security price system is arbitrage-free if and only if there exists a state-price vector.*

Unlike the model we considered in Chapter 2, 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 riskfree 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 riskfree asset. Note that when a riskfree 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 riskfree 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

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

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 riskfree 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* $D^\top \theta_c = c$.

Proposition 8. *Let an arbitrage-free system* (π, D) *and a redundant security c by given. The augmented system* $(\widehat{\pi}, \widehat{D})$ *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*

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

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*

$$D^\top \theta = y \tag{4.3}$$

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

Proposition 9. *An arbitrage-free market is complete if and only if the state-price vector is unique.*

The proof is exactly as in Chapter 2 and we are ready to price new securities in the financial market; also known as pricing of *contingent claim.*[4]
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

$$\pi(c) = \frac{1}{R_0} E^q(c) \equiv \frac{1}{R_0} \sum_{i=1}^{S} q_i c_i, \tag{4.4}$$

where $q_i = \frac{\psi_i}{d_0} \equiv R_0 \psi_i$. Again, the method in Equation (4.4) (and the generalizations of it we'll meet in Chapters 5 and 6) is referred to as *risk-neutral pricing*. Arbitrage-free prices are calculated as discounted expected values (with some new or artificial probabilities, the qs), i.e. as if agents were risk-neutral. But the "as if" is important to note:

[4] A contingent claim just a random variable describing pay-offs; the pay-off is *contingent* on ω. The term (financial) derivative (asset, contract, or security) is largely synonymous, except thatwe are usually more specific about the pay-off being contingent on another financial asset such as a stock. We say option, even when we more specific pay-off structures in mind.

> **No assumption of actual agent risk-neutrality is used to derive the risk neutral pricing formula (4.4) - just that they prefer more to less (see the next subsection). As a catch-phrase: "Risk-neutral pricing does not assume risk-neutrality".**

The measure Q might therefore be construed as a mathematical convenience tool, which allows us to do arbitrage free valuation. Of course, the existence of Q in turn depends on whether the financial market (we, the agents) have valued existing assets consistently (i.e. without arbitrage). Indeed, its uniqueness (our ability to extract just one arbitrage free price) depends on whether the market is complete. But in the real world this assumption is obviously extremely hard to check, which has brought some skeptics to voice their dissatisfaction with the risk-neutral pricing enterprise.

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

$$(\pi, D) = \left(\begin{pmatrix} 1 \\ S \end{pmatrix}, \begin{pmatrix} R & R \\ uS & dS \end{pmatrix} \right).$$

For this to be arbitrage-free, proposition (7) 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 D has full rank. $u > d$ can be assumed without loss of generality. We find

$$\psi_1 = \frac{R - d}{R(u - d)}, \quad \text{and} \quad \psi_2 = \frac{u - R}{R(u - d)},$$

and note that the solution is strictly positive precisely when $u > R > d$ (given our assumption that $u > d > 0$). The risk-free asset has a rate of return of $R - 1$, and

$$q_1 = R\psi_1 = \frac{R - d}{u - d}, \quad \text{and} \quad q_2 = R\psi_2 = \frac{u - R}{u - d},$$

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.

Example 18. Consider an arbitrage-free market comprised of three securities all valued at 2 (units of some currency) with associated pay-offs $\binom{2}{3}$, $\binom{1}{5}$, and $\binom{3}{1}$. Is the market complete? Is it arbitrage free? Suppose we introduce a fourth security with pay-off $\binom{0}{10}$. What is its fair price?

The security price system is

$$(\pi, D) = \left(\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 3 \\ 1 & 5 \\ 3 & 1 \end{pmatrix} \right).$$

Evidently the market is complete as we can form a basis in $\mathbb{R}^2$ from the existing securities (they do not all lie on the same line). Indeed, one security is redundant, which can be seen by noting that $2\binom{2}{3} - \binom{1}{5} = \binom{3}{1}$. To check for the absence of arbitrage, is equivalent to checking whether there exists a strictly positive vector $\psi = (\psi_1, \psi_2)^\top$ such that $\pi = D\psi$. To this end we notice that D qua its dimensionality is non-invertible. However, upon multiplying $\pi = D\psi$ through by $D^\top$ we have a system which is solvable. Indeed,

$$\psi = (D^\top D)^{-1} D^\top \pi \in \mathbb{R}^2_{++},$$

so the system is arbitrage-free (the reader should check that this is in fact the case). Finally, let's put a fair price on the (redundant) security $y = \binom{0}{10}$. Now we might try to do this by solving equation (4.3) as $\theta = (DD^\top)^{-1} Dy$, however, the matrix $DD^\top$ is singular. Instead, we pick an invertible sub-matrix D and perform the valuation accordingly (specifically, we pick two (independent) assets and solve the problem). E.g. using assets 1 & 2 we find that

$$\begin{pmatrix} 2 & 3 \\ 1 & 5 \end{pmatrix}^\top \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 10 \end{pmatrix} \Leftrightarrow \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} -\frac{10}{7} \\ \frac{20}{7} \end{pmatrix}.$$

Hence the no-arbitrage price is $-2 \cdot \frac{10}{7} + 2 \cdot \frac{20}{7} = \frac{20}{7}$. The important point is that we arrive at this price irrespective of which replicating securities we choose. Thus, the reader might like to verify that the security pairs $\{\binom{2}{3}, \binom{3}{1}\}$ and $\{\binom{1}{5}, \binom{3}{1}\}$ both entail the same price for $\binom{0}{10}$. ■

Example 19. Consider the following curious set-up: suppose there are three securities on the market with $t = 0$ prices 1, 1 and γ (measured in, say, £), where γ is a positive constant. At time $t = 1$ their pay-offs are determined based on the local temperature (T) in London: if $T \geq 20°$ the securities respectively pay out 1, 2 and 1 [£]. If $20° > T \geq 15°$ they pay out 1, 1 and γ [£]. Finally, if $T < 15°$ the securities pay out 1, 0 and 1 [£]. Is this market arbitrage-free? The security price system is

$$(\pi, D) = \left(\begin{pmatrix} 1 \\ 1 \\ \gamma \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 0 \\ 1 & \gamma & 1 \end{pmatrix} \right). \tag{4.5}$$

Note that the first security is risk free. To check whether the system is arbitrage free, we must establish whether there exists a strictly positive state price vector $\psi = (\psi_1, \psi_2, \psi_3)^\top$ such that $\pi = D\psi$. To this end, consider the inverse matrix

$$D^{-1} = \frac{1}{2(\gamma - 1)} \begin{pmatrix} 1 & \gamma - 1 & -1 \\ -2 & 0 & 2 \\ 2\gamma - 1 & 1 - \gamma & -1 \end{pmatrix}.$$

This is well-defined if and only if $\gamma \neq 1$. Assuming this to be the case, we readily find that $\psi = D^{-1}\pi = (0, 1, 0)^\top$ which clearly does **not** meet the requirement of strict positivity. $\gamma \neq 1$ thus corresponds to a complete market with arbitrage opportunities. What about the case $\gamma = 1$? From (4.5) we see that the market becomes incomplete: specifically, assets one and three are now identical (both in price and in pay-off). Any residual arbitrage should therefore be between the risk free asset and asset 2. To check if arbitrage obtains, let us look for a vector θ such that $\pi \cdot \theta < 0$ and $D^\top \theta \geq 0$. Without loss of generality set $\theta_3 = 0$, then we may recast this problem as the linear programme

$$\begin{array}{ll} \text{min} & \theta_1 + \theta_2 \\ \hline \text{s.t.} & \theta_1 + 2\theta_2 \geq 0, \\ & \theta_1 + \theta_2 \geq 0, \\ & \theta_1 \geq 0. \end{array}$$

Clearly, the second constraint is incompatible with a situation in which $\min \theta_1 + \theta_2 < 0$. Hence, $\gamma = 1$ corresponds to an incomplete market without arbitrage opportunities. ■

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 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, iPhones, ...) 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 an agent who has an endowment $e = (e_1, \dots, e_S) \in \mathbb{R}^S_+$. 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

$$\begin{aligned} \max_{\theta} \quad & U(e + D^\top \theta) \\ \text{s.t.} \quad & \pi^\top \theta \leq 0. \end{aligned} \tag{4.6}$$

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.

Proposition 10. *If there exists a portfolio θ_0 with $D^\top \theta_0 > 0$ then the agent can find a solution to the maximization problem if and only if (π, D) is arbitrage-free.*

Proof. **The "only if" part:** We want to show that if (4.6) admits an optimal solution, then there is no arbitrage - or, equivalently, if there is an arbitrage, then there is no solution to (4.6). Suppose there is an arbitrage portfolio $\tilde{\theta}$ and that $c^* = e + D^\top \theta^*$ is an optimal solution to (4.6). Let the arbitrage be of the first kind, i.e. $\pi^\top \tilde{\theta} \leq 0$ and $D^\top \tilde{\theta} > 0$, then the agent can add the arbitrage portfolio to θ^* and be strictly better of – which contradicts the assumption that θ^* is optimal. Now suppose the arbitrage is of the second kind, specifically the case where $\pi^\top \tilde{\theta} < 0$ and $D^\top \tilde{\theta} = 0$. Then the agent may invest the proceeds from the arbitrage portfolio into the portfolio θ_0 for which $D^\top \theta_0 > 0$ (the assumption that such a portfolio exists is a very mild condition). Once again, this will allow us to contradict the assumption that θ^* is the optimal portfolio.

The "if" part: We will now show that in the absence of arbitrage, there exists a solution to (4.6). To this end, it would be convenient to use the *extreme value theorem* which establishes that a continuous real-valued function on a nonempty compact space is bounded above and attains its supremum. That $X = \{e + D^\top \theta \in \mathbb{R}^S_+ | \theta \in \mathbb{R}^N, \pi^\top \theta \leq 0\}$ constitutes a non-empty compact (convex) space is readily demonstrated: by assumption it is not empty, and closure follows from the "$\leq$". Convexity is likewise trivial: if θ_1 and θ_2 are two arbitrary portfolios which both satisfy the conditions of X, then so does the portfolio $\lambda\theta_1 + (1-\lambda)\theta_2 \; \forall \lambda \in (0,1)$. Finally, we can argue for boundedness by contradiction: suppose the convex space X is unbounded, then each $c \in X$ has an associated *ray* i.e. an infinite straight line which can be traversed without leaving X.[5] However, such a ray corresponds to the existence of an arbitrage (why?), which contradicts our assumptions. Hence, X must be non-empty and compact and the extreme value theorem entails that (4.6) has a solution. □

In the proposition above we have made life considerably simpler for ourselves by defining the utility function for consumption ≥ 0; Having the $=$ there makes the proof much easier, e.g. concavity is not used. The results also holds in the "consumption must be '> 0'"-case, but we omit the proof.

The following Theorem 3 is an important and versatile result that we will refer to as the *state-price utility theorem.*

Theorem 3. *Assume that there exists a portfolio θ_0 with $D^\top\theta_0 > 0$. If there exists a solution θ^* to (4.6) and the associated optimal consumption is given by $c^* := e + D^\top\theta^* \gg 0$, 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 that is orthogonal to π is also orthogonal to $D\nabla U(c^*)$. So (in finite-dimensional vector space language) we have an inclusion for orthogonal complements ($\text{span}\{\pi\}^\perp \subseteq \text{span}\{D\nabla U(c^*)\}^\perp$) and hence the converse inclusion for the spans themselves. Or in concrete terms: We have

[5] The precise statement is: $\forall c \in X \; \exists h \in \mathbb{R}^S$ such that $h \neq 0$ and $\ell = \{x \in \mathbb{R}^S | x = c + th, t \geq 0\} \subset X$.

$\mu\pi = D\nabla U(c^*)$ for some $\mu \in \mathbb{R}$ showing that $\nabla U(c^*)$ is proportional to a state-price vector. Choosing a θ_0 with $D^\top\theta_0 > 0$ we know from no arbitrage that $\pi^\top\theta_0 > 0$ and from the assumption that the utility function is strictly increasing, we have $\nabla U(c^*)D^\top\theta_0 > 0$. Hence μ must be positive. □

To understand the implications of the state-price utility theorem, 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

$$\psi_i = \lambda p_i u'(c_i^*), \ \ i = 1, \ldots, S. \tag{4.7}$$

where λ is some constant of proportionality.

The martingale method of portfolio optimization. Suppose the market is complete. Then Theorem 3 effectively reduces the optimal portfolio problem (4.6) to a one-dimensional problem. Specifically, the left-hand side of (4.7) is determined from market data independently of the agent. So we divide λ and p_i over and take $(u')^{-1}$ (u is concave and smooth, so u' is a continuous and decreasing function, hence it has an inverse) thus determining c_i^*. (At least up to knowledge of the scalar λ, which in practical cases would be determined from the agent's budget constraint; λ has to be such that the time zero price of the agent's consumption is no more than what he has to spend.) This is known as Pliska's martingale method — and wit orks in multi-period models too.

Zero-level pricing. If the market is incomplete we may construe Theorem 3 in reverse order as a way to pin down "reasonable" risk-neutral probabilities. Assuming that a certain representative agent has chosen a certain portfolio as her optimal choice will help use pick a ψ among the many. If a new asset is introduced and priced according to the ψ coming from that specific agent's marginal utilities, then that agent will demand exactly 0 units of the new asset (to prove this check the first-order conditions for utility maximizationin the extended model with the specific price and 0 units of the new asset). This principle is therefore known as zero-level pricing.

Intuition. 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 (4.7) 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 – an insight similar to what the CAPM gave us.

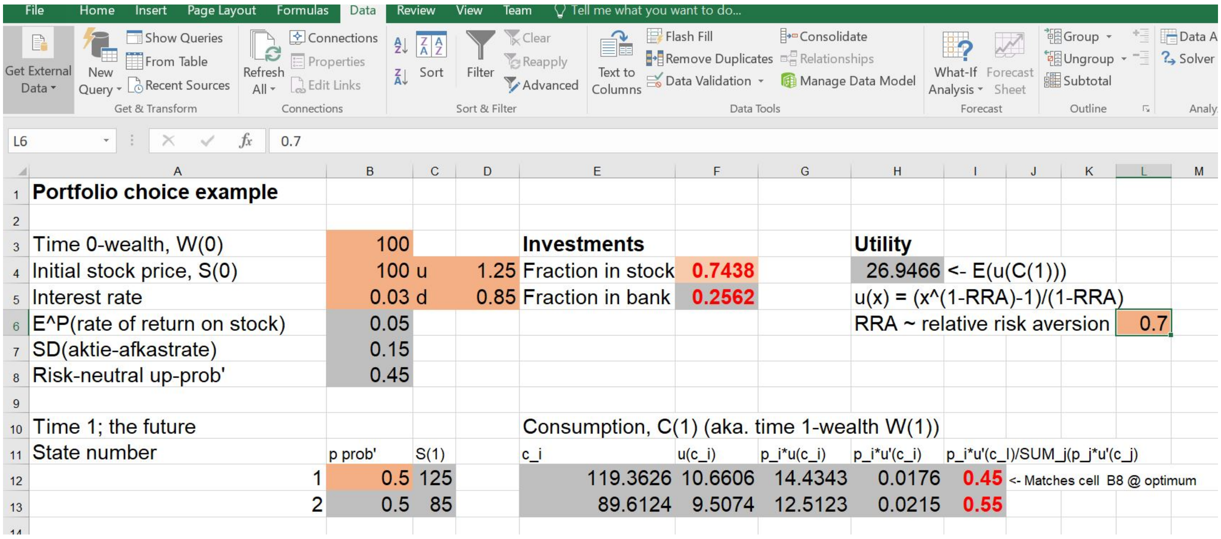
Example 20. Consider an investor who can choose between a (risky) stock and a (risk-free) bond investment, with the aim of maximising his expected terminal wealth. Specifically, he wants to solve the following optimisation problem

$$\max_{x_S,x_b} E(u(W(1))) \quad \text{s.t.} \quad \begin{aligned} &x_S S(0) + x_b \leq W(0), \\ &x_S S(1) + x_b(1+r) = W(1), \end{aligned} \tag{4.8}$$

where $W(1)$ denotes his (stochastic) time-1 wealth and his criterion or utility function has the form $u(x) = \frac{x^{(1-\gamma)}-1}{1-\gamma}$. The seemingly strange γ-parametrization is because in this way the agent's relative risk-aversion (generally defined as $-xu''/u'$) becomes exacty γ. In other words, this ustility function has constant relative risk-aversion. (For $\gamma = 1$, the u-defining expression should be read as $u(x) = \ln(x)$ – by l'Hopital's rule.) To get the ball rolling, let us look at a two-state model for $S(1)$

$$S(1) = \begin{cases} uS(0) \text{ with probability } p, \\ dS(0) \text{ with probability } 1-p, \end{cases}$$

and as default, assume $W(0) = 100$, $S(0) = 100$, $u = 1.25$, $d = 0.85$, $r = 0.03$, $p = 0.5$ and $\gamma = 0.7$. We can formulate the problem in Excel and solve it using Solver. The default case is shown in Figure 4.1. For numerical convenience we have reformulated the problem such that we optimize over the fraction of time 0-wealth invested in the stock – and only over that. We see for this agent it is optimal to invest 74.38% of his wealth in the stock (and the rest in the risk-free). Interestingly for our purposes we see the state



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Portfolio choice example**												
2													
3	Time 0-wealth, W(0)	100			**Investments**			**Utility**					
4	Initial stock price, S(0)	100	u	1.25	Fraction in stock	0.7438		26.9466	<- E(u(C(1)))				
5	Interest rate	0.03	d	0.85	Fraction in bank	0.2562		u(x) = (x^(1-RRA)-1)/(1-RRA)					
6	E^P(rate of return on stock)	0.05						RRA ~ relative risk aversion				0.7	
7	SD(aktie-afkastrate)	0.15											
8	Risk-neutral up-prob'	0.45											
9													
10	Time 1; the future				Consumption, C(1) (aka. time 1-wealth W(1))								
11	State number	p prob'	S(1)		c_i	u(c_i)	p_i*u(c_i)	p_i*u'(c_i)	p_i*u'(c_I)/SUM_j(p_j*u'(c_j)				
12	1	0.5	125		119.3626	10.6606	14.4343	0.0176	0.45	<- Matches cell B8 @ optimum			
13	2	0.5	85		89.6124	9.5074	12.5123	0.0215	0.55				
14													

Figure 4.1: An optimal portfolio choice problem solved by Excel. The orange cells are input cells; the grey cells are intermediary calculations. File: https://tinyurl.com/n738w8dt

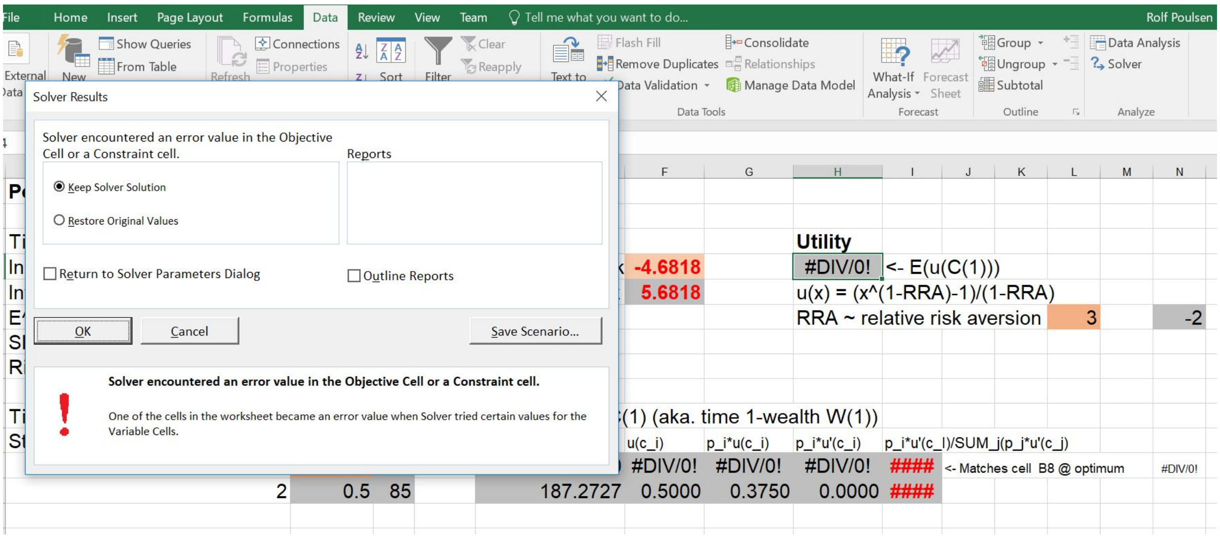
price utility theorem in action: the normalized, probability weighted marginal utilities at the optimum matches risk-neutral probabilities, compare cell I12 to cell B8. In fact, this gives us an alternative way to solve the portfolio choice problem: Put the difference between cell I12 and cell B8 in some cell and make Solver make that 0 by changing the fraction invested in the stock (cell F4). Playing around with the Excel file gives us further insight into the problem (remember that Solver has to be re-run manually when you change inputs):

- (Invariance) The optimal fraction invested in the stock is independent of (i) $S(0)$ (because of the multiplicative (u, d)-structure), (ii) $W(0)$ (because of the constant relative risk aversion).

- (Risk-aversion) The more risk-averse the agent is, the less he invests in the stock; for $\gamma = 0.9$ the optimal stock fraction is 58%, for $\gamma = 0.5$ it is 104% – the last result showing that we have not put in any short-sale or borrowing contraints. If we do (which is easy with Solver), the state-price utility theorem no longer applies (if constraints bind).

- (Good divergence) For $\gamma = 0$, Solver diverges; an unrestricted risk-neutral agent will invest all he can in the asset (here the stock) with the highest expected rate of return and nothing is stopping him here. For $r = 0.3$ and $r = -0.2$, Solver diverges, which Proposition 10 says should happen because this model has arbitrage.

- (Bad divergence) For $\gamma = 3$, the optimal stock fraction (found by the alternative



Solver Results

Solver encountered an error value in the Objective Cell or a Constraint cell.

Reports

● Keep Solver Solution

○ Restore Original Values

□ Return to Solver Parameters Dialog

□ Outline Reports

OK Cancel Save Scenario...

Solver encountered an error value in the Objective Cell or a Constraint cell.

One of the cells in the worksheet became an error value when Solver tried certain values for the Variable Cells.

-4.6818

5.6818

Utility

#DIV/0! <- E(u(C(1)))

u(x) = (x^(1-RRA)-1)/(1-RRA)

RRA ~ relative risk aversion 3 -2

(1) (aka. time 1-wealth W(1))

u(c_i) p_i*u(c_i) p_i*u'(c_i) p_i*u'(c_I)/SUM_j(p_j*u'(c_j))

#DIV/0! #DIV/0! #DIV/0! #### <- Matches cell B8 @ optimum #DIV/0!

2 0.5 85 187.2727 0.5000 0.3750 0.0000 ####

Figure 4.2: Solver diverging.

method described above) is 17%. However, when I try to make Solver find that by maximizing cell H4, it diverges even when I start it very close to the true optimum – see Figure 4.2. The reason is that in this case the criterion function is very, very flat around the optimal value; in 0.16, 0.17, or 0.18 in cell F4 and you will see no change in cell H4. This confuses the numerical optimization algorithm. Thinking carefully about the scaling of the problem and putting in constraints can help, but: Optimization can be very tricky. ■

Chapter 5

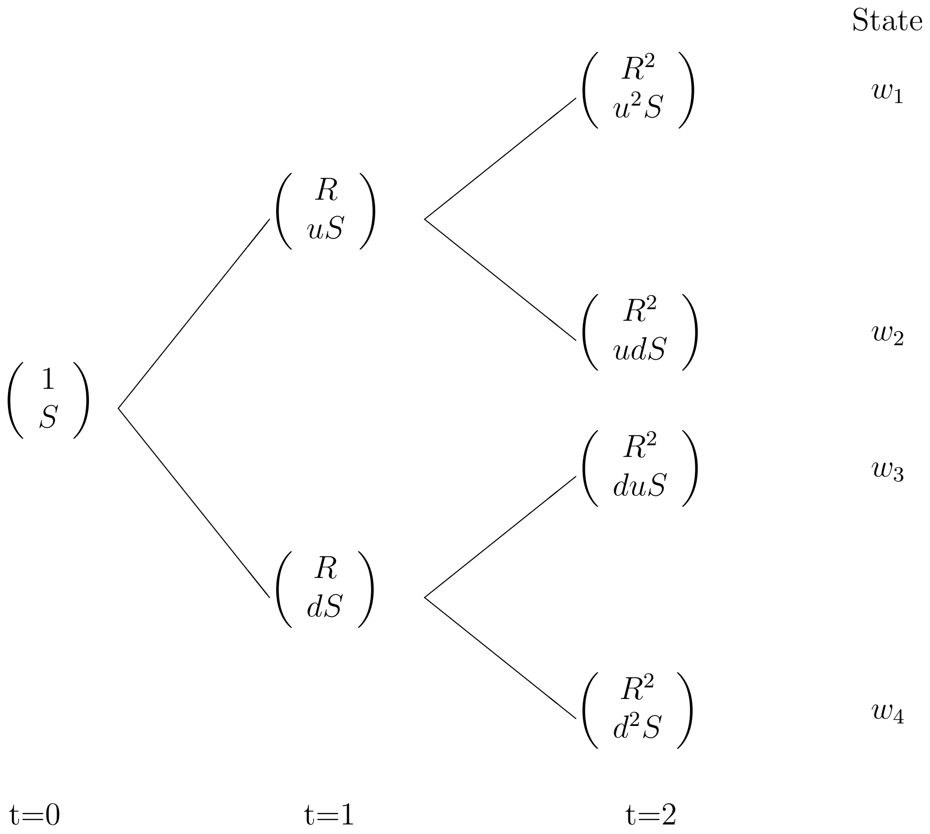
Stochastic multi-period models, dynamic trading, and the fundamental theorems of asset pricing

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 expiry 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 unique arbitrage-free 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.

Suppose there are three dates: $t \in \{0, 1, 2\}$ and that $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ – where we use the probabilistic tradition states of the world by ω. If you want, think "four future scenarios that we have numbered". We specify the behavior of the stock and the risk-free bank account as the following graph or tree:


State
$\begin{pmatrix} R^2 \\ u^2S \end{pmatrix}$
w_1
$\begin{pmatrix} R \\ uS \end{pmatrix}$
$\begin{pmatrix} R^2 \\ udS \end{pmatrix}$
w_2
$\begin{pmatrix} 1 \\ S \end{pmatrix}$
$\begin{pmatrix} R^2 \\ duS \end{pmatrix}$
w_3
$\begin{pmatrix} R \\ dS \end{pmatrix}$
$\begin{pmatrix} R^2 \\ d^2S \end{pmatrix}$
w_4
t=0
t=1
t=2

At time 0 the stock price is S, the bank 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 bank 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 expiry $T = 2$. 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 replication or hedging argument is the key to pricing derivative securities (for which the terms "contingent claims" or "options" are largely synonymous) 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 $\mathcal{F}$: $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}_T$.*

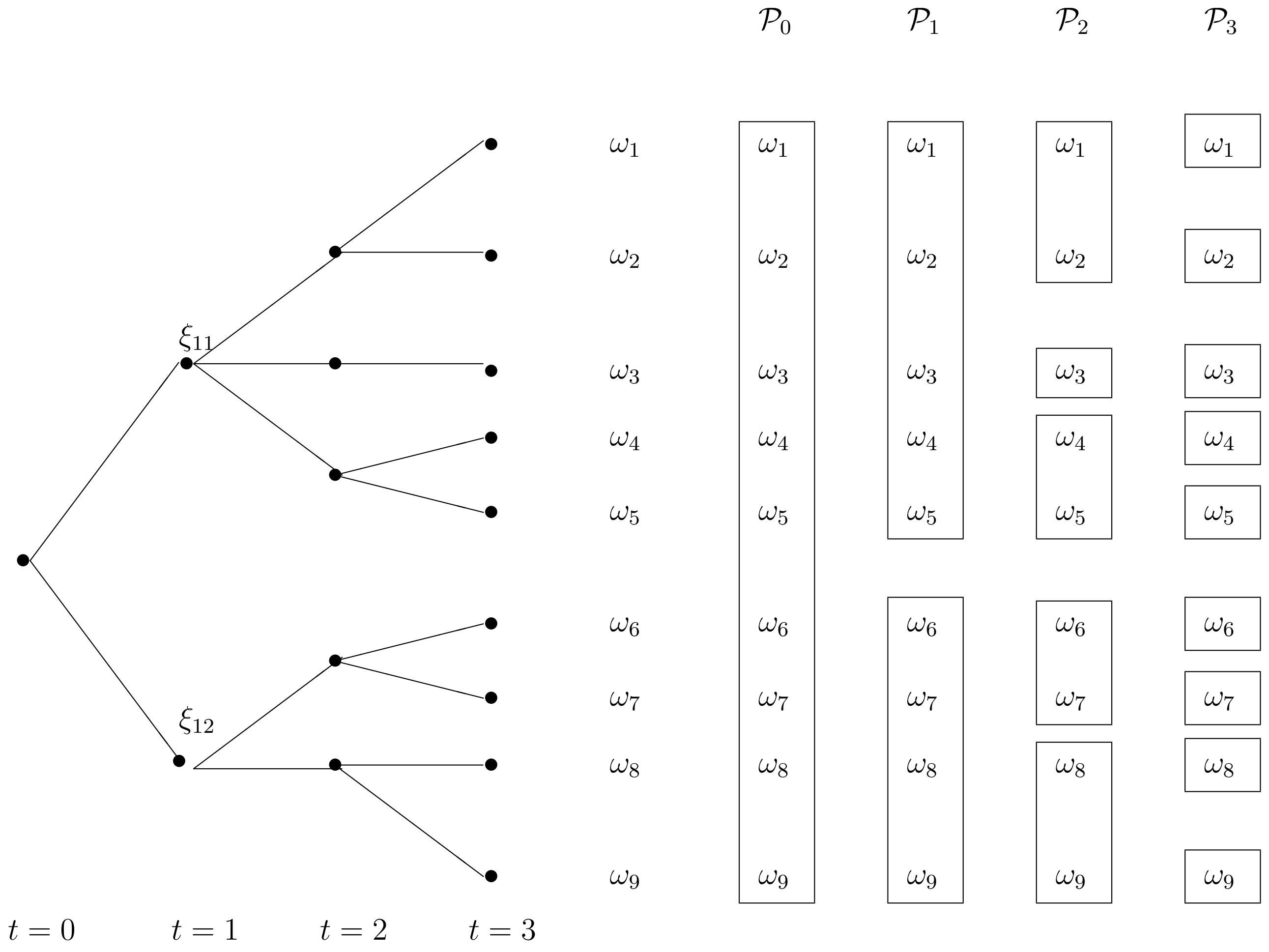
We will always assume that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_T = \mathcal{F}$. Since Ω is finite, it will be easy to think of the $\sigma-$algebras in terms of partitions:

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 3. *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 $\mathbb{F}$, 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 $\{\omega_1, \omega_2\}$.

Example 21. A classical exercise in *filtrations and measurability* runs along the following lines: Suppose we have a fair coin which we flip twice. The outcome ω of this experiment is the sequence $\omega = \omega_1\omega_2 \in \Omega$ where each individual coin flip ω_i $(i = 1, 2)$ can come out either heads (H) or tails (T). We also imagine that we monitor the evolution of the experiment: first at time $t = 1$ after the coin is flipped for the first time, and subsequently at time $t = 2$ when the coin has been flipped again. Question: what is the associated probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t=0,1,2}$ for this experiment? Clearly, the experiment has a totality of $2^2 = 4$ possible outcomes, which we represent by the sample space $\Omega = \{HH, HT, TH, TT\}$. The associated event space is the power set 2^Ω :

$$\begin{aligned}\mathcal{F} = \{&\emptyset, \{HH\}, \{HT\}, \{TH\}, \{TT\}, \{HH, HT\}, \{HH, TH\}, \{HH, TT\},\\ &\{HT, TH\}, \{HT, TT\}, \{TH, TT\}, \{HH, HT, TH\},\\ &\{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, TT\}, \Omega\}.\end{aligned} \tag{5.1}$$

You should check for yourself that $\mathcal{F}$ satisfies the σ-algebra properties of closure under complementation and countable unions. Notice that the cardinality of the filtration is $\#\mathcal{F} = 16$ or, equivalently, $|\mathcal{F}| = 2^{\#\Omega} = 2^4$ – which explains the notation for the power set. Finally, the real world probability measure $\mathbb{P}$ specifies the probability of every event in $\mathcal{F}$: e.g. since the coin is fair, we have for each of the elementary events $\omega \in \{HH, HT, TH, TT\}$ that $\mathbb{P}(\omega) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Other probabilities follow from the

axioms of probability: e.g. $\mathbb{P}(\Omega) = 1, \mathbb{P}(\emptyset) = 0, \mathbb{P}(\{HH \cup HT\}) = \mathbb{P}(\{HH) + \mathbb{P}(\{HT\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ and so forth.
As for the filtration $\mathbb{F} = \{\mathcal{F}_t\}_{t=0,1,2}$, the three σ-algebras $\mathcal{F}_0$, $\mathcal{F}_1$ and $\mathcal{F}_2$ effectively encode the information available to us at times $t = 0$, $t = 1$ and $t = 2$. Clearly, at $t = 0$, before any observation is made, we can only deduce the trivial events *something happened* or *nothing happened*, whence $\mathcal{F}_0 = \{\emptyset, \Omega\}$. At $t = 1$ the outcome of the first coin flip has been revealed (either $\omega_1 = H$ occurred, or $\omega_1 = T$ occurred) while ω_2 remains undisclosed. Hence, $\mathcal{F}_1 = \{\emptyset, \{HH \cup HT\}, \{TT \cup TH\}, \Omega\}$ where $\{HH \cup HT\}$ and $\{TT \cup TH\}$ are the *atoms* of $\mathcal{F}_1$. Finally, at $t = 2$ the outcome of the second coin flip has been revealed, thus resolving any ambiguity about the experiment. The σ-algebra of identifiable events is therefore $\mathcal{F}_2 = \mathcal{F}$, where $\mathcal{F}$ is given by (5.1).

Now suppose we (costlessly) enter a game which pays out \$1 every time the coin comes out heads, but deducts \$1 every time the coin comes out tails. Our cumulative gains are thus represented by the stochastic process $G_t : \Omega \times \{0, 1, 2\} \mapsto \mathbb{R}$ where $G_0 = 0$ and

$$G_1(\omega) = \begin{cases} +1, & \text{if } \omega \in \{HH, HT\} \\ -1, & \text{if } \omega \in \{TH, TT\} \end{cases} \qquad G_2(\omega) = \begin{cases} +2, & \text{if } \omega = HH \\ 0, & \text{if } \omega \in \{HT, TH\} \\ -2, & \text{if } \omega = TT \end{cases}$$

What is the smallest σ-algebras $\mathcal{F}_{G_1}$ and $\mathcal{F}_{G_2}$ generated by the random variables G_1 and G_2? I.e. what are the sets of possible outcomes that can be deduced solely by monitoring our cumulative gains and not the actual coin flips? With respect to which of the six σ-algebras $\mathcal{F}, \mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_{G_1}, \mathcal{F}_{G_2}$ are the random variables G_1 and G_2 measurable?

It is quite clear that $\mathcal{F}_{G_1} = \mathcal{F}_1$: there is no information difference between calling out $+1/-1$ or calling out H/T after the first coin flip. On the other hand, $\mathcal{F}_{G_2} \neq \mathcal{F}_2$: clearly, G_2 encodes *less* information since the outcome 0 tells us nothing about whether $\omega = HT$ or $\omega = TH$ occurred. The correct σ-algebra is readily shown to be

$$\begin{aligned} \mathcal{F}_{G_2} = \{&\emptyset, \{HH\}, \{TT\}, \{HT \cup TH\}, \{HH \cup TT\}, \{HH \cup HT \cup TH\}, \\ &\{TT \cup HT \cup TH\}, \Omega\}. \end{aligned}$$

To determine the measurability of a random variable X with respect to a given σ-algebra $\mathcal{F}$, we must check that the σ-algebra generated by X, $\mathcal{F}_X$, is a subset of $\mathcal{F}$. Since $\mathcal{F}_{G_1} \subseteq \mathcal{F}, \mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_{G_1}$ the random variable G_1 is measurable with respect to those filtrations. Analogously, you should check that G_2 is measurable with respect to $\mathcal{F}, \mathcal{F}_2$ and $\mathcal{F}_{G_2}$. ■

We now move to the multi-period modelling af financial markets using the concepts we have just introduced. We consider as exogenously given – but random – a vector of adapted dividend processes

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

and a vector of adapted security price processes

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

The interpretation is as follows: $S^i_t(\omega)$ 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 $\delta^i_{t+1}, \delta^i_{t+2}, \dots, \delta^i_T$.[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 bank account which provides locally risk-free borrowing and lending. This is modeled as follows: Given an adapted process - the short rate process

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

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. Now we may define the money market account as follows:

Definition 25. *The bank account has the security price process*

$$\begin{aligned} S^0_t &= 1, \qquad t = 0, 1, \dots, T-1 \\ S^0_T &= 0. \end{aligned}$$

and the dividend process

$$\begin{aligned} \delta^0_t(\omega) &= \rho_{t-1}(\omega) \textit{ for all } \omega \textit{ and } t = 1, \dots, T-1, \\ \delta^0_T(\omega) &= 1 + \rho_{T-1}(\omega). \end{aligned}$$

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 riskfree asset. You may of course also choose to keep the money in the bank 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*

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

with the 'glue on'-convention that we put $\phi_T = 0$. The value process of a trading strategy, V^δ, is $V^\delta(t) = \phi(t) \cdot S(t)$.

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*

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

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.

Definition 29. *(Replication) We say that we can replicate an adapted process X if there exists a trading strategy ϕ such that $\delta_t^\phi = X_t$ for all $t \geq 1$.*
(Completeness) The security model is called complete if every *adapted process X can be replicated.*

5.3 Conditional expectations and martingales

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 30. *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 $A_t \in \mathcal{P}_t$.

Note that we obtain an $\mathcal{F}_t-$measurable random variable since it is constant over elements of the partition $\mathcal{P}_t$. We can move $1/P(A_t)$ inside the sum, and since $A_v \subseteq A_t$ we have $P(A_v)/P(A_t) = P(A_v \cap A_t)/P(A_t) = P(A_v|A_t)$. This shows that our conditional expectation is expectation with conditional probabilities – and a lot of bookkeeping that we hide in the notation.

We must stress that the definition above does not work when the probability space becomes uncountable. Conditional expectation can still be defined in a more general setting. We will not need that here, so we will not do it, but just note that the more general definition is completely consistent with our definition.

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

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 — all of which can be derived by elementary methods from the definition :

$$\begin{aligned} E(E(X_u \,|\mathcal{F}_t)) &= EX_u && (5.2)\\ E(Z_t X_u \,|\mathcal{F}_t) &= Z_t E(X_u \,|\mathcal{F}_t) \text{ whenever } Z_t \in \mathcal{F}_t && (5.3)\\ E(E(X_u \,|\mathcal{F}_t) \,|\mathcal{F}_s) &= E(X_u \,|\mathcal{F}_s) \text{ whenever } s \leq t \leq u && (5.4) \end{aligned}$$

Equation (5.4) is called (the rule of) iterated expectations or the tower law. It is very useful. (But the so-called useful rule is something different!) Using Equation (5.3) is sometimes referred to as "taking out what is known". A consequence of (5.3) obtained by letting $X_u = 1$, is that

$$E(Z_t \,|\mathcal{F}_t) = Z_t \quad \text{whenever} \quad Z_t \in \mathcal{F}_t. \tag{5.5}$$

Another fact that we will often need: If a random variable Y is independent of the σ-algebra $\mathcal{F}$ (which means exactly what you think it means), then conditional expectation reduces to ordinary expectation, $E(Y|\mathcal{F}) = E(Y)$.

Example 22. The conditional expectation can be interpreted as "our best estimate given the available information". Or expressed mathematically: For any random variable X and any σ-algebra $\mathcal{F}$ we have that

$$E(X|\mathcal{F}) = \arg\min_{Z:\mathcal{F}-\text{measurable}} E((X-Z)^2).$$

Mathematicians would refer to this as a projection property. To prove it let us first note that for any $\mathcal{F}$-measurable random variable Y we have that

$$\begin{aligned} E(Y(X - E(X|\mathcal{F}))) &= E(E(Y(X - E(X|\mathcal{F}))|\mathcal{F})) \\ &= E(YE(X - E(X|\mathcal{F})|\mathcal{F})) \\ &= E(Y(E(X|\mathcal{F}) - E(X|\mathcal{F}))) = 0, \end{aligned}$$

where the first equality comes from iterated expectations and the second from taking out the known Z. Now let us write

$$\begin{aligned} E((X-Z)^2) &= E(((X - E(X|\mathcal{F})) + (E(X|\mathcal{F}) - Z))^2) \\ &= E((X - E(X|\mathcal{F}))^2) \\ &\quad +2E((X - E(X|\mathcal{F}))(E(X|\mathcal{F}) - Z)) \\ &\quad +E((E(X|\mathcal{F}) - Z)^2). \end{aligned}$$

The first term on the (last) right hand side does not depend on Z. By using $E(X|\mathcal{F}) - Z$, which is $\mathcal{F}$-measurable, in the role of Y above, we see that the second term is 0. The third term is positive, but by choosing $Z = E(X|\mathcal{F})$, which is allowed in the minimization problem, we can make it 0, which is as small as it can possibly get. Thus the desired result follows. ■

Now we can state the important definition:

Definition 31. *A stochastic process X is a martingale with respect to the filtration $\mathbb{F}$ if it satisfies*

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

Intuitively, a martingale is a stochastic process for which the expectation of tomorrow's value of the process is always equal to today's value. Martingales show up everywhere when modelling random dynamical systems. On the one hand the class of martingales is large, i.e. martingales can be quite different, but on the other hand not so large that all stochastic processes are martingales.

Example 23. (Three standard pieces of martingality.) The tower property of conditional expectation gives us the following:

- If X is an $\mathcal{F}_T$-measurable random variable then the stochastic process defined by $X_t = E(X\,|\mathcal{F}_t\,)$ is a martingale. Note: X could be "complicated". Martingales of this type are called Levy martingales. With our finite time-horizon, all martingales are Levy-type (put $X = M_T$). When working with an infinite time horizon, this is not true, as Example 24 effectively shows.

- If (and clearly only if) M is a martingale then it holds that

$$M_t = E(X_u\,|\mathcal{F}_t\,) \text{ for all } 0 \leq t \leq u \leq T. \tag{5.6}$$

  So the martingale property holds not only for adjacent time-points, but "globally". We also see if M is a martingale, then $E(M_t)$ is constant (i.e. independent of t). (The converse is not true.) In continuous time, where we can't talk rigorously about adjacent time-points, equation (5.6) is used to define martingales.

- Let $\{X_t\}_{t=1,2,\ldots T}$ be a sequence of independent, identically distributed random variables with common mean μ. Put $Y(0) = 0$ and $Y(t) = \sum_{i=1}^t X_i$ for $1 \leq t \leq T$. Then Y is a martingale (wrt. both the natural filtration of the X's and the natutral filtration of the Y's) if and only if $\mu = 0$. (Note: Unless the X's are all the same deterministic constant, then the process $\{X_t\}_{t=1,2,\ldots T}$ is not a martingale) This shows that martingales are intimately connected to the cumulative winnings in fair games.

■

Example 24. (A martingale-like process with a strange property) Let $\{X_t\}_{t=1,2,3,\ldots}$ be an *infinite* sequence of independent, identically distributed random variables where each X_t takes the values $\pm \ln 2$ with probability $\frac{1}{2}$. All conditional expectations are wrt. the X's natural filtration. Put $M(0) = 1$ and

$$M(t) = \prod_{i=1}^t e^{X_i + c} \text{ for } t \geq 1,$$

where c is some constant. We see that if c has the property that

$$E(\exp(X_{t+1} + c)) = 1 \tag{5.7}$$

then it holds that

$$E_t(M(t+1)) = M(t) \text{ for alle } t \geq 0, \tag{5.8}$$

which would be the definition of something being a martingale were we to work with an infinite time-horizon. Plugging in, we see that equation (5.7) is solved by $c = \ln(4/5) = -0.2231$. Note that we can rewrite

$$M(t) = \exp\left(tc + \sum_{i=1}^t X_i\right) = \exp\left(t\left[c + \frac{\sum_{i=1}^t X_i}{t}\right]\right).$$

By the law of large numbers, the second term inside the square brackets goes to 0 almost surely for $t \to 0$. Because the first term is negative we get that $M(t) \to 0$ almost surely for $t \to \infty$ — even though $E(M(t)) = 1$ for all t. This is an example of a martingale that is not uniformly integrable – and it is the stuff of which paradoxes are made. ■

5.4 The fundamental theorems of asset pricing

In this section we state and prove what is known as the fundamental theorems of asset pricing. These are an indispensable tool for constructing arbitrage-free models and pricing contingent claims in these models.
We maintain the setup with a filtered probability space $(\Omega, \{\mathcal{F}\}_{t=0,1,\ldots,T}, \mathcal{F}, P)$ on which we have N securities with price (S) and dividend processes (δ) as well as a bank account or locally risk-free asset generated by the short rate process ρ; (S, δ, ρ) for short. We define the corresponding discounted processes $\widetilde{S}, \widetilde{\delta}$ by defining for each $i = 1, \ldots, N$

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

We shall need to refer to (all) one-period sub-models. These are defined or constructed in the following way: Let $\mathcal{P}_u$ denote the partition of $\mathcal{F}_u$ for all u, look at an $A_t \in \mathcal{P}_t$ and put

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

This number is called the splitting index at A_t. In our graphical representations 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 define a one-period submodel by first letting

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

and denoting by $B_1, \ldots, B_{N(A_t)}$ the members of $\mathcal{F}_{t+1}$ that are subsets of A_t . We put

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

and the $(\pi(t, A_t), D(t, A_t))$'s are what we call the one-period sub-models.

Two probability measures are said to be equivalent when they assign zero probability to the same sets; we denote such equivalence by $\sim$.

Definition 32. *A probability measure $Q \sim P$ is an equivalent martingale measure, we have*

$$\widetilde{S}_t^i = E_t^Q \left(\sum_{j=t+1}^{T} \widetilde{\delta}_j^i \right) \qquad t = 0, \ldots, T-1 \textit{ and } i = 1, \ldots, N. \tag{5.9}$$

We can rewrite the definition of an equavalent martingale measure into the follow local characterization:

Theorem 4. *A measure Q is an equivalent martingale measure if and only if the following holds*

$$S_t^i = \frac{1}{1+\rho_t} E_t^Q \left(S_{t+1}^i + \delta_{t+1}^i \right) \textit{ for all } i \textit{ and } t \textit{ (and } \omega).$$

Proof. (Short form. The reader is encouraged to "cross the dot the i's and cross the t's" him- or herself.) Rewrite Equation (5.9) as

$$\frac{S_t^i}{R_{0,t}} = E_t^Q \left(\frac{\delta_{t+1}^i}{R_{0,t+1}} + \sum_{j=t+2}^{T} \widetilde{\delta}_j^i \right).$$

Now use linearity of conditional expectation, use iterated expectations to write "$E_t^Q = E_t^Q E_{t+1}^Q$", use the definition of R (a couple of times), and finally use Equation (5.9) for $t+1$. ■

The local charaterization theorem explains why we use the phrase "martingale measure": Discounted prices of non-dividend-paying sequrities are Q-martingales. More generally, the local characterization and the properties of conditional expectation gives us the following "martingality" result.

Theorem 5. *The discounted value process of any self-financing trading strategy is a Q-martingale.*

We are now ready to formulate and prove our most general version of the first fundamental theorem of asset pricing:

Theorem 6. *For the model (S, δ, ρ) the following statements are equivalent:*

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

Proof. ("$\exists\, Q \Rightarrow$ no arbitrage") Assume an equivalent martingale measure Q exists. Suppose – contrarily – that ϕ is an arbitrage. Because the locally risk-free asset has a strictly positive price process, we may without loss of generality take ϕ to be self-financing on the relevant time-interval. (Any intermediate cash-flows we just invest in the locally risk-free asset.) The arbitrage property implies that $V^\phi(0) = V^\phi(0)/R_{0,0} \leq 0$ (the dividend-stream, which is non-negative, starts out as minus the value process) while (by

the self-financing property) we have $E^Q(V^\phi(T)) > 0$ – strict inequality being guaranteed by the equivalence of P and Q – and thus $E^Q(V^\phi(T)/R_{0,T}) > 0$. Since V^ϕ/R is a Q-martingale, this is a contradiction.
("No arbitrage $\Rightarrow \exists\, Q$") Assume the model is arbitrage-free. Then all 1-period sub-models ("$D = S(t+1) + \delta(t+1)$") must be arbitrage-free. Thus – by our knowledge of one-period models, Proposition 7 and analysis thereafter – the local characterization holds in all one-period sub-models, and so a martingale measure Q exists ($>>$ 0-positivity of stare price vectors ensuring equivalence) ■

Working from the local characterization we also get the second fundamental theorem of asset pricing:

Theorem 7. *Assume the model (S, δ, ρ) is arbitrage-free. Then the market is complete if and only if the equivalent martingale measure is unique.*

Other two immediate consequences are:

Corollary 1. *Assume the security model defined by (S, δ, ρ) is 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*

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

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

Corollary 2. *The multi-period model (S, δ, ρ) is (a) arbitrage-free if and only if all one-period sub-models are arbitrage-free, (b) complete if and only if all one-period sub-models are complete.*

Corollary 2 tell us that we can check absence of arbitrage and completeness by looking at one-period sub-models instead of the whole tree. This is useful because we often build multi-period models by repeating the same one-period structure. Notice also: The one-period sub-models do not have to be in any way similar, e.g. "uniqueness" does not mean that "q is the same in all sub-models".

5.5 The standard binomial model

We now take a closer look at a frequently used workhorse: The standard binomial model. To this end suppose first that the relevant time interval $[0;T]$ has been spilt into n pieces of length $\Delta t = T/n$ and write $t_i = i\Delta$. All our previous analysis carries over when we change time-points from $\{0, 1, 2 \dots T\}$ to $\{0, \Delta, 2\Delta t, \dots n\Delta t\} = \{t_i\}_{i=0}^n$.

Definition 33. *A multiplicatively homogeneous model has the form*

$$S(t_{i+1}) = S(t_i) \times \begin{cases} u & \textit{with probability } p \\ d & \textit{with probability } 1-p \end{cases}$$

where

$$\begin{aligned} u &= \exp(\alpha\Delta t + \sigma\sqrt{\Delta t}) \\ d &= \exp(\alpha\Delta t - \sigma\sqrt{\Delta t}), \end{aligned}$$

and α and σ constants. The case where $p = \frac{1}{2}$ we call the standard binomial model.

There is not complete consistency in the literature about what consitutes the standard binomial model. For instance, some sources allow $p \neq \frac{1}{2}$ while others force α to be 0. Such choices are usually made for good reasons, some of which we will see at the end of this section. An important part of the definition is the explicit assumption about how up- and down-moves scale with time-steps. It is not a priori clear why we want a Δt, $\sqrt{\Delta t}$, and nothing else in there. Our subsequent analysis will illuminate and justify this structure. Notice also that the model structure means that relative price changes $(S(t_{i+1})/S(t_i))$ are independent (under both P and Q) and identically distributed (*iid* for short). (But prices themselves or "raw" price changes are neither.) Figure 5.1 shows the standard binomial model for a particular choice of α and σ. The price development can be described by a lattice, a recombining tree; "up, down" leads to the same stock price as "down, up", but note that the lattice is (exponentially) curved.

Moments and empirics. To further investigate the model's stucture let us look at conditional first and second moments. By using the form of u and d and Taylor expanding the exponential function to the second order around 0 ($\exp(\pm x) \approx 1 \pm x + x^2/2$ — with the role of x played by $\alpha\Delta t + \sigma\sqrt{\Delta t}$ and $\alpha\Delta t - \sigma\sqrt{\Delta t}$) we get that

$$\mathrm{E}_t^P\left(\frac{S_{t+\Delta t} - S_t}{S_t}\right) = \mu\Delta t + \mathrm{o}(\Delta t),$$

where $\mu = \alpha + \sigma^2/2$ and the order-notation "$\mathrm{o}(\Delta t)$" means that the remainder goes to 0 faster than Δt when $\Delta t \to 0$. Taylor expanding to the second order ($\exp(\pm x) = 1 \pm x + \frac{1}{2}x^2$) gives us the conditional variance of the rates of return,

$$\mathrm{var}_t^P\left(\frac{S_{t+\Delta t} - S_t}{S_t}\right) = \sigma^2\Delta t + \mathrm{o}(\Delta t).$$

We can use this along with the independence of the rates of return to estimate σ and α (and/or μ if we like) by basic statistics; find sample moments and scale them appropriately by Δt to get parameters. (It is common to measure time in years and say that $\Delta t = 1/252$ corresponds to daily observations.) But we can be more cunning. To this end let us look at logarithmic rates of return defined

$$a_{t,t+h} = \ln(S(t+h)/S(t)),$$

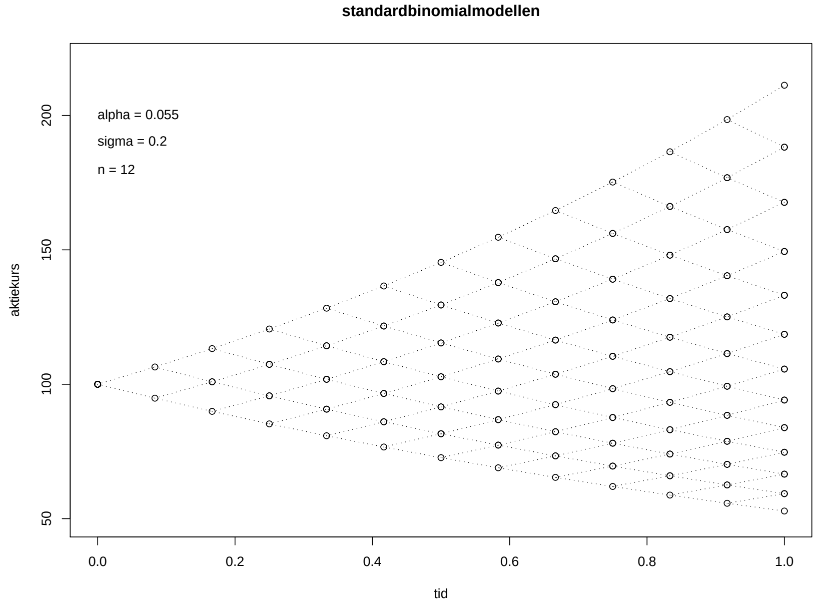


Figure 5.1: A standard binomial model lattice

Horizon (h)	average of a_h's	$\hat{\alpha}$	std. dev. of a_h's	$\hat{\sigma}$
1 day	0.000399	0.101 [0.042]	0.0123	0.195 [0.0019]
5 days (∼ 1 week)	0.00181	0.101 [0.043]	0.0280	0.198 [0.0042]
21 days (∼ 1 month)	0.00734	0.100 [0.043]	0.0578	0.202 [0.0089]

Table 5.1: Estimation of parameters of the standard binomial model on Danish data 1994-2014. Numbers in square brackets are standard errors.

where h is called the horizon. The logarithmic rates of return are *iid* and additive in the horizon,

$$a_{t,t+n*\Delta t} = \sum_{i=1}^{n} a_{t+(i-1)*\Delta t,t+i*\Delta t}. \tag{5.10}$$

This enables us to estimate the model's parameters (which horizon independent) from different observation frequencies corresponding to different choices of horizon and thereby justify the suitablility of its structure. Doing this for the leading Danish stock index over the period 1994-2014 (shown in Figure 5.2). , estimation results are shown in Table 5.1. Notice that the "raw" moments depend strongly on the horizon, the parameters much less so. Because of the telescoping nature of the sum in equation (5.10), the α-estimates do not depend on the horizon (the slight difference for $h = 21$ is becuase the number of observations is not a multiple of 21). This also means that the accuracy of the α-

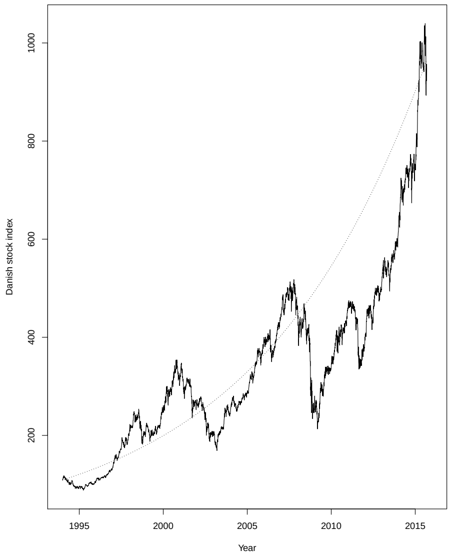


Figure 5.2: The leading Danish stock index 1995-2014

estimator is not increased by sampling more frequently. On the other hand, the fact that σ-estimates are stable acoss horizons indicates that the model gets the time-scaling about right and that the rates of return are close to uncorrelated; the standard deviations of rates of return grow like the square-root of the horizon. When comparing standard errors of the estimators, we see that the mean-parameter α is much less accurately estimated than the volatility parameter σ; in fact the error on the former decreases to zero as we increase the observation frequency. There is a reason for this – that has to do with the quadratic variation of Brownian motion! (The reason the $h = 1$ and $h = 5$ α-estimators have slightly different standard errors is that these involve the samle standard deviation of the rates of return.)

Option pricing. The analysis from earlier of one-period binomial models still holds, in particular the usual formula $q = (R - d)/(u - d)$ applies with $R = \exp(r\sqrt{\Delta t})$. So the core of an implementation of call-option pricing is this (here: R-code)

```
call[,n+1]<-pmax(S0*u^(0:n)*d*^(n:0)-strike,0)
for (i in n:1) {
  for (j in 1:i){
    call[j,i]<-(q*call[j+1,i+1]+(1-q)*call[j,i+1])/R
  }
}
```

From the fact that the discounted stock price is a Q-martingale we have

$$\mathrm{E}_t^Q\left(\frac{S_{t+\Delta t}-S_t}{S_t}\right)=\exp(r\sqrt{\Delta})-1=r\Delta t+\mathrm{o}(\Delta t).$$

By second order Taylor expansions on u, d, and q (conceptually straightforward but computationally tedious) we find that

$$\mathrm{var}_t^Q\left(\frac{S_{t+\Delta t}-S_t}{S_t}\right)=\sigma^2\Delta t+\mathrm{o}(\Delta t).$$

This means that the conditional variances of the rates of return are unaffected by the change from P to Q, at least up to terms that a quite small when time-steps are small. (This is a discrete version of a result known as Girsanov's theorem.) This – along with the graphical anlysis in Figure 5.3 – leads us to the following conjectures:

- Call option prices converge to some limit when $n \to \infty$.

- This limit depends neither on p nor on α. (That is the reason some sources have a laissez faire approach to their values, setting $p = 1/2$ and $\alpha = 0$.)

Both conjectures are true. The limit is the so-called and celebrated Black-Scholes *formula*, which is supported by the continuous-time Black-Scholes *model.* We shall return to these matters in Chapter 7 .

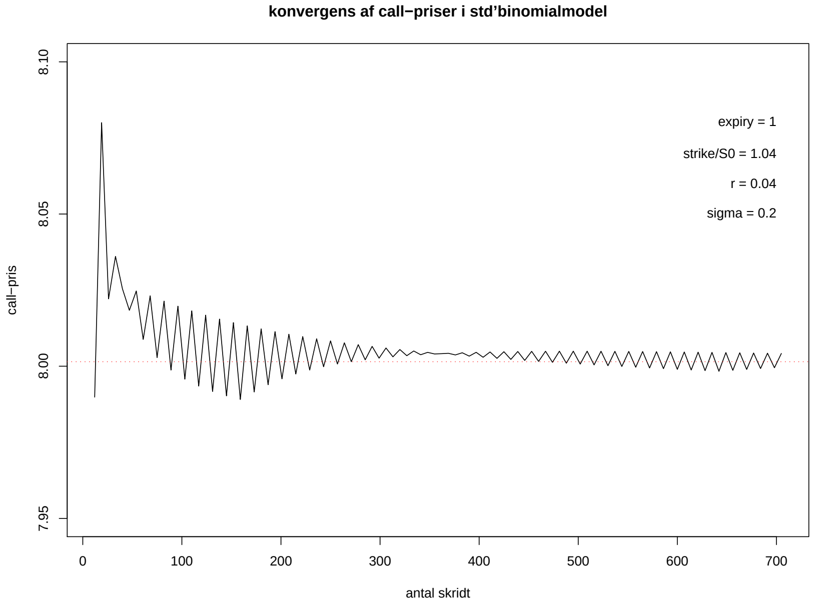


Figure 5.3: Call option prices in the standard binomial model as a function of the number of steps n keeping expiry fixed; $\Delta t = T/n$.

Chapter 6

Other options: Pricing and hedging financial derivatives

The financial contracts of which we speak in this chapter go by many names; derivatives, contingent claims, options, or structured products. Unless we are trying to make a very specific point, we will use these terms indiscriminately. Market participants may be even more cryptic, giving specific strategies funny names (straddle, butterfly, ...) and using phrases such as "I'm long volatility" or "I've shorted Gamma". We shall refrain from such formulations – although again: unless we are trying to make a very specific point.

The material in this chapter is classical applications of the arbitrage pricing machinery we have developed in the previous chapters. The models, results, and techniques we present are applied with minor modifications every day all over the world as the basis for trading billions of dollars worth of contracts. For a student intent on becoming a financial quant or trader, that should give plenty of motivation for learning these models. But the recent (or perhaps more accurately: recurrent) turmoil in financial markets is a reminder that financial managers and executives must also 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.

Throughout the chapter we take as given an arbitrage-free (price, dividend, interest rate) $= (S, \delta, \rho)$ model on a filtered probability space; π is used generically for derivative prices and Q denotes a martingale measure. We do not assume completeness (i.e. uniqueness of Q) unless explicitly stated. The same goes for dividends being 0 and the interest rate being constant and/or positive.

6.1 Forward and futures contracts

A forward contract is an agreement to buy the underlying security at the expiry date T of the contract at a price of $\mathrm{Fwd}(t,T)$. Note that $\mathrm{Fwd}(t,T)$ is specified at time t and that unlike an option, the forward contract forces the holder to buy. In other words you can lose money at expiry on a forward contract. The forward price is decided so that the value of the forward contract at the initiation 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. So which value of $\mathrm{Fwd}(t,T)$ gives the contract a value of 0 at initiation? The following propositions tell us.

Proposition 11. *(Basic forward prices.) Consider an expiry-T forward contract on a traded asset that does not pay dividends over the life of the forward contract,* $]t;T]$. *The unique arbitrage-free forward price is*

$$\mathrm{Fwd}(t,T) = \frac{S_t}{P(t,T)},$$

where $P(t,T)$ *denotes the time-t price of a maturity-T zero coupon bond.*

Proof, I: A simple portfolio argument, this is often called a carry argument. Consider the following strategy and its cash-flows:

strategy / cash-flow	date t	date T
buy 1 stock	$-S_t$	S_T
sell $\frac{S_t}{P(t,T)}$ maturity-T zero coupon bonds	S_t	$-\frac{S_t}{P(t,T)}$
sell 1 expiry-T forward	0 (by definition)	$\mathrm{Fwd}(t,T) - S_T$
total cash-flow	0	$\mathrm{Fwd}(t,T) - \frac{S_t}{P(t,T)}$

The strategy's cash-flow at time T is known at time t and its cash-flow at time t is 0. Thus the former must be 0 as well. Hence to avoid arbitrage we must have $\mathrm{Fwd}(t,T) = S_t/P(t,T)$ ■

Proof, II: An abstract proof. From general no arbitrage pricing we have

$$0 = \mathrm{E}_t^Q\left(\frac{S_T - \mathrm{Fwd}(t,T)}{R_{t,T}}\right).$$

Using linearity of conditional expectation and moving the t-measurable quantity $\mathrm{Fwd}(t,T)$ outside gives us

$$\mathrm{Fwd}(t,T)\mathrm{E}_t^Q\left(\frac{1}{R_{t,T}}\right) = \mathrm{E}_t^Q\left(\frac{S_T}{R_{t,T}}\right).$$

The second factor on the left-hand side is by definition the zero coupon bond price. The 0-dividend assumption, the local characterization of Q, and repeated application of the tower property shows that the right-hand side is simply S_t. The desired formula follows ■

With non-zero dividends, things get more complicated.

Proposition 12. *For an expiry-T forward contract on a traded asset that pays deterministic dividends, $\delta(t_j)$ at time t_j, the unique time-t arbitrage-free forward price is*

$$\mathrm{Fwd}(t,T) = \frac{S(t) - \sum_{j|t_j \in]t;T]} P(t,t_j)\delta(t_j)}{P(t,T)}.$$

Proof. Extend the carry strategy from the proof of Proposition 11. At time t and for each $t_j \in]t;T]$ we sell $\delta(t_j)$ units of maturity t_j-zero coupon bonds. For a specific t_j, this gives income of $P(t,t_j)\delta(t_j)$. We use this money to buy maturity T-zero coupon bonds, thus making the net cash-flow at time t zero; each component giving us $P(t,t_j)\delta(t_j)/P(t,T)$ units.. Intermediary cash-flows net out, and compared to the basic carry strategy, we get an extra cash-flow of $\frac{\sum_{j|t_j \in]t;T]} \delta(t_j)P(t,t_j)}{P(t,T)}$ at time T – which is known at time t. So just as before, to avoid arbitrage, the total time T-cashflow must be zero, and the desired formula for $\mathrm{Fwd}(t,T)$ follows ■

Proposition 12 can handle forward contracts on coupon-bearing bonds. While the assumption of deterministic dividends is crucial for the exactness of the result, the proposition is also useful for approximating forward prices; we may switch dividends by their expected values, or in empirical applications by their ex-post realized values.

In some settings a more reasonable assumption is that dividends are proportional to the price of the underlying, $\delta(t_j) = \delta S(T_j)$. Explicit, exact and unique expressions for forward prices can also be derived in this case without assuming market completeness. Particularly relevant are exchange rate forward contracts:

Proposition 13. *Let X denote the exchange rate between two currencies, economies; i.e. $X(t)$ is the number of units of domestic currency needed at time t to buy 1 unit of foreign currency. The time-t forward price of an expiry-T forward contract on the exchange rate is*

$$\mathrm{Fwd}^{FX}(t,T) = e^{(r_d - r_f)(T-t)} X(t),$$

where r_d and r_f are, respectively, the domestic and the foreign interest rate. For a stock that pays a constant continuous dividend yield δ, the formula above holds with δ playing the role of r_f.

Proof. Adjust the carry argument but buying a suitable number of units of currency and depositing it in a foreign bank; details left as an exercise to the reader ■

The explicit results in Propositions 11-13 do not assume market completeness, but they do assume specific dividend structures. The best general result we can state is this:

Proposition 14. *The arbitrage-free forward price is*

$$\mathrm{Fwd}(t,T) = \frac{\mathrm{E}_t^Q\left(S(T)/R_{t,T}\right)}{P(t,T)}.$$

Proof. Similar to the first part of the abstract proof of Proposition 11.

For an incomplete model with a general dividend structure, there is *not* a unique arbitrage-free forward price. However, for a specific model specification an arbitrage-free interval of forward prices (typically quite tight) can be found using the method described in Section 2.7.2. Other complications arise if the forward contract is related to an underlying that is not a traded or easily storable asset, examples of the latter being energy and agricultural commodities.

There is one further twist to the story. If you consult the homepages of major stock exchanges you will find quotes of stocks, bonds, interest rates, and a plethora options. But not forward prices. You will, though, find quotes of futures (yes, there is an s at the end) prices. These are almost, but not quite, the same. Let us explain – in a way intended to separate the *what* (are futures contracts and the somewhat mysterious futures prices) from the *why* (are these the traded objects).

Like a forward contract, a futures contract can be entered into free of cost. Doing that at time t and holding the futures contract until its expiry T generates these cash-flows:

	t	$t+1$	$t+2$	$\cdots$	$T-1$	T
Futures	0	$\Phi_{t+1}-\Phi_t$	$\Phi_{t+2}-\Phi_{t+1}$	$\cdots$	$\Phi_{T-1}-\Phi_{T-2}$	$S_T-\Phi_{T-1}$

Here, the adapted process Φ is the so-called futures price (which technically is not really a price, but a cumulative dividend process) and the $\Phi_{t+1}-\Phi_t$-terms are called margin payments. This is a somewhat convoluted way of defining the contract and its cash-flows; what is the futures price, what must it be? To analyze this we start working our way backwards from $T-1$. First, Φ_{T-1} is such that the cash-flow promised when the futures contract is entered into (alternative words: bought, "go long") at $T-1$ has value 0, i.e.

$$0 = \mathrm{E}_{T-1}^Q\left[\frac{S_T-\Phi_{T-1}}{R_{T-1,T}}\right],$$

but since $R_{T-1,T}$ is $\mathcal{F}_{T-1}$-measurable this implies

$$0 = \frac{1}{R_{T-1,T}}\mathrm{E}_{T-1}^Q\left[S_T-\Phi_{T-1}\right] ,$$

i.e.

$$\Phi_{T-1} = \mathrm{E}_{T-1}^Q\left[S_T\right].$$

Now for Φ_{T-2}. By definition Φ_{T-2} should be set such that the cash-flow of the futures contract entered into at $T-2$ has value 0,

$$0 = \mathrm{E}^Q_{T-2}\left[\frac{\Phi_{T-1}-\Phi_{T-2}}{R_{T-2,T-1}} + \frac{S_T-\Phi_{T-1}}{R_{T-2,T}}\right] \tag{6.1}$$

Using the tower property of conditional expectations and the expression for Φ_{T-1}, we find

$$\mathrm{E}^Q_{T-2}\left[\frac{S_T-\Phi_{T-1}}{R_{T-2,T}}\right] = \frac{1}{R_{T-2,T-1}} E^Q_{T-2}\left[E^Q_{T-1}\left[\frac{S_T-\Phi_{T-1}}{R_{T-1,T}}\right]\right] = 0,$$

so (6.1) holds precisely when

$$0 = E^Q_{T-2}\left[\frac{\Phi_{T-1}-\Phi_{T-2}}{R_{T-2,T-1}}\right] = \frac{1}{R_{T-2,T-1}} E^Q_{T-2}\left[\Phi_{T-1}-\Phi_{T-2}\right],$$

i.e. we have

$$\Phi_{T-2} = E^Q_{T-2}\left[\Phi_{T-1}\right] = E^Q_{T-2}\left[S_T\right].$$

Continuing this argumentation backwards we get:

Proposition 15. *The arbitrage-free time-t futures price for a futures contract with expiry T is*

$$\Phi_t = \mathrm{E}^Q_t\left[S_T\right].$$

In general, forward and futures prices are not equal. However, by combining Propositions 14 and 15 we see that if the interest rate is deterministic, then the forward and the futures price do coincide. If interest rates are stochastic, forward and futures prices will rarely be the same (note that it is quite difficult for $\frac{1}{R_{t,T}}$ and S_T to be uncorrelated under Q, because the interest rate is itself the conditional Q-expected one-period rate of return on the underlying). However, the practical differences only manifest themselves for long times to expiry and/or peculiar underlying assets, so in applications it is fairly safe to substitute futures price for forward price or vice versa.

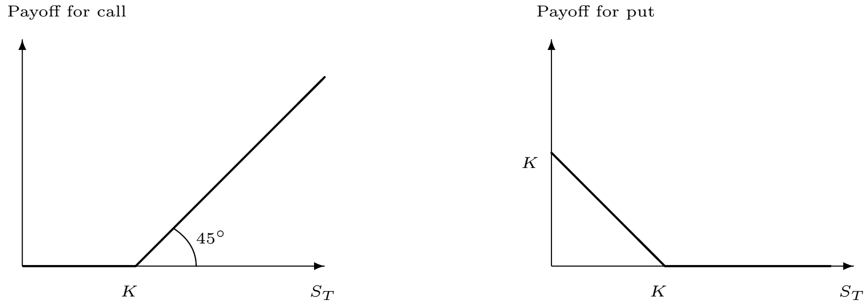
Note that even with a deterministic interest rate – where forward and futures *prices* are equal – the forward and the futures *contracts* are not the same; they have different cash-flows. And strictly speaking the futures contract's payments are path-dependent; the time $t+1$ payment depends on what the stock price was at time t. In binomial model terms, the cash-flow is different depending on whether we got to the current stock price node via an up-move or a down-move.

To understand why futures, not forwards, are traded in practice, note that even though a forward contract has value 0 at initiation, its value may subsequently become (quite) negative or positive depending on the movement of the underlying. One party in the trade will thus have a substantial liability to the other. In the futures contract construction, the running margin payments ensure that the contract's value is always kept at zero; gains and losses are settled along the way. From a credit risk point of view, it is easy to see the benefits of this pay-as-you-go system. (Also futures contracts have a bookkeeping advantage; to settle a futures contract we only have to look one day back, to settle forward contracts at the expiry date, we have to keep track of all the different forward prices at which the contracts were entered into at different times.)

6.2 Option basics: Diagrams, strategies and put-call parity

A call option on some underlying asset (which we generically refer to as the stock) gives the right, but not the obligation to buy the underlying asset at a certain date[1] (T; the expiry) for a certain price (K; the strike or exercise price). Conversely, a put option gives the right but not the obligation to sell. Calls and puts – often referred to as *plain vanilla* – are the basic building blocks of options markets.

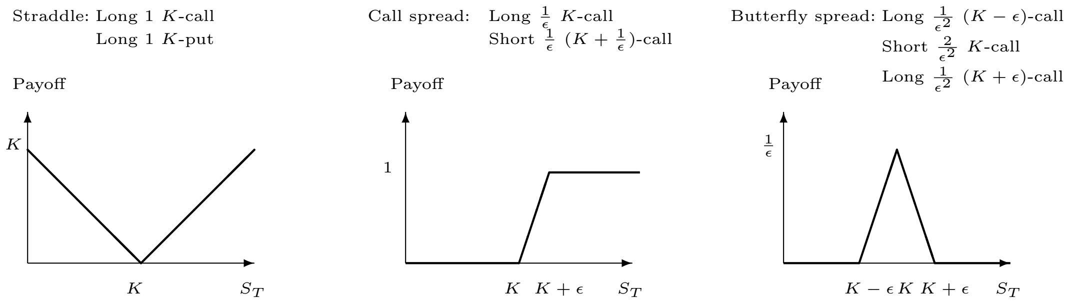
We think of options in terms of their payoffs;[2] for calls and puts in particular their payoff *functions*; $f_{call}(x) = (x - K)^+$, $f_{put}(x) = (K - x)^+$. We can depict these in payoff diagrams, which show the payoff, i.e. the time T-value, of the option as a function of the value of the underlying. For the basic call and put they look like this:



When combining calls and puts (and bonds and the underlying), payoff diagrams become very useful. Such strategies or spreads are given different, often colorful, names. For instance the diagrams below show (and define) the straddle, the call spread and the butterfly spread:

[1] Note that exercise can take place only at the expiry date T, not prior to expiry. We will later look at options where the holder also chooses when to exercise the option. The latter are called options of American type, the former European.

[2] Strictly speaking, to us pay-off functions are what define the option, not the “right, not obligation” formulation. In doing so, we implicitly but prudently assume that our counterpart is not a complete idiot, e.g. he does not insist on buying the stock from us when it is cheaper in the market.



Better than words or formulas, the payoff diagrams tell us what different strategies do; what we are betting on. The straddle pays off positively if the stock moves (think of $K \approx S_t$), irrespective of the direction; the call spread is a 0/1-bet on the stock price going up (think of $K \approx S_t$ and a small ϵ); the butterfly allows us to generate a positive payoff in a target range of future stock prices, where the width of the range is controlled by ϵ (the seemingly strange $1/\epsilon^2$-scaling comes in handy in a little while).

Sometimes, payoff diagrams for strategies may be shown in a break-even fashion, i.e. by adjusting the payoff for the initial cost of the strategy (taken forward to time T by dividing by $P(t,T)$).

To value an individual call or put at some date prior to expiry, we need a dynamic, stochastic model. However, there are model-free restrictions across different options, the most important of which is the put-call parity.

Proposition 16. *(Base-case put-call parity) Let $call_t$ and put_t denote prices of (European) expiry-T, strike-K calls and the prices on a certain stock. If the stock does not pay dividends during the life of the options, then we have*

$$call_t - put_t = S_t - KP(t,T).$$

Proof, I A simple portfolio argument. At time t, buy 1 call and K T-zero coupon bonds, sell 1 put and 1 unit of the stock. The strategy has no intermediary cash-flows, and its time-T payoff we can find thus: (1) If $S_T \geq K$, then is $S_T - K$ (from call) + K (from bonds) - S_T (to stock) - 0 (to put) – or in other words 0. (2) If $S_T \geq K$, then is 0 (from call) + K (from bonds) - S_T (to stock) - $(K - S_T)$ (to put) – also 0. So to avoid arbitrage, the time-t cost of the strategy most also be 0. From this the put-call parity follows. ■

Proof, II Abstractly. Note that $x - K = (x-K)^+ - (K-x)^+$. Using S_T in the role of x, dividing by $R_{t,T}$ and taking E_t^Q gives us (by linearity of conditional expectation)

$$\mathrm{E}_t^Q\left(\frac{S_T - K}{R_{t,T}}\right) = \mathrm{E}_t^Q\left(\frac{(S_T-K)^+}{R_{t,T}}\right) - \mathrm{E}_t^Q\left(\frac{(K-S_T)^+}{R_{t,T}}\right).$$

The right-hand side is $call_t - put_t$ by definition. The last term on the left-hand side is $KP(t,T)$ (also by definition; we can move K outside the expectation) and by repeated use of the local characterization for the 0-dividends case and the tower property the first term on the left-hand side is S_t. ■

The portfolio proof shows that the put-call parity must hold under very mild conditions; we did not assume a particular dynamic model, nor even completeness. Put differently, it means that the put-call parity is a *hard barrier*: If your model or formula or intuition or data go against the put-call parity, then if you are not simply wrong, you at least have to tread very, very carefully. As an example: One would easily 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. With similar reasoning, the put price would be lower 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; the right-hand side does not change, hence the call and the put price are affected in the same way. (Which we happen to know from previous chapters is "not at all".)

As was the case for forward prices, non-zero dividends complicate things, but by combining the abstract proof of the basic put-call parity with the forward price reasoning in Proposition 14, we get:

Corollary 3. *(General put-call parity.) Expiry-T, strike-K put and call prices must satisfy*

$$call_t - put_t = P(t,T)(\mathrm{Fwd}(t,T) - K).$$

Propositions 12-13 can then be used to treat specific dividend structures.

We continue with two more useful and largely model-free properties of call options.

Corollary 4. *(Merton's tunnel.) The value of a call on a stock that does not pay dividends during the life of the option must satisfy*

$$S_t \geq call_t \geq \max\left(0, S_t - KP(t,T)\right).$$

Proof. The first inequality follow as $S_T \geq (S_T - K)^+$ (stock and strike price are assumed positive). The first part of the second inequality is obvious and the last part of it comes from using the put-call parity $S_t - P(t,T)K = call_t - put_t \leq call_t$ ■

It is possible to construct models in which arbitrage-free call prices are as close as we want to either bound; take a standard binomial (or a Black-Scholes) model and let σ go to infinity or 0.

Proposition 17. *Consider (European) call options on some underlying. Then:*

1. *For a fixed strike K, call prices are increasing with time to expiry, T provided interest rates are positive and the stock does not pay dividends during the life of the options.*

2. *The price of European call is a convex function the strike, K. The function is also positive and decreasing.*

Proof. For 1.: Buy $T+\epsilon$ call, sell T-call. By Merton's tunnel, this portfolio's value at time T is positive, which its time t-price must then also be. For 2.: Assume that call prices are smooth (twice differentiable) in strike. (This simplifying assumption can be relaxed, but at considerable notational cost.) Consider the butterfly spread. It's time-t price is (in obvious notation) $(call_t(K+\epsilon)-2call_t(K)+call_t(K-\epsilon))/\epsilon^2$, which must be non-negative. Letting $\epsilon \to 0$, the price converges to second derivative of the call price wrt. K. This follows from a second order Taylor expansion, $f(x \pm \epsilon) = f(x) \pm f'(x)\epsilon + \frac{1}{2}f''(x)\epsilon^2 + \mathrm{o}(\epsilon^2)$ – and the smoothness assumption. All terms are non-negative, then so is the limit, and positive a second derivative is (for smooth functions) exactly the same as convexity ■

For removal "provided interest rates"-part of the first condition to have practical effect, a combination of substantially negative interest rates, substantially positive dividends, and long time to expiry is needed.

The two conditions in Proposition 17 are not just necessary for absence of arbitrage, they are in fact sufficient in the following sense: If at time t we observe a surface of (positive) call prices that is increasing in T and decreasing and convex in K, then it is possible to construct a complete arbitrage-free dynamic stochastic model for which the theoretical call prices match the observed call prices. If we want to match with a complete model, the construction is (essentially) unique, but allowing for incomplete models gives us more degrees of freedom. The model construction is (surprisingly) … constructive[3], but fraught with practical intricacies.

6.3 Dynamic hedging and the Greeks

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 bank 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

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

[3] For an introduction see Derman & Kani, (1994) ”Riding on a Smile”, Risk, 7(2) Feb.1994, pp. 139-145.

Proposition 18. *A self-financing trading strategy replicating the dividend process of the option from time* 1 *to* T *is constructed recursively as follows: Find* $\phi_{T-1} = (\phi^0_{T-1}, \phi^1_{T-1})$ *such that*

$$\phi^0_{T-1}R + \phi^1_{T-1}S_T = \delta^c_T.$$

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

$$\phi^0_t R + \phi^1_t S_{t+1} = \phi^0_{t+1} + \phi^1_{t+1}S_{t+1}.$$

The trading strategy is self-financing by definition, replicates the call and its initial price of $\phi^0_0 + \phi^1_0 S_0$ 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 $\phi_{T-1} = (\phi^0_{T-1}, \phi^1_{T-1})$ such that

$$\phi^0_{T-1}R + \phi^1_{T-1}(S_T + \delta_T) = \delta^c_T.$$

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

$$\phi^0_t R + \phi^1_t(S_{t+1} + \delta_{t+1}) = \phi^0_{t+1} + \phi^1_{t+1}S_{t+1} + \delta^c_{t+1}. \tag{6.2}$$

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 $\phi^0_0 + \phi^1_0 S_0$.

In general, Equation (6.2) is compact notation for a whole bunch of linear equations, namely one for each submodel; the number of unknowns is equal to $\dim(S)$, the number of equations is the splitting index, i.e. the number of future states. And the equations must be solved recursively backwards. But in many applications things are simple. If S is 1-dimensional, the splitting index is 2 (so we have simple binomial submodels; refer to their states as "up" and "down"), and neither S nor the derivative pay dividends, then

- The RHSs of (6.2) are just the option's price in different states, say π^u and π^d.
- Subtract the up-equation from the down-equation to get the replication portfolio('s stock holdings)
$$\phi^1_t = \frac{\pi^u - \pi^d}{S^u - S^d} = \text{``}\frac{\Delta\pi}{\Delta S}\text{``} := \Delta.$$
- This procedure/technique is called delta hedging. A very good mnemonic. The Δ is the equation above is called the option's (delta) hedge ratio.

Further 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

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

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

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

The following can be shown (but we omit the proof):

Proposition 19. *Assume that the stock pays no dividends during the life of the option. The hedging strategy which replicates δ^c is computed as follows:*

$$\begin{aligned}\phi^1_t = \frac{cov^Q\left(\widetilde{S}_{t+1}, \widetilde{C}_{t+1} + \widetilde{\delta}^c_{t+1}\right|\mathcal{F}_t\right)}{var^Q\left(S_{t+1}\left|\mathcal{F}_t\right.\right)} & \qquad t = 0,1,\ldots,T-1 \\ \phi^0_t = \widetilde{C}_t - \phi^1_t \widetilde{S}_t & \qquad t = 0,1,\ldots,T-1\end{aligned}$$

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.

6.4 Lattices and trees

The fundamantal theorems of asset pricing absolutely do not care about properties such as Makovianity or path-independence. However, anyone having to calculate numbers does.

If the number of time periods in a model is large the Chapter 5 tree representing the stock price evolution grows very rapidly. For a binomial tree the number of nodes at time t is 2^t, and since for example $2^{20} = 1048576$ we see that when we implement this model in a spreadsheet and we wish to follow the price process π_t and the associated hedging strategy over time, we will soon run out of space. Fortunately, in many cases there is a way around this problem: If the price (and dividend and short rate) process(es) is Markov and the contingent claim we wish to price is path-independent, we can use a recombining tree – known as a lattice – to do all of our calculations. Let us look at each property in turn.

The process S is a Markov chain under Q if it satisfies

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

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 process at time $t+1$. The standard binomial model is Markov. As an example of how a non-Markov model might come about: S up, volatility down. 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

$$E^Q\left(f(S_u)\,|\mathcal{F}_t\right) = g(S_t). \tag{6.3}$$

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.

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 (S, δ, ρ) is Markov under P and the model is complete and arbitrage-free, then the model is Markov under the equivalent martingale measure Q as well.

A second condition for using a lattice to price a contingent claim is a condition on the contingent claim itself. The lattice keeps track of the number of up-jumps that have occurred for a Markovian stock price process, not the order in which they occurred. A full event tree would keep track of the exact timing of the up-jumps. That may or may not be important for a particular option.

Definition 34. *A contingent claim with dividend process δ^c is path-independent if $\delta_t = f_t(S_t)$ for some function f.*

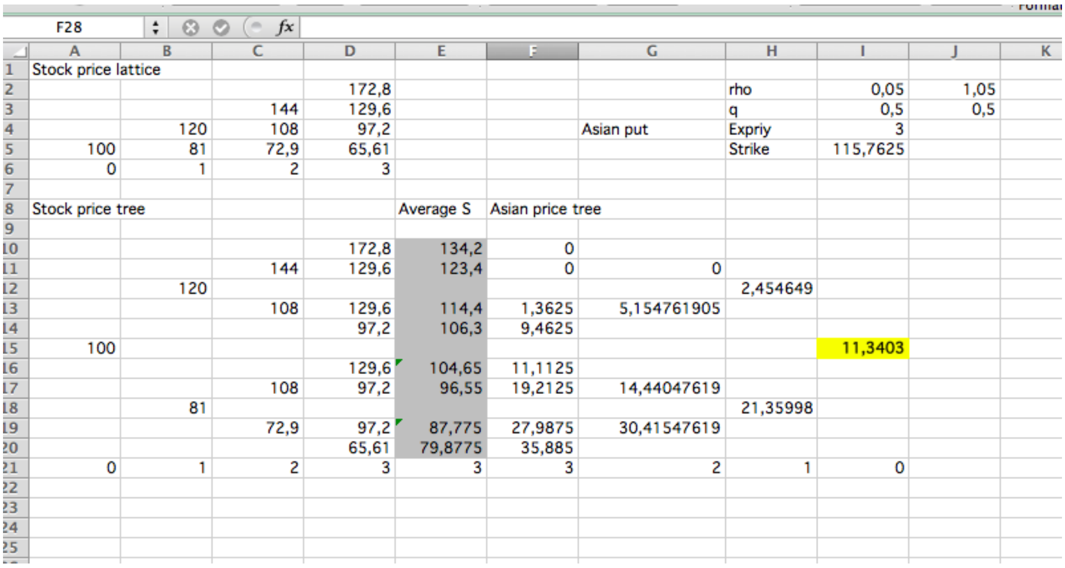
Indeed, if the claim is path-independent, the interest rate is deterministic, and the underlying process is Markov, we have

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

and the last expression is a function of S_t by the Markov property.

A European call option with expiration date T is path-independent since its only dividend payment is at time T and is given as $\max(S_T - K, 0)$. However not all options are path-independet, in which case we have to "fold out the lattice" as the next example shows.

Example 25. An Asian option Options that depend on the time-average of the stock price are – for no particular resaon – called Asian. Such options are path-dependent. Figure 6.1 shows how we price an Asian put option by turning the stock lattice into a tree. The Asian option in question is a put-type option that pays $\max(\text{Strike} - A(3), 0)$ where $A(3) = (1/4)\sum_{t=0}^{4} S(t)$ ■



F28 fx

	A	B	C	D	E	F	G	H	I	J	K
1	Stock price lattice										
2				172,8				rho	0,05	1,05	
3			144	129,6				q	0,5	0,5	
4		120	108	97,2			Asian put	Expriy	3		
5	100	81	72,9	65,61				Strike	115,7625		
6	0	1	2	3							
7											
8	Stock price tree				Average S	Asian price tree					
9											
10				172,8	134,2	0					
11			144	129,6	123,4	0	0				
12		120						2,454649			
13			108	129,6	114,4	1,3625	5,154761905				
14				97,2	106,3	9,4625					
15	100								11,3403		
16				129,6	104,65	11,1125					
17			108	97,2	96,55	19,2125	14,44047619				
18		81						21,35998			
19			72,9	97,2	87,775	27,9875	30,41547619				
20				65,61	79,8775	35,885					
21	0	1	2	3	3	3	2	1	0		
22											
23											
24											
25											

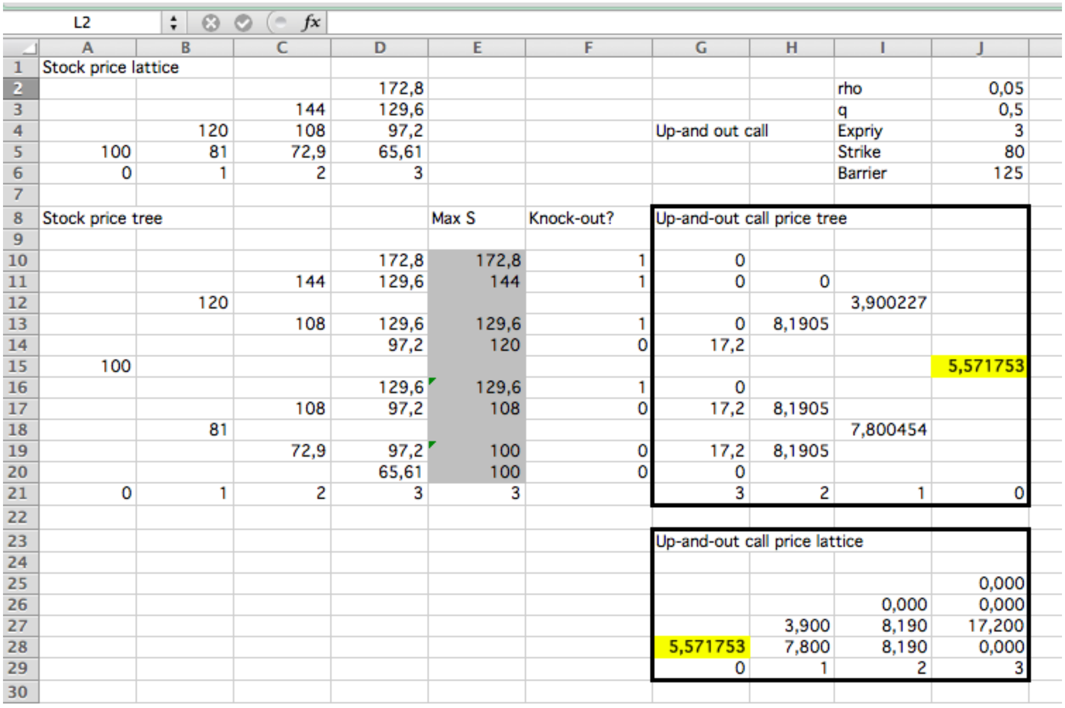
Figure 6.1: Pricing an Asian put option. File: https://tinyurl.com/nkhcsnp9

Example 26. Barrier options A barrier is an option whose payoff depends on the stock price hitting – or not hitting a certain barrier during the whole lift of the option. An up-and-out call-option with barrier B and strike B for instance would pay $\mathbf{1}_{\{\max_{0\leq u\leq T} S(u)<B\}} \times (S(T)-K)^{+}$ at T. This option is also path-dependent. We can price it by folding the stock price lattice out to a tree and keeping track of the maximal stock price along each path. Figure 6.2 shows an example. However, the barrier option is only weakly path-dependent. Given that the options has not previously been knocked out we can determine its value locally in each node when working backwards through the lattice. Or concretely for the up-and-out barrier call: We put everything above the barrier to 0 and work as usual on the rest of the lattice. The bottom right-hand part of Figure 6.2 illustrates ■

6.5 American options

An American option is one that can be exercised by its holder at any time he wishes, not just at the expiry date T. In a complete model valuing an American option is easier than it sounds: At any point in the tree (or lattice) defining the model check if the option is worth more exercised (*dead)* than unexercised (*alive*) and act accordingly. Or formulated as a recursive algorithm: Let g be the option's payoff – e.g. $g(x)=(K-x)^{+}$ for a put-option – and set $\pi_T^{AMR}=g(S_T)$, where T is the expiry date. Prior to expiry, the value of the American option is

$$\pi_t^{AMR}=\max\left(g(S_t),\frac{1}{1+\rho_t}\mathrm{E}_t^{Q}(\pi_{t+1}^{AMR})\right). \tag{6.4}$$



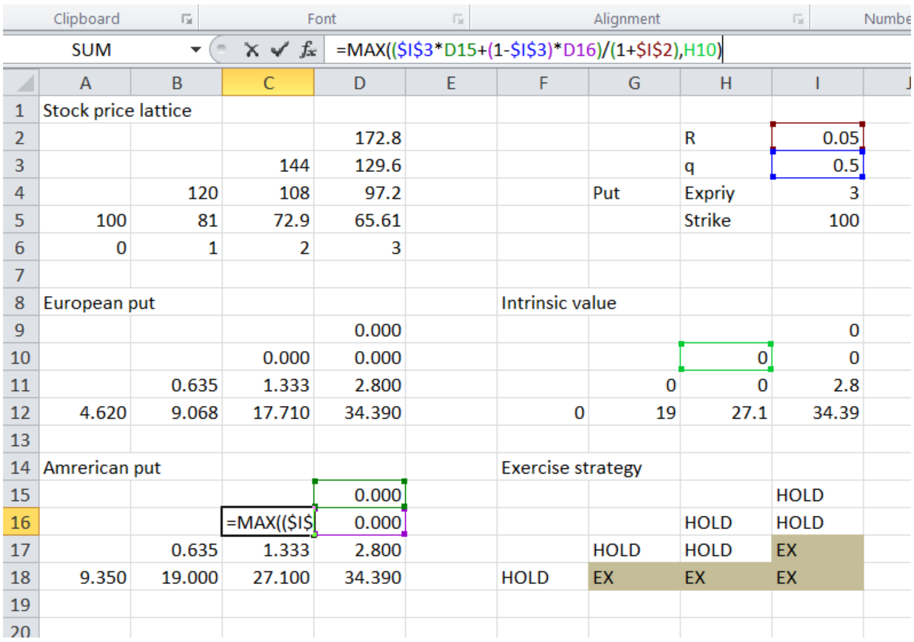
	A	B	C	D	E	F	G	H	I	J
1	Stock price lattice									
2				172,8					rho	0,05
3			144	129,6					q	0,5
4		120	108	97,2			Up-and out call		Expriy	3
5	100	81	72,9	65,61					Strike	80
6	0	1	2	3					Barrier	125
7										
8	Stock price tree				Max S	Knock-out?	Up-and-out call price tree			
9										
10				172,8	172,8	1	0			
11			144	129,6	144	1	0	0		
12		120							3,900227	
13			108	129,6	129,6	1	0	8,1905		
14				97,2	120	0	17,2			
15	100									5,571753
16				129,6	129,6	1	0			
17			108	97,2	108	0	17,2	8,1905		
18		81							7,800454	
19			72,9	97,2	100	0	17,2	8,1905		
20				65,61	100	0	0			
21	0	1	2	3	3		3	2	1	0
22										
23							Up-and-out call price lattice			
24										
25										0,000
26									0,000	0,000
27								3,900	8,190	17,200
28							5,571753	7,800	8,190	0,000
29							0	1	2	3
30										

Figure 6.2: Two ways to price a barrier option. File: https://tinyurl.com/nkhcsnp9

Notice that we can also find the optimal exercise strategy; we exercise in nodes where the first term on the right-hand side of Equation (6.4) is the larger one, otherwise we hold the option alive. To be precise, the strategy tells us whether or not we should exercise at a given node provided we have not previously exercised the option. This means that even though we can represent all things related to the American put option in a lattice, the payoff from following the optimal strategy is path-dependent. Another subtlety is the truly recursive nature of Equation (6.4); the last term on the right-hand side contains the American option price itself, we are not just taking node-wise maximum af European option prices and intrinsic values. Intuitively, this is because if we chose not to exercise our American option today, we still have the chance to do so tomorrow; it's not now or never (or: now or at expiry).

Example 27. Pricing an American put. The pricing algorithm in equation (6.4) is easy to use for computations; Figure 6.3 shows an example of an American put option calculation ■

To exercise or not? Consider an American call on a stock that does not pay out any dividends over the remaining life of the option, $[t; T]$, and assume positive interest rates,



=MAX((I3*D15+(1-I3)*D16)/(1+I2),H10)

	A	B	C	D	E	F	G	H	I	J
1	Stock price lattice									
2				172.8				R	0.05	
3			144	129.6				q	0.5	
4		120	108	97.2			Put	Expriy	3	
5	100	81	72.9	65.61				Strike	100	
6	0	1	2	3						
7										
8	European put					Intrinsic value				
9				0.000					0	
10			0.000	0.000				0	0	
11		0.635	1.333	2.800			0	0	2.8	
12	4.620	9.068	17.710	34.390		0	19	27.1	34.39	
13										
14	Amrerican put					Exercise strategy				
15				0.000					HOLD	
16			=MAX((I	0.000				HOLD	HOLD	
17		0.635	1.333	2.800			HOLD	HOLD	EX	
18	9.350	19.000	27.100	34.390		HOLD	EX	EX	EX	
19										
20										

Figure 6.3: Pricing an American put. File: https://tinyurl.com/nkhcsnp9

$P(t,T) < 1$. We then have

$$\begin{aligned} Call^{AMR}(t) \geq Call^{EU}(t) &\geq Call^{EU}(t) - Put^{EU}(t) \\ &= S(t) - P(t,T)K > S(t) - K, \end{aligned}$$

where we get the equality from the put-call parity. Comparing the left-most term to the right-most term we see that prior to expiry the call is worth strictly more *alive* then *dead* (clearly we would never exercise an out-of-the-money option). Hence it is optimal to wait as long as possible to exercise the American call which must then have the same value as its European counterpart. The previous example shows that this "hold until expiry"-argument does not work for American puts. However, there is a dual result for American puts: Zero interest rates and positive dividends make us hold them until expiry. It must be said, though, that these no-exercise results are dependent on a frictionless market; no borrowing or short-selling constraints, no transactions costs. Without a liquid market for the American option – or if replication is costly – then the most efficient way for the holder to cash in on his option may to exercise it.

Optimal stopping theory. The argument for Equation (6.4) sounds convincing. But is it really that simple and obvious? Yes, it is that simple. No, it is not that obvious. To see the latter, note that unless the whole thing were trivial (which the put price calculation example shows that it isn't), what we claim to be the American option price in Equation (6.4) violates the local characterization from Theorem 4. Where the rabbit goes into the hat is that the American option is not a *contingent claim*, which is something whose cash-flows and prices, although random, are determined exogenously by our model; i.e. by nature or the market, not specifically by our counterpart in the trade. So there is work to do in proving that Equation (6.4) is indeed the only arbitrage-free price of the American option. To do this let us first assume that the interest rate is zero and define Z as the so-called Snell envelope of the intrinsic value,

$$Z_t = \max\Big(g(S_t), \mathrm{E}_t^Q(Z_{t+1})\Big).$$

Now, Z is a Q-supermartingale and so it has a unique Doob decomposition, $Z_t = M_t - A_t$, where M is a martingale and A is a non-decreasing, predictable, 0-at-0 process. The stopping time $\tau^* = \min\{t|Z_t = g(S_t)\}$ has the very particular property that the stopped process defined via $Z_t^{\tau^*} = Z_{\min(t,\tau^*)}$ is a martingale. This is surprising because in general stopped supermartingales are only supermartingales. By considering the following two cases, we conclude that equation (6.4) gives the only arbitrage-free price:[4]

Case 1 If $\pi_t^{AMR} < Z_t$: We buy the American option, finance it (with money to spare) by selling a self-financing trading strategy that replicates Z^{τ^*} (for this to be possible, the martingale property is crucial), and exercise according to τ^*. This is an arbitrage strategy.

Case 2 If $\pi_t^{AMR} > Z_t$: We sell the American option. This can (with money to spare) finance buying a self-financing trading strategy that replicates the martingale part of Z's Doob decomposition. Because Z dominates the option's payoff and A is positive (its predictability isn't used) we can always cover our liabilities by liquidating the trading strategy. Again, an arbitrage. This differs from case 1 because we no longer control the exercise of the American option.

Applying the optional sampling theorem (twice) gives us an abstract representation of the American option price via an optimal stopping problem,

$$\pi_t^{AMR} = \sup_{\tau \in \mathcal{T}_{t,T}} \mathrm{E}_t^Q((g(S_\tau))/R_{t,\tau}), \tag{6.5}$$

where $\mathcal{T}_{t,T}$ denotes the set of stopping times with values in $[t;T]$ and R is the usual bank-account discount factor. An advantage of Equation (6.5) – whose validity is a theorem, a result; not a definition or an assumption – is pattern recognition; essentially the same formulation works in continuous-time models. Contrarily, the continuous-time analogue of Equation (6.4) isn't obvious but turns out to be a free-boundary value problem.

[4] Strictly speaking, we must also show that in case $\pi_t^{AMR} = Z_t$ there is no arbitrage, and we must cover non-zero interest rates. We leave both these things to the reader.

Chapter 7

To infinity and beyond: Black-Scholes

7.1 Black-Scholes as a limit of binomial models

Suppose we construct a binomial model covering T years, and that we divide each year into n periods. This gives a binomial model with nT periods. In each 1-period submodel choose

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

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 an expiry-T call option in the binomial model,

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

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 $\ln\left(K/(S_0 d_n^{Tn})\right)/\ln\left(u_n/d_n\right)$. Note that alternatively we may write Equation (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 $j \overset{Q'}{\sim} \text{bi}(Tn, q_n')$, with 'bi' being short for the binomial distribution. Therefore we have

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

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: $\exp(\pm x) = 1 \pm x + x^2/2 + \mathrm{o}(x^2)$. From this we get

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

Inserting this in the M_n^Q expression yields

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

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 $T \cdot n$ 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.

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

[1] Actually, the most basic De Moivre-version will not quite do because we do not have a scaled sum of *identically* distributed random variables; the two possible outcomes depend on n. You need the notion of a triangular array and the Lindeberg-Feller-version of the Central Limit Theorem.

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:

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

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

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

where Φ is the standard normal distribution function and

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

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 there 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 section we will indicate how the formula is derived, or with the previous limiting argument in mind: A different way to derive the formula. 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 comes from a continuous time framework with random variables that have continuous distributions. 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:

$$\text{If } Y \sim LN(\alpha, \sigma^2) \text{ then } E(Y) = \exp\left(\alpha + \frac{\sigma^2}{2}\right).$$

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 35. *A Brownian motion (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.*

[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.

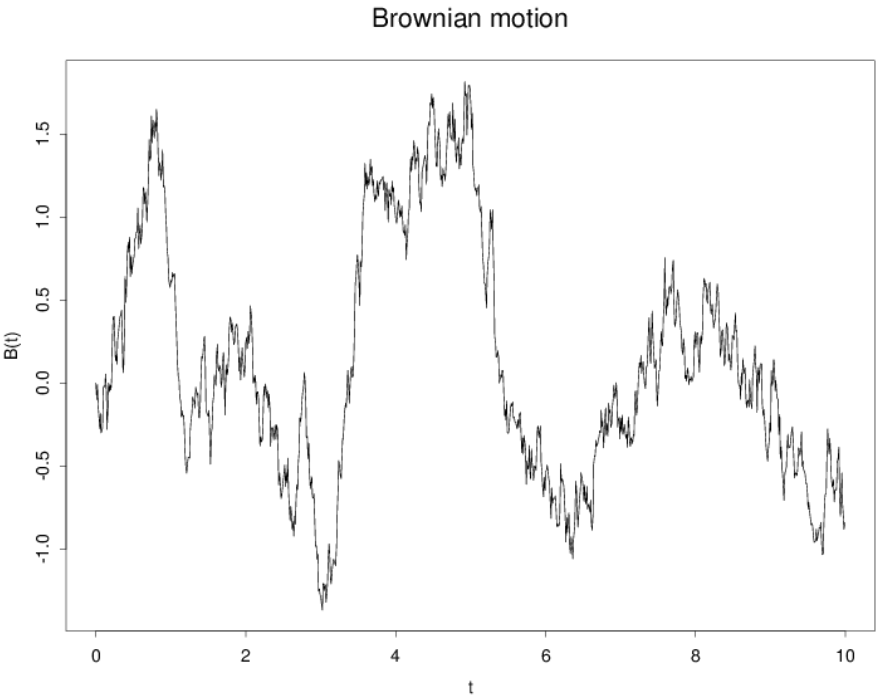


Figure 7.1: A simulated path of Brownian motion.

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 Brownian motion are continuous everywhere but differentiable nowhere. Fgure 7.1 shows a simulated sample path of a BM and should give an indication of this.

A useful fact following from the independent increment property is that for any measurable $f : \mathbb{R} \to \mathbb{R}$ for which $E\left[\left|f\left(B_t - B_s\right)\right|\right] < \infty$ we have

$$E\left[f\left(B_t - B_s\right) \middle| \mathcal{F}_s\right] = E\left[f\left(B_t - B_s\right)\right] \tag{7.3}$$

where $\mathcal{F}_s = \sigma\left\{B_u : 0 \leq u \leq s\right\}$.

The fundamental assumption of the Black-Scholes model is that the stock price can be represented by

$$S_t = S_0 \exp\left(\alpha t + \sigma B_t\right) \tag{7.4}$$

where B_t is a Brownian motion. Such a process is called a geometric Brownian motion. Furthermore, it assumes that there exists a risk-free bank account that behaves like

$$\beta_t = \exp(rt) \tag{7.5}$$

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

$$\ln\left(\frac{S_{t_1}}{S_0}\right) = \alpha t_1 + \sigma B_{t_1},$$

$$\ln\left(\frac{S_{t_2}}{S_{t_1}}\right) = \alpha(t_2 - t_1) + \sigma(B_{t_2} - B_{t_1})$$

Since αt, $\alpha(t_2 - t_1)$, 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:

$$\begin{aligned} S_t &= S_0 \exp(\alpha t + \sigma B_t) \\ \beta_t &= \exp(rt) \end{aligned}$$

[3] But skeptics would say many empirical analyses of financial data is a case of "believing is seeing" rather than the other way around.

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 risk-free asset. Equation (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 $\sigma\widetilde{B}_t \sim N\left(0, \sigma^2 t\right)$ 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]

[4] If you want to try it yourself, use

$$E\left[\frac{S_t}{\beta_t}\middle| \mathcal{F}_s\right] = E\left[\frac{S_t\beta_s S_s}{S_s\beta_t\beta_s}\middle| \mathcal{F}_s\right] = \frac{S_s}{\beta_s} E\left[\frac{S_t\beta_s}{S_s\beta_t}\middle| \mathcal{F}_s\right]$$

Accepting the fact that the call price at time 0 is

$$C_0 = \exp\left(-rT\right) E^Q \left[S_0 \exp\left(\left(r - \frac{1}{2}\sigma^2\right) T + \sigma \widetilde{B}_T\right) - K\right]^+$$

we can get the Black-Scholes formula: We know that $\sigma B_T \sim N\left(0, \sigma^2 T\right)$ and also "the rule of the unconscious statistician", which tells us that to compute $E\left[f\left(X\right)\right]$ for some random variable X which has a density $p\left(x\right)$, we compute $\int f\left(x\right) p\left(x\right) dx$. 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

$$\begin{aligned} C_0 &= e^{-rT} \int_d^\infty \left(S_0 e^{(r-\frac{1}{2}\sigma^2)T+x} - K\right) \frac{1}{\sqrt{2\pi}\sigma\sqrt{T}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2 T}} dx \\ &= \underbrace{e^{-rT} S_0 \int_d^\infty \frac{1}{\sqrt{2\pi}\sigma\sqrt{T}} e^{(r-\frac{1}{2}\sigma^2)T+x} e^{-\frac{1}{2}\frac{x^2}{\sigma^2 T}}\, dx}_{:=A} \\ &\quad \underbrace{- K e^{-rT} \int_d^\infty \frac{1}{\sqrt{2\pi}\sigma\sqrt{T}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2 T}} dx}_{:=B} . \end{aligned}$$

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

$$B = Ke^{-rT}\Phi\left(d_2\right),$$

where (as before)

$$d_2 = -\frac{d}{\sigma\sqrt{T}} = \frac{\ln\left(\frac{S_0}{K}\right) + \left(r - \frac{1}{2}\sigma^2\right) T}{\sigma\sqrt{T}}.$$

and then see if you can bring (7.3) into play and use

$$E\left[\exp\left(\sigma(B_t - B_s)\right)\right] = \exp\left(\frac{1}{2}\sigma^2(t-s)\right).$$

[5] This should bring up memories of the quantity a which we defined in the binomial model.

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 8. *The unique arbitrage-free price of a European call option on a non-dividend paying stock in the Black-Scholes model is given by*

$$C_0 = S_0\Phi\left(d_1\right) - Ke^{-rT}\Phi\left(d_2\right)$$

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 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 the time to maturity, $T - t$, i.e.

Theorem 9. *The unique arbitrage-free price at some time $t \in [0,T)$ of a European call option on a non-dividend paying stock in the Black-Scholes model is given by*

$$C_t = S_t\Phi\left(d_1\right) - Ke^{-r(T-t)}\Phi\left(d_2\right) \tag{7.7}$$

where

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

and

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

You may want to "try your hand" with conditional expectations and properties of Brownian motion by proving this.

7.3.1 Hedging the call

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 continuously adjusted strategy consisting of (the two last equations are just notation; deliberately deceptive as all good notation)

$$\phi^1(t) = \frac{\partial F}{\partial x}(t,x)\bigg|_{x=S_t} = \frac{\Delta\pi}{\Delta S} := \Delta(t)$$

units of the stock and $\phi^0(t) = \pi(t) - \phi^1(t)S_t$ Dollars in the bank account. This is called delta hedging.

For the Black-Scholes model this applies to the call with

$$\begin{aligned} F^{BScall}(t,x) = x\Phi &\left(\frac{\ln\left(\frac{x}{K}\right) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \right) \\ &- Ke^{-r(T-t)}\Phi\left(\frac{\ln\left(\frac{x}{K}\right) + (r - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \right). \end{aligned} \tag{7.8}$$

The remarkable result (and what you must forever remember) is that the partial derivative (wrt. x) of this lengthy expression is simple:

$$\frac{\partial F^{BScall}}{\partial x}(t,x) = \Phi\left(\frac{\ln\left(\frac{x}{K}\right) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \right) = \Phi(d_1), \tag{7.9}$$

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.

The proof of Equation (7.9) is prima facie a tedious exercise in differentiation, which the reader is encouraged to attempt. Here, we will provide a more subtle (but at the end of the day, also more straightforward) proof based on Euler's theorem of homogenous functions. First, observe that viewed as a function of x and K, F^{BScall} is *first order homogenous* which is to say that for any $a \in \mathbb{R}$: $F^{BScall}(ax, aK) = aF^{BScall}$. Thus, from Euler's theorem of homogenous functions[6] we then have that

$$F^{BScall}(x, K) = x\frac{\partial F^{BScall}}{\partial x} + K\frac{\partial F^{BScall}}{\partial K}. \tag{7.10}$$

Comparing Equations (7.10) and the BS formula (7.8) it is tempting to conclude that the Delta formula (7.9) obtains. While this is indeed the case, we have to be somewhat careful here: it is **NOT** true that merely by writing a first order homogenous function $g(x_1, x_2)$ on the form $x_1 h_1(x_1, x_2) + x_2 h_2(x_1, x_2)$ we can identify $h_1(x_1, x_2)$ as $\partial g/\partial x_1$ and $h_2(x_1, x_2)$ as $\partial g/\partial x_2$ (to see this, consider $g(x_1, x_2) = 0$ which we can write as both $g(x_1, x_2) = x_1 \times (-x_2/x_1) + x_2 \times 1$ and $g(x_1, x_2) = x_1 \times (-x_2^2/x_1) + x_2 \times x_2$). However, in our case it is in fact fairly obvious that

$$\frac{\partial F^{BScall}}{\partial K} = -e^{-r(T-t)}\Phi\left(d_2\right), \tag{7.11}$$

whence the unique result (7.9) follows. Specifically, the correctness of (7.11) emerges from the risk neutral pricing formula:

$$\begin{aligned}\frac{\partial}{\partial K}e^{-r(T-t)}E^Q[(S_T - K)^+] &= e^{-r(T-t)}E^Q[-\mathbf{1}\{X_T \geq K\}]\\ &= -e^{-r(T-t)}Q(S_T \geq K)\\ &= -e^{-r(T-t)}(1 - Q(S_T < K)),\end{aligned}$$

where $\mathbf{1}\{...\}$ is the indicator function.[7] Writing S_T on the form $S_t \exp((r - \frac{1}{2}\sigma^2)(T - t) + \sigma\sqrt{T - t}Z)$, where $Z \sim N(0, 1)$ we thus have

$$\begin{aligned}\frac{\partial}{\partial K}e^{-r(T-t)}E^Q[(S_T - K)^+] &= -e^{-r(T-t)}(1 - Q(Z < -d_2))\\ &= -e^{-r(T-t)}(1 - \Phi(-d_2))\\ &= -e^{-r(T-t)}\Phi(d_2).\end{aligned}$$

[6] Recall that $g(x_1, x_2)$ is said to be homogenous of degree n if $g(ax_1, ax_2) = a^n g(x_1, x_2)$. Let $x_1' = nx_1$ and $x_2' = nx_2$ then we find upon differentiating g with respect to a that

$$na^{n-1}g = \frac{\partial g}{\partial x_1'}\frac{\partial x_1'}{\partial a} + \frac{\partial g}{\partial x_2'}\frac{\partial x_2'}{\partial a} = \frac{\partial g}{\partial(ax_1)}x_1 + \frac{\partial g}{\partial(ax_2)}x_2.$$

In particular, upon setting $a = 1$ we get Euler's result for homogenous functions

$$ng = \frac{\partial g}{\partial x_1}x_1 + \frac{\partial g}{\partial x_2}x_2.$$

[7] Recall that $\mathbf{1}\{y \in \mathcal{Y}\}$ is unity just in case $y \in \mathcal{Y}$ and zero otherwise.

7.4 Implied volatility

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, we 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 36. *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

$$c_0(S_0, K, T, \sigma, r) = c^{obs}. \tag{7.12}$$

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.

There is no closed-form expression for implied volatility; Equation (7.12) must be solved numerically. Bisection works nicely (whereas a Newton-Raphson search without safety checks may diverge for deep out-of-the money options).

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 risk-free 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.

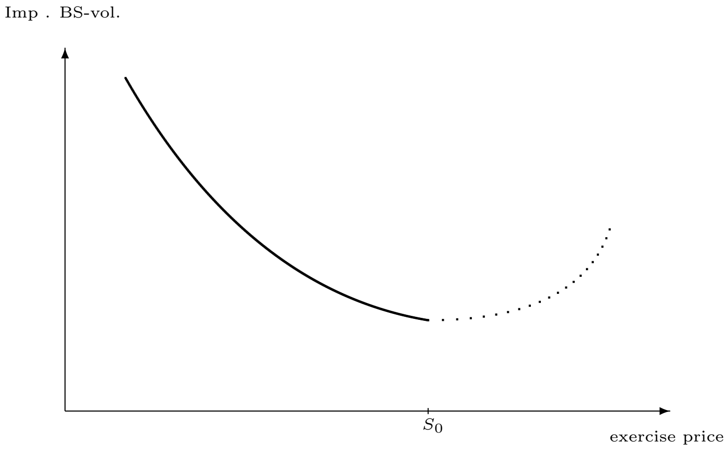
However, taking this knowledge of future volatility into account one could still imagine that all implied volatilities of options on the same underlying stock were the same (and equal to the 'anticipated' volatility). In practice this is not observed either. To get an idea of why, we consider the notion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

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.

Chapter 8

Models with curves: Stochastic interest rates

8.1 Constructing an arbitrage free model

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 multiperiod finite state space model is in the area of term structure modelling. We saw in Chapter 2 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 2 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.

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 short 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 37. *The return of the bank 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 6) we know that the system consisting of the money market account and zero coupon bonds will be arbitrage free if and only if

$$\left(\frac{P(t,T_i)}{R_{0,t}}\right)_{0\le t\le T_i}$$

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 zero coupon bond prices as expected discounted values of their terminal price 1 under a measure Q which we specify in advance (as opposed to derived from bond prices). More precisely:

Proposition 20. *Given a short 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

$$P(t,T_i) = E_t^Q\left[\frac{1}{R_{t,T_i}}\right] \textit{ for } 0 \le t \le T_i \le T,$$

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.

*P*roof. 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 we know from Example 23 that this is a martingale for each T_i ■

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.

Example 28. (Impossibility of only flat shifts of flat yield curves) If the yield curve 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, then there exists a random variable $\widetilde{r}$ such that

$$P(1,2) = \frac{1}{1+\widetilde{r}} \text{ and } P(1,3) = \frac{1}{(1+\widetilde{r})^2}$$

However, in an arbitrage-free model we have that

$$P(0,2) \quad = \quad \frac{1}{1+r} E^Q\left[P(1,2)\right] = \frac{1}{1+r} E^Q\left[\frac{1}{1+\widetilde{r}}\right]$$

and

$$P(0,3) \quad = \quad \frac{1}{1+r} E^Q\left[P(1,3)\right] = \frac{1}{1+r} E^Q\left[\frac{1}{(1+\widetilde{r})^2}\right]$$

Combining these results, we have

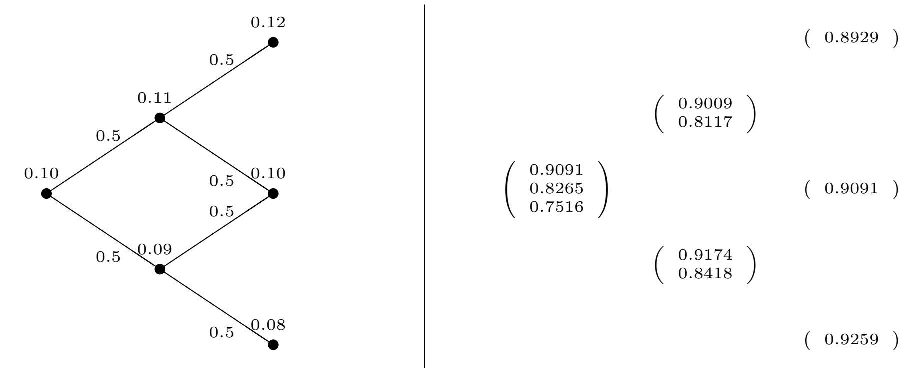
$$\frac{1}{1+r} = E^Q\left[\frac{1}{(1+\widetilde{r})}\right]$$

and

$$\frac{1}{(1+r)^2} = E^Q\left[\frac{1}{(1+\widetilde{r})^2}\right]$$

which – because $\mapsto u^2$ is strictly convex and $\widetilde{r}$ is not constant – contradicts Jensen's inequality. This explains what goes wrong in the example in Section 2.6.1. 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.

Example 29. Here is a simple illustration of the procedure in a model where the short rate follows a binomial process.



A ρ-tree with Q-probabilties

ZCB prices

The short 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 20. Note that a consequence of Proposition 20 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 $\rho_{T_i - 1}$, 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 2. 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 29 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:

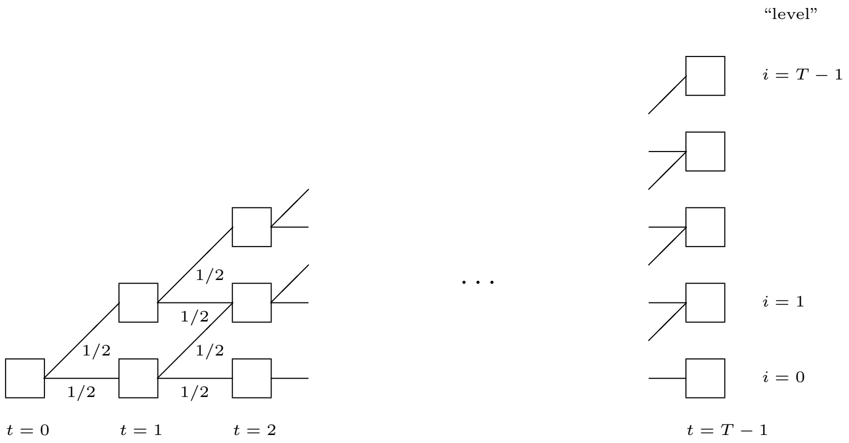


Figure 8.1: The ρ-lattice we must complete.

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 29-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.)

8.1.1 Constructing a Q-tree/lattice 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

$$P(0,2) = \frac{1}{1+\rho_0(0)}\left(\frac{1}{2}\times\frac{1}{1+\rho_1(0)} + \frac{1}{2}\times\frac{1}{1+\rho_1(1)}\right), \tag{8.1}$$

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 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, Equation (8.1) can be solved to determine 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 we 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 from a known ρ-lattice). Fortunately there is a way of doing this. We need the following lemma.

Lemma 1. *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 $s \le t$) think of

[1] Of course there is a reason for the names attached. As so often before, this is for later courses to explain.

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

$$\Pi_{t-1} = \underbrace{\left.\begin{bmatrix} \frac{1-q}{1+\rho_{t-1}(0)} & \frac{q}{1+\rho_{t-1}(0)} & & & 0 \\ & \frac{1-q}{1+\rho_{t-1}(1)} & \frac{q}{1+\rho_{t-1}(1)} & & \\ & & \ddots & \ddots & \\ 0 & & & \frac{1-q}{1+\rho_{t-1}(t-1)} & \frac{q}{1+\rho_{t-1}(t-1)} \end{bmatrix}\right\}}_{t+1 \text{ columns}} t \text{ rows}$$

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 1. Solve $\lambda_t(a_{imp}(t-1)) = P(0,t)$ numerically for $a_{imp}(t-1)$.

3. Increase t by one. If $t \leq T$ 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.

8.2 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*δ)-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:

[2] Simple *objects* are often referred to as plain vanilla *objects*. But what is seen as simple depends very much on who is looking.

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	$1/P(T_{i-1},T_i) - 1$

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)

$$\kappa_n(t) = \frac{P(t,T_0) - P(t,T_n)}{\delta \sum_{i=1}^{n} P(t,T_i)}. \tag{8.2}$$

In practice (8.2) 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

[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.

Hull's book, but 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 \$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 conceivable that someone would want. The firms contact their banks and receive the following loan offers: (Lingo: "bp" means basispoints (pronounced "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 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:

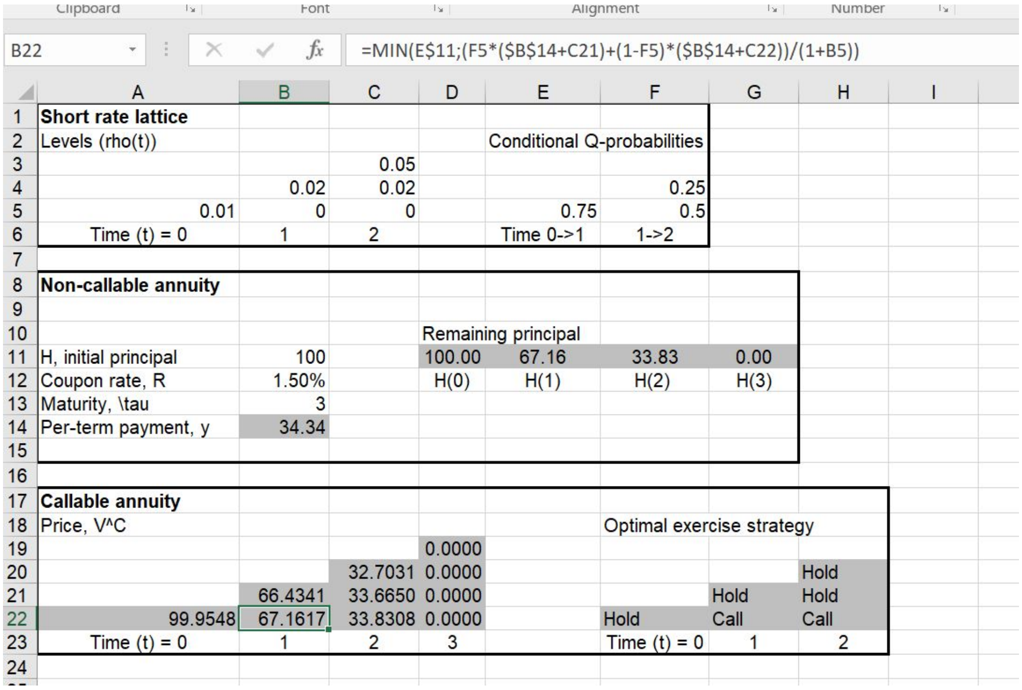
A: Pays (5Y-ZCB-rate + 50bp) (to bank), Pays 1Y-ZCB-rate (to B) and receives (5Y-ZCB-rate + 45bp) (from B). In net-terms: Pays 1Y-ZCB-rate+5bp

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



B22 =MIN(E$11;(F5*($B$14+C21)+(1-F5)*($B$14+C22))/(1+B5))

	A	B	C	D	E	F	G	H	I
1	Short rate lattice								
2	Levels (rho(t))				Conditional Q-probabilities				
3			0.05						
4		0.02	0.02			0.25			
5	0.01	0	0		0.75	0.5			
6	Time (t) = 0	1	2		Time 0->1	1->2			
7									
8	Non-callable annuity								
9									
10				Remaining principal					
11	H, initial principal	100		100.00	67.16	33.83	0.00		
12	Coupon rate, R	1.50%		H(0)	H(1)	H(2)	H(3)		
13	Maturity, \tau	3							
14	Per-term payment, y	34.34							
15									
16									
17	Callable annuity								
18	Price, V^C					Optimal exercise strategy			
19				0.0000					
20			32.7031	0.0000				Hold	
21		66.4341	33.6650	0.0000			Hold	Hold	
22	99.9548	67.1617	33.8308	0.0000		Hold	Call	Call	
23	Time (t) = 0	1	2	3		Time (t) = 0	1	2	
24									

Figure 8.2: Pricing a callable three-period annuity bond. File: https://tinyurl.com/6t8hnerr

through which the swapping takes place. This institution should ensure payments to the non-defaulting party (hence taking "credit risk" × 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.

8.3 Callable mortgage bonds

Often mortgage bonds – the bonds use to finance the loan taken by somebody buying a house or flat – are what is known as callable: At any time the borrower has the right (but not the obligation) to get out of his future commitments by (pre)paying the lender the remaining principal on the loan, say $H(t)$. This is known as calling the bond or exercising the right to prepay and it is a situation where stochastic interest rate models are crucial for determining the optimal prepayment strategy and the associated value of the bond. The lender, i.e. the buyer of the bond that the borrower issues, is no fool so she knows that a callable bond is worth less to her than the similar non-callable bond. The right to prepay gives the borrower extra possibilities. More specifically, a long position in a callable bond is equivalent to a long position in the similar non-callable bond and a short position in an American call option on the non-callable bond, where said call option has the time-varying strike $H(t)$. If he exercises, the borrower pays $H(t)$ and nothing more – which we can think of as him buying the non-callable bond for $H(t)$ from the lender and having all future payments net out. The amount $H(t)$ needed to repay the loan will often come from issuing a new bond at a lower coupon rate – prepaying is attrative when interest rates fall – but that is not a direct concern to us when valuing the callable bond.

Suppose the non-callable bond – whose time t-value we denote by $V^{NC}(t)$ – has payment $y(t)$ at time t. Let $\pi^A(t)$ denote the value of the strike-$H(t)$ American call option and $V^C(t)$ the value of the callable bond. From the argumentation above and from previously derived local charaterization and American option pricing methodology, we have for all t that

$$\text{(a)} V^C(t) = V^{NC} - \pi^A(t),$$

$$\text{(b) } V^{NC}(t) = E_t^Q\left(\frac{V^{NC}(t+1) + y(t+1)}{1+\rho_t}\right),$$

and

$$\text{(c) } \pi^A(t) = \max\left\{V^{NC}(t) - H(t), E_t^Q\left(\frac{\pi^A(t+1)}{1+\rho_t}\right)\right\}.$$

Combining these results (and using $z - \max(x,y) = \min(z-x, z-y)$) we get

$$\begin{aligned} V^C(t) &= V^{NC}(t) - \max\left\{V^{NC}(t) - H(t), E_t^Q\left(\frac{\pi^A(t+1)}{1+\rho_t}\right)\right\} \\ &= V^{NC}(t) + \min\left\{H(t) - V^{NC}(t), -E_t^Q\left(\frac{\pi^A(t+1)}{1+\rho_t}\right)\right\} \\ &= \min\left\{H(t), V^{NC}(t) - E_t^Q\left(\frac{\pi^A(t+1)}{1+\rho_t}\right)\right\} \\ &= \min\left\{H(t), E_t^Q\left(\frac{V^{NC}(t+1) + y(t+1) - \pi^A(t+1)}{1+\rho_t}\right)\right\} \\ &= \min\left\{H(t), E_t^Q\left(\frac{V^C(t+1) + y(t+1)}{1+\rho_t}\right)\right\} \end{aligned}$$

This has a nice intuitive interpretation: The borrower has two choices, he can prepay or he can keep the loan alive. If he does the latter, i.e. waits, things may get better, but he will have to make the next schedules payment on the loan. Being rational, he does what is cheaper for him.

Figure 8.2 gives a computation example of the implementation of the callable bond pricing method in a three-period model.

Appendices

Appendix A

Useful facts about probability

A.1 Expectations

Definition 38. *Let $X : \Omega \mapsto \mathbb{R}$ be a discrete random variable with possible outcomes $\{x_1, x_2, ..., x_n\}$ and the associated probability mass function $p_i \equiv P(X = x_i)$ s.t. $\sum_{i=1}^n p_i = 1$ and $\forall i : p_i \geq 0$. We then define the* ***expected value*** *of X by*

$$E[X] \equiv \sum_{i=1}^{n} x_i p_i. \tag{A.1}$$

On the other hand, if $X : \Omega \mapsto \mathbb{R}$ is a continuous random variable with density function $f : \mathbb{R} \mapsto \mathbb{R}_+$ s.t. $\int_{\mathbb{R}} f(x)dx = 1$ then the expectation is defined by

$$E[X] \equiv \int_{\mathbb{R}} x f(x) dx. \tag{A.2}$$

Note that these definitions extend trivially to conditional expectations. For discrete random variables, X and Y, $E[X|Y = y_j]$ is defined as $E[X|Y = y_j] \equiv \sum_{i=1}^n x_i p_{i|j}$, where $p_{i|j} = P(X = x_i|Y = y_j)$. For continuous random variables, X and Y, $E[X|Y = y]$ is defined as $E[X|Y = y] \equiv \int_{\mathbb{R}} x f_{X|Y}(x|y)dx$, where $f_{X|Y}(x|y) = f(x, y)/f_Y(y)$ and $f_Y(y) = \int f(x, y)dx$.
Furthermore, we have the following useful result known as the **Law of the Unconscious Statistician** (LOTUS):

Theorem 10. *The expected values of a function $g : \mathbb{R} \mapsto \mathbb{R}$ of a random variable X can be evaluated as*

$$\textit{(discrete)} \qquad E[g(X)] = \sum_{i=1}^{n} g(x_i) p_i, \tag{A.3}$$

$$\textit{(continuous)} \qquad E[g(X)] = \int_{\mathbb{R}} g(x) f(x) dx, \tag{A.4}$$

where we reiterate that p is the probability mass function of X, and f is the probability density of X. Note that this is particularly useful when one knows the distribution of X but not of $g(X)$.

Remark 4. Repeatedly below, we will refer to X, Y, and $\{X_i\}_{i\in\mathbb{N}}$ as random variables (not necessarily statistically independent unless otherwise stated). In a similar vein, $a, b, c, d, ...$ will be used to refer to arbitrary constants.

Properties: The expectation operator satisfies the following properties:

- $E[aX + b] = aE[X] + b.$

- $E[X_1 + X_2 + ... + X_n] = E[X_1] + E[X_2] + ... + E[X_n].$

- $E[X] = E[E[X|Y]]$, or in sigma-algebraic terms: $E[X|\mathcal{F}_1] = E[E[X|\mathcal{F}_2]|\mathcal{F}_1]$ where $\mathcal{F}_1 \subseteq \mathcal{F}_2$ (this is known as the **law of iterated expectations**).

- If X and Y are independent then $E[XY] = E[X]E[Y]$ and indeed we have $E[g(X)f(Y)] = E[g(X)]E[g(Y)]$ for any functions g and f. The converse statement, that "$E[XY] = E[X]E[Y]$ implies independence of X and Y", is **false**.

- For any **convex** function ξ: $\xi(E[X]) \leq E[\xi(X)]$ (Jensen's inequality).

A.2 Variance-Covariance

Definition 39. *We define the **variance** of a random variable X as*

$$Var[X] \equiv E[(X - E[X])^2] \tag{A.5}$$

which trivially can be rewritten as

$$Var[X] = E[X^2] - E[X]^2. \tag{A.6}$$

*The **standard deviation** or **volatility** of X is defined as*

$$\sigma(X) \equiv \sqrt{Var[X]}. \tag{A.7}$$

Properties: The variance operator satisfies the following properties:

- $Var[X] \geq 0.$

- $Var[a] = 0.$

- $Var[aX + b] = a^2 Var[X]$. In particular, $Var[-X] = Var[X]$.

- The variance of n random variables is

$$\begin{aligned} Var[X_1 + X_2 + ... + X_n] &= \sum_{i=1}^{n} \sum_{j=1}^{n} Cov[X_i, X_j] \\ &= \sum_{i=1}^{n} Var[X_i] + 2 \sum_{i>j} Cov[X_i, X_j]. \end{aligned}$$

The last line introduces the covariance operator. It is a straight-forward generalisation of the variance concept and is defined thusly:

Definition 40. *Let X and Y be two random variables, then we define the* ***covariance*** *between them as*

$$Cov[X, Y] \equiv E[(X - E[X])(Y - E[Y])], \tag{A.8}$$

or identically

$$Cov[X, Y] = E[XY] - E[X]E[Y]. \tag{A.9}$$

Properties: The covariance operator satisfies the following properties:

- $Cov[X, a] = 0$.
- $Cov[X, X] = Var[X]$.
- $Cov[X, Y] = Cov[Y, X]$.
- $Cov[aX, bY] = abCov[X, Y]$.
- $Cov[X + a, Y + b] = Cov[X, Y]$.
- $Cov[aX_1 + bX_2, cX_3 + cX_4] = acCov[X_1, X_3] + adCov[X_1, X_4] + bcCov[X_2, X_3] + bdCov[X_2, X_4]$.
- If X, Y are independent then $E[XY] = E[X]E[Y]$ and therefore $Cov[X, Y] = 0$ (the converse statement is again **false**).

Definition 41. *We define the* ***Pearson product-moment correlation coefficient*** *between random variables X and Y as*

$$\rho_{XY} \equiv \frac{Cov[X, Y]}{\sigma(X)\sigma(Y)}. \tag{A.10}$$

Clearly, $\rho_{XY} \in [-1, 1]$.

A.3 Higher Dimensional Quantities

The results above can readily be generalised to higher dimensional random variables. Specifically, let $\mathbf{X} = (X_1, X_2, ..., X_n)^T$ be a random vector codifying n (not necessarily identically distributed) random variables, then the **expectation** is defined as

$$E[\mathbf{X}] \equiv (E[X_1], E[X_2], ..., E[X_n])^T, \tag{A.11}$$

and the **covariance matrix** Σ of $\mathbf{X}$ is defined as

$$\Sigma \equiv E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T] \tag{A.12}$$

meaning that

$$\Sigma \equiv \begin{pmatrix} Var[X_1] & Cov[X_1, X_2] & \cdots & Cov[X_1, X_n] \\ Cov[X_2, X_1] & Var[X_2] & \cdots & Cov[X_2, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & Cov[X_n, X_2] & \cdots & Var[X_n] \end{pmatrix}. \tag{A.13}$$

Notationally, $\Sigma \equiv Var[\mathbf{X}] \equiv Cov[\mathbf{X}]$. The latter is particularly used when considering the covariance matrix between two different random vectors: $Cov[\mathbf{X}, \mathbf{Y}] \equiv E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{Y} - E[\mathbf{Y}])^T]$.

Properties: Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors of dimensionality n and m respectively. Furthermore, let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$ be constant matrices and $\boldsymbol{k} \in \mathbb{R}^m$ a constant vector.

- $\Sigma = E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}]E[\mathbf{X}^T]$.

- Symmetry: $\Sigma = \Sigma^T$.

- Positive-semidefiniteness: $\forall \mathbf{w} \in \mathbb{R}^n : \mathbf{w}^{\intercal}\Sigma\mathbf{w} \geq 0$. This is particularly important since every positive definite matrix is invertible and its inverse is also positive definite, i.e. Σ^{-1} exists and $\forall \mathbf{w} \in \mathbb{R}^n : \mathbf{w}^T\Sigma^{-1}\mathbf{w} \geq 0$.

- $Cov[\boldsymbol{A}\mathbf{X} + \boldsymbol{k}] = \boldsymbol{A}Cov[\mathrm{X}]\boldsymbol{A}^T$.

- $Cov[\mathbf{X}, \mathbf{Y}] = Cov[\mathbf{Y}, \mathbf{X}]^T$.

- $Cov[\boldsymbol{A}\mathbf{X}, \boldsymbol{B}^T\mathbf{Y}] = \boldsymbol{A}Cov[\mathbf{X}, \mathbf{Y}]\boldsymbol{B}$.

A.4 Important Distributions: Normality and Log-normality

A.4.1 The Normal Distribution

Definition 42. *A random variable X is said to be* ***normally distributed*** *with mean μ and variance σ^2, $X \sim N(\mu, \sigma^2)$, if X has the probability density function (pdf)*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{A.14}$$

In particular, the random variable Z is said to be ***standard normally distributed*** *if it is normally distributed with mean 0 and variance 1, i.e. $Z \sim N(0,1)$.*

In the literature it is customary to refer to the **standard** normal pdf by ϕ, and to the associated cumulative distribution function (cdf) by Φ.

Properties:

- The normal distribution is symmetric about the mean. In particular, this entails that the standard normal cdf has the property $\Phi(z) = 1 - \Phi(-z)$.
- Let a, b be constants and $X \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.
- If X and Y are *independent* normal random variables, then so is the sum $X + Y$. Specifically, if $X \sim N(\mu_x, \sigma_x^2)$, and $Y \sim N(\mu_y, \sigma_y^2)$ then

  $$aX + bY \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2). \tag{A.15}$$

- The central moment of $X \sim N(\mu, \sigma^2)$ is

  $$E[(X-\mu)^p] = \begin{cases} 0, & p \text{ is odd} \\ \sigma^p(p-1)!!, & p \text{ is even} \end{cases} \tag{A.16}$$

  where !! is the double factorial which runs in decrements of two: e.g. $7!! = 7\cdot5\cdot3\cdot1$.

A.4.2 The Log-normal Distribution

What is the distribution of an exponentiated normal random variable? The answer is the log-normal distribution:

Definition 43. *Suppose $X \sim N(\mu, \sigma^2)$ then $Y = e^X$ is* ***log-normally distributed****, $Y \sim \ln N(\mu, \sigma^2)$, with pdf*

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma y} \exp\left(-\frac{(\ln y-\mu)^2}{2\sigma^2}\right). \tag{A.17}$$

The mean and variance are given by

$$E[Y] = e^{\mu+\frac{1}{2}\sigma^2}, \qquad Var[Y] = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}. \tag{A.18}$$

Note that you need only memorise the first expression: the latter follows immediately from $Var[Y] = E[Y^2] - E[Y]^2$.
A particularly important result from a financial (derivative pricing) perspective is the relation

Theorem 11. *Let $X \sim N(\mu, \sigma^2)$ and $Y = e^X$ then*

$$E[\max\{Y - K, 0\}] = E[Y]\Phi\left(\frac{\mu - \ln K}{\sigma} + \sigma\right) - K\Phi\left(\frac{\mu - \ln K}{\sigma}\right), \tag{A.19}$$

where K is a positive constant.

Appendix B

A crash course in convex analysis

It is almost inevitable that an undergraduate syllabus in mathematics will fail to disseminate many of the central theorems underpinning the incredibly rich field of convex analysis. Indeed, people who major in other (more applied) quantitative subjects may have had little or no exposure to this area at all. This is unfortunate for many reasons: most pressingly the fact that the reader might potentially be left in the woods when we in Chapter 2 call on a standard result known as the *the strong separation theorem.* Thus, for the reader's convenience we here provide a brief survey of everything that is "nice to know" about convex analysis, for the purpose of reading this book.[1]

B.1 The Basics

We will start slow.

Definition 44. *Let V be a vector space. We say that a subset $S \subset V$ is convex if $\forall \lambda \in [0,1]$ and $\forall x, y \in S$:*

$$\lambda x + (1-\lambda) y \in S.$$

The basic idea is thus that if we connect any two points in S by a straight line, then the entirety of that line will be embedded by S. Thus, an n-ball is a convex set, while the n-torus is not.

Properties: Let's highlight a few elementary properties of convex subsets of a vector space:

1. The empty set and the whole vector space are convex.
2. If S and T are convex subsets, then $S \cap T$ is convex. On the other hand, $S \cup T$ generally won't be convex.

[1] There are many excellent resources dealing with convex analysis. Our own account draws pedagogical inspiration from Mark Dean's course on *Maths for Economists* (Brown University).

3. If S and T are convex subsets, then $S+T$ is convex. On the other hand, $S-T$ generally won't be convex.

Proof. We will prove the third item. Let $x = s_1 + t_1$ and $y = s_2 + t_2$ be any two points in $S+T$, where $s_1, s_2 \in S$ and $t_1, t_2 \in T$. To prove convexity we must show that $z = \lambda x + (1-\lambda)y \in S+T$, $\forall \lambda \in [0,1]$. Rewriting z as $(\lambda s_1 + (1-\lambda)s_2) + (\lambda t_1 + (1-\lambda)t_2)$, we readily see (by convexity of S and T) that $\lambda s_1 + (1-\lambda)s_2 \in S$ and $\lambda t_1 + (1-\lambda)t_2 \in T$. Thus, $z \in S+T$. To see that $S-T$ need not be convex, we provide a simple counterexample. Let $S = [0,3]$ and let $T = [1,2]$ then $S-T = [0,1) \cup (2,3]$. Let $x \in [0,1)$ and let $y \in (2,3]$. Clearly, $\lambda x + (1-\lambda)y$ generally won't lie in $S-T$ for all values of $\lambda \in [0,1]$. □

B.2 Orthogonal Projections

Through linear algebra, the reader is likely to have had prior expose to *orthogonal projections*. Here, we briefly expose how the concept can be extended to convex sets, which in turn will come in handy when we introduce the notion of *separation*.

Definition 45. *Let $S \subset \mathbb{R}^n$ and $y \in \mathbb{R}^n$. If there exists an $s^* \in S$ such that $\forall s \in S$*

$$||y - s^*|| \leq ||y - s||, \tag{B.1}$$

then s^* is the orthogonal projection of y onto S. We write $s^* = P_s(y)$.

Here, $||x||$ of a vector $x \in \mathbb{R}^n$ is the *Euclidian norm*, i.e. $||x|| = (x_1^2 + x_2^2 + ... + x_n^2)^{1/2}$. Alternatively, we write $||x|| = (x^T x)^{1/2} = (x \cdot x)^{1/2}$, where T is the *transpose* and $\cdot$ is the *dot product*.

Lemma 2. *If S is convex, then there is at most one $s^* \in S$ which satisfies (B.1).*

Proof. We will argue by contradiction: suppose there are two (unequal) vectors $s_1 \in S$ and $s_2 \in S$ which satisfy the condition. From the elementary identity $||x+y||^2 = ||x||^2 + ||y||^2 + 2x^T y$ we have

$$\begin{aligned} ||s_1 - s_2||^2 &= ||(s_1 - y) - (s_2 - y)||^2 \\ &= ||s_1 - y||^2 + ||s_2 - y||^2 - 2(s_1 - y)^T (s_2 - y). \end{aligned}$$

Similarly,

$$\begin{aligned} 4||\tfrac{s_1+s_2}{2} - y||^2 &= ||(s_1 - y) + (s_2 - y)||^2 \\ &= ||s_1 - y||^2 + ||s_2 - y||^2 + 2(s_1 - y)^T (s_2 - y). \end{aligned}$$

Adding these expressions we find

$$||s_1 - s_2||^2 = 2||s_1 - y||^2 + 2||s_2 - y||^2 - 4||\tfrac{s_1+s_2}{2} - y||^2.$$

By assumption, $||s_1 - y||$ and $||s_1 - y||$ are the same minimal distance (say δ). Furthermore, by convexity of S it follows that $\frac{s_1+s_2}{2} \in S$ whence $||\frac{s_1+s_2}{2} - y|| \geq \delta$. Thus, the only possibility is that $||s_1 - s_2||^2 = 0$ which shows that $s_1 = s_2$, contrary to our assumption. □

Uniqueness of the orthogonal projection is therefore guaranteed. Existence, of course, is a different matter entirely (if S is the convex set $(0, 1)$ and $y = 2$, then the orthogonal projection does not exist.)

Theorem 12. *Let $S \subset \mathbb{R}^n$ be a closed convex set, and let $y \in \mathbb{R}^n$. Then $P_s(y)$ exists, is unique and $s^* = P_s(y)$ if and only if $s^* \in S$ and $(y - s^*)^T(s - s^*) \leq 0$, $\forall s \in S$.*

Proof. We leave the existence proof as an exercise for the reader. Uniqueness follows from the lemma above. To show that $s^* = P_s(y) \Rightarrow (y - s^*)^T(s - s^*) \leq 0\ \forall s \in S$, we pick an $s \in S$ and a $\lambda \in [0, 1]$. Since $||s^* - y||$ is the minimal distance it follows that

$$\begin{aligned} ||s^* - y||^2 &\leq ||s^* - (\lambda s + (1 - \lambda)s^*)||^2 \\ &= ||y - s^* - \lambda(s - s^*)||^2 \\ &= ||y - s^*||^2 + \lambda^2||s - s^*||^2 - 2\lambda(y - s^*)^T(s - s^*) \end{aligned}$$

I.e. $(y - s^*)^T(s - s^*) \leq \frac{\lambda}{2}||s - s^*||^2$. Since this holds for all values of λ we can consider the case where $\lambda \to 0$, which gives us the desired result: $(y - s^*)^T(s - s^*) \leq 0$. Finally, to show that $s^* = P_s(y) \Leftarrow (y - s^*)^T(s - s^*) \leq 0\ \forall s \in S$, observe that for any $s \in S$

$$\begin{aligned} ||y - s||^2 &= ||(y - s^*) - (s - s^*)||^2 \\ &= ||y - s^*||^2 + ||s - s^*||^2 - 2\lambda(y - s^*)^T(s - s^*). \end{aligned}$$

Since $||s - s^*||^2 \geq 0$, and, by assumption, $\lambda(y - s^*)^T(s - s^*) \geq 0$ it follows that $||y - s||^2 \geq ||y - s^*||^2$ (i.e. $s^* = P_s(y)$). □

B.3 Separating Hyperplanes

Armed with the machinery above, we are now in a position to speak cogently about separating hyperplanes. Simply put, a hyperplane of an n dimensional vector space is a subspace of dimensionality $n - 1$.[2] E.g., a hyperplane in $\mathbb{R}^2$ is a 1-dimensional line, while a hyperplane in $\mathbb{R}^3$ is a 2-dimensional plane.

Definition 46. *In Cartesian coordinates, an (affine) hyperplane in $\mathbb{R}^n$ can be described by the set*

$$H(n, d) := \{x \in \mathbb{R}^n | x^T n = d\}, \tag{B.2}$$

where $n \in \mathbb{R}^n \backslash \{\emptyset\}$ and d is a real-valued constant.

[2] Equivalently, we say that it is a subspace of *codimension* 1.

Let x_0 and y_0 be two unequal coordinates in $\mathbb{R}^n$ which satisfy the hyperplane equation, then $x_0^T n = d$ and $y_0^T n = d$, whence $(x_0 - y_0)^T n = 0$. All vectors in the hyperplane are thus orthogonal to n (we say that n is *normal* to the hyperplane).

Notice that (B.2) effectively cleaves the full vector space $\mathbb{R}^2$ into two resulting half-spaces, viz. that for which $\forall x : x^T n < d$ and that for which $\forall x : x^T n > d$. This prompts us to consider the concept of *hyperplane separation.* Specifically,

Definition 47. *Let X and Y be subspaces in $\mathbb{R}^n$. We say that*

1. *X and Y are **separated** by $H(n, d)$ if*

   $$x^T n \geq d \geq y^T n,$$

   $\forall x \in X$ and $\forall y \in Y$.

2. *X and Y are **properly separated** by $H(n, d)$ if they are not both subspaces of $H(n, d)$.*

3. *X and Y are **strictly separated** by $H(n, d)$ if*

   $$x^T n > d > y^T n,$$

   $\forall x \in X$ and $\forall y \in Y$.

4. *Finally, X and Y are **strongly separated** by $H(n, d)$ if $\exists \varepsilon > 0$ s.t. $\forall x \in X$, $\forall y \in Y$*

   $$x^T n - \varepsilon ||n|| > d > y^T n + \varepsilon ||n||,$$

   or, equivalently,

   $$\inf_{x \in X} x^T n > \sup_{y \in Y} y^T n.$$

To see the difference between the latter two, consider the boxes $(0, 1) \times (-1, 0)$ and $(0, 1) \times (0, 1)$. Since the boxes do no contain their limit points, we can just about fit a hyperplane (a line) in between them. Thus, they are strictly separated. However, as there is no room to "wiggle the hyperplane", the boxes are not strongly separated. For this to be the case, we would have to displace (one of) them ever so slightly, say, by moving $(0, 1) \times (-1, 0)$ to $(0, 1) \times (-1 - \varepsilon, -\varepsilon)$, where $\varepsilon > 0$.

Lemma 3. *Consider the closed, non-empty, convex space $S \subset \mathbb{R}^n$, and the point $y \in \mathbb{R}^n \backslash \{S\}$. Then there exists an $n \in \mathbb{R}^n \backslash \{\emptyset\}$ s.t.*

$$y^T n > \sup_{x \in S} x^T n. \tag{B.3}$$

Proof. Let $x^* = P_s(y)$, and let $n := y - x^*$. From theorem 12 it follows that $\forall x \in S$: $(y - x^*)^T(x - x^*) \leq 0$, or equivalently

$$\begin{aligned} 0 &\geq n^T(x + n - y) \\ &= n^T x + n^T n - n^T y \\ &= x^T n + ||n||^2 - y^T x, \end{aligned}$$

or

$$\Leftrightarrow y^T x \geq x^T n + ||n||^2.$$

Since, by assumption, $n \notin \emptyset$ the result follow. □

Finally, we arrive at the so-called **strong separation theorem**:

Theorem 13. *Let S and T be disjoint convex subsets of $\mathbb{R}^n$. Assume S is closed, and T is compact, then they can be strongly separated.*

Proof. Recall from exercise **??** that $S - T$, under the listed conditions, will be closed, non-empty and convex. Thus, we may invoke lemma 3 to show that

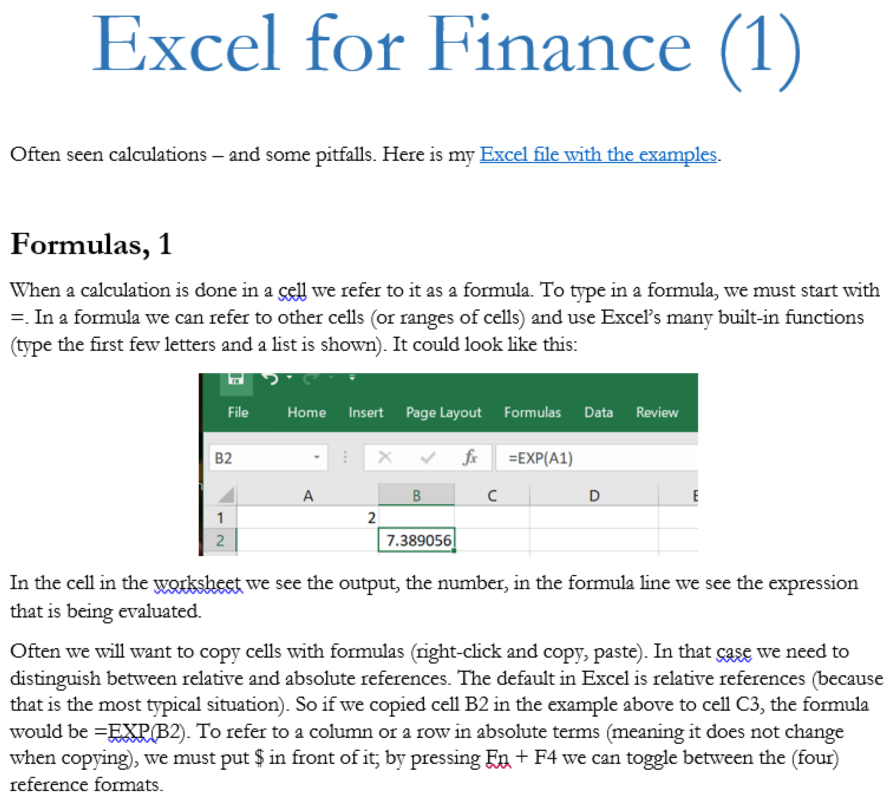
$$\begin{aligned} 0 = 0^T n &> \sup_{q \in (S-T)} q^T n = \sup_{x \in S, y \in T} (x - y)^T n \\ &= \sup_{x \in S} x^T n + \sup_{y \in T} (-y)^T n \\ &= \sup_{x \in S} x^T n - \inf_{y \in T} y^T n \end{aligned}$$

or, identically, $\inf_{y \in T} y^T n > \sup_{x \in S} x^T n$ which was to be proven. □

Appendix C

A primer on Excel for finance

Follow this link https://tinyurl.com/deu9vyr7 to get the document shown in Figure C.1.



Excel for Finance (1)

Often seen calculations – and some pitfalls. Here is my Excel file with the examples.

Formulas, 1

When a calculation is done in a cell we refer to it as a formula. To type in a formula, we must start with =. In a formula we can refer to other cells (or ranges of cells) and use Excel's many built-in functions (type the first few letters and a list is shown). It could look like this:

In the cell in the worksheet we see the output, the number, in the formula line we see the expression that is being evaluated.

Often we will want to copy cells with formulas (right-click and copy, paste). In that case we need to distinguish between relative and absolute references. The default in Excel is relative references (because that is the most typical situation). So if we copied cell B2 in the example above to cell C3, the formula would be =EXP(B2). To refer to a column or a row in absolute terms (meaning it does not change when copying), we must put $ in front of it; by pressing Fn + F4 we can toggle between the (four) reference formats.

Figure C.1: A document with a primer on Excel for finance. File: https://tinyurl.com/deu9vyr7

Index