1

Introduction and Overview

Derivative securities are financial instruments whose values are tied contractually to values of other assets, called underlying assets, at one or more points in time. One of the simplest examples is a forward contract. This is an agreement between two parties to exchange some specified quantity of a commodity or asset for a certain sum of money at a stated future time and location. For example, a forward buyer may agree to accept delivery of $1,000,000 face value of U.S. Treasury bills of six months' maturity from the forward seller on a specific date and to pay $968,000—the forward price—at that time. Delivery would likely be to an account with a bank or a broker. Once the bills are delivered and paid for, the agreement expires. The forward contract is thus a derivative on the underlying Treasury bills. To the buyer, the per-unit value of the contract at expiration is the difference between the underlying asset's spot-market price and the forward price of $968,000; to the seller, it is the negative of this. The spot-market price is the price at which one could buy 6-month Treasury bills on the open market for immediate delivery.

Derivatives are distinguished from primary assets, such as stocks, bonds, currencies, and gold. Values of these are not tied contractually to other prices, although they and other prices are surely interdependent as a consequence of the connections between markets in general economic equilibrium. We make no distinction between primary financial assets, such as stocks, bonds, and currencies, and primary "real" assets or commodities, such as gold. It is sometimes necessary to distinguish commodities that are held primarily for investment purposes—gold being a prime example—from other commodities like wheat and petroleum that are often held in inventory for use in production. Forward contracts and certain other derivatives on such underlying assets commonly come into being as firms and individuals do business and attempt to limit risk.

Forward contracts, and derivatives generally, fall into the broader class of contingent claims—contracts that mandate payments that are contingent on uncertain events. As regards the forward contract, it is the commodity's future spot price that is uncertain. Casualty and life insurance contracts, in which damage and death are the contingencies, are members of this broader class that are not normally considered derivative assets. However, the distinction begins to blur as the field of financial engineering becomes more creative. For example, one can now buy and sell bets on outcomes of temperature and snowfall in specific localities, and these are often referred to loosely as "derivative" products. Our use of the term *derivative* and our concern in this book will, however, be limited to those financial instruments described in the first sentence of this chapter.

It is precisely because derivatives' values are tied contractually to other assets that "pricing" them is a distinct subfield within economic science. While economics helps us understand the determinants of prices of goods, services, and primary assets—for example, Coca Cola and the price of Coke stock—economists might disagree profoundly were they asked to specify the prices at which these commodities "should" sell. Indeed, since no good answer is apt to be forthcoming, the question is not often asked. On the other hand, financial economists who agreed about the dynamic behavior of the price of an underlying asset would likely supply very similar answers when asked to value, or "price", a derivative. Before we can begin to see why this is so, we will need to know more about the kinds of derivatives that are available, how they are used, and how they are bought and sold.

1.1 A Tour of Derivatives and Markets

1.1.1 *Forward Contracts*

Forward agreements to deliver and accept delivery of commodities or financial instruments at a future date are alternatives to simply waiting and trading cash for commodity at the spot price that prevails at that date. Such agreements are made because at least one party finds the risk of an adverse price fluctuation undesirable. For example, the parties to the exchange of Treasury bills discussed in the introduction might be a bank and a dealer in government securities. The bank's purpose in contracting to accept forward delivery—that is, to take a "long" forward position—would be to lock in the prices of these instruments until the funds to purchase them became available. The dealer's purpose would be to earn a fee, in the form of a markup, for taking the opposite or "short" side of the transaction. Similarly, a U.S. importing firm that had contracted for a shipment of goods from a Japanese supplier might enter into a forward agreement with a bank for the purchase of Japanese yen, thus protecting against an increase in the $/¥ exchange rate. In the same way, a producer of cotton yarn might contract to deliver a shipment of such goods to a textile manufacturer at a specified price, allowing both firms to avoid the risk of adverse price fluctuation. As these examples illustrate, forward agreements are typically made between financial institutions, between productive enterprises and financial institutions, and between enterprises that produce commodities and those that use them as inputs. Because the delivery terms—date, place, and quantity—are tailored to suit the demands of the originating parties, positions in forward agreements are rarely resold to others. In particular, they are not traded on any organized exchange.

Given the delivery terms, the forward price is set so that neither party requires compensation for entering the agreement. That is, the forward price is such that the initial value of the contract is zero to both parties.[1] However, once the agreement is concluded and the spot price of the underlying commodity (or other relevant factors) begins to change, the agreement starts to acquire positive or negative value. Letting T be the time to delivery for a contract initiated at time $t = 0$, we will represent the forward price itself as $\mathsf{f}(0, T)$ (or simply as f if the times are understood) and the time-t value of the long position in the contract as $\mathfrak{F}(t, T; \mathsf{f})$, $t \in [0, T]$. Thus, $\mathfrak{F}(0, T; \mathsf{f}) = 0$ and $\mathfrak{F}(T, T; \mathsf{f}) = S_T - \mathsf{f}$, where S_T is the spot price at T. In chapter 4 we will see how to price a forward contract at any $t \in [0, T]$ and to find the forward price, $\mathsf{f}(0, T)$, that satisfies $\mathfrak{F}(0, T; \mathsf{f}) = 0$. We will also see that the special feature that makes these tasks very easy is that the contract's terminal value is linear in the spot price.

[1] But why would they enter the agreement if it had no value to them? The statement in the text, like many others to follow in this book, pertains to conditions that would apply in markets free of impediments to borrowing and transacting. We shall see in chapter 4 that in such frictionless markets either party could replicate the agreement by taking appropriate positions in the underlying commodity and riskless bonds. It is the common value of this replicating portfolio that would be zero at the agreed-upon forward price.

1.1.2 *Futures*

Futures contracts involve the same general commitments that forward contracts entail; namely, to exchange a commodity for cash at a future date. Unlike forward agreements, which are made through direct negotiation between private parties, futures commitments are bought and sold in active markets conducted on organized exchanges. Examples of these are the Chicago Board of Trade (CBOT), the Chicago Mercantile Exchange, the London International Financial Futures Exchange, the New York Futures Exchange, the Tokyo International Financial Futures Exchange, and the Toronto Futures Exchange. Most of the distinguishing features of futures contracts can be traced to the need to assure that markets for them will be active and liquid, and therefore cheap to trade in.

One such feature that distinguishes futures from forwards is that the terms of futures contracts are standardized as to quantity and quality of the commodity, as to delivery dates, and as to delivery locations.[2] For example, wheat contracts on the CBOT call for delivery of 5,000 bushels of wheat of specified types and grades to approved warehouses in Chicago (and certain other specific locations) during a particular delivery month. Besides wheat contracts, one can buy and sell futures commitments for many other agricultural products, for certain industrial commodities, and for various financial instruments. Among the last are U.S. Treasury bonds and bills, certain stock indexes, and major currencies. Trading of futures is typically done through a broker, who transmits clients' orders to the exchange. The transactions arising from the continual flow of buy and sell orders determine the futures price, which in turn determines the net cash outlay that the buyer must provide at delivery. This is explained further below.

Besides standardizing the contracts, an important step to promote market liquidity is assuring buyers and sellers that counterparties will comply with their obligations to deliver or take delivery. To this end, the exchanges operate clearing houses that serve as intermediaries to the parties and back the compliance ("performance") of buyers and sellers with their own capital and that of member brokers. Thus, one who buys (takes a long position in) wheat futures on the CBOT deals not directly with the seller of the

[2]As do some forward agreements, futures contracts typically allow the delivering party some flexibility in certain of these terms; for example, some contracts allow for delivery on any business day of the expiration month.

contract but with the CBOT's clearinghouse. Since the house simultaneously commits to deliver to the buyer and to take delivery from the seller, it has no net exposure to price risk; but it *is* exposed to risk of default by parties to the trade. It is to limit this risk that exchanges follow a practice known as marking to market, which from our perspective constitutes the major difference between futures contracts and forward contracts.

Here is how marking to market works. An individual who places an order to buy or sell futures must post a surety deposit, called the initial margin, with the broker who handles the trade. This deposit usually amounts to just a small proportion of the total value of the commodity for which delivery is contracted, because, like forward contracts when they are first arranged, futures positions at first have no net value to either party. This changes, however, as the futures price fluctuates through time and one of the parties acquires a liability—the obligation to sell at a price below the spot or to buy at a price above. To help keep unsecured losses from getting out of hand, futures contracts are, in effect, renegotiated at the end of each trading day to provide for delivery at the current futures price. If this is lower than the original price, then the long position acquires negative value equal to the product of quantity and price difference, while the short position acquires a positive value of equal magnitude. These losses and gains are then deducted from and added to the respective traders' accounts, which are thereby "marked" to market.

For example, suppose one buys a contract for 5,000 bushels of July wheat at futures price \$3.00, posting initial margin of \$500. At the end of the first day, with the futures price settling at \$2.97, the loss of (\$3.00-\$2.97) $\times$ 5,000 = \$150 is deducted from the margin balance, the original commitment to buy at \$3.00 is terminated, and a new contract is opened for delivery at the futures price \$2.97. When losses reduce the balance below some threshold, another deposit is required. If this deposit is not forthcoming, the position is closed out by the broker and the balance, if any, returned to the client. A large intraday price fluctuation that exhausted the margin balance would have to be made up by the broker were the client unable to do so.[3]

The net effect of marking to market can be seen as follows. Let $\{S_t\}_{t\geq 0}$ be the spot-price process and $\{\mathsf{F}(t,T)\}_{t\in[0,T]}$ be the futures price for a

[3]To reduce the chance of such an occurrence, exchanges limit daily price fluctuations by halting trading for the day when prices make "limit moves" from the previous day's settlement value.

contract expiring at T. Of course, at $t = T$ the futures price and spot price coincide, so that $\mathsf{F}(T, T) = S_T$. Taking the trading day as the unit of time, a long position in a contract initiated at price $\mathsf{F}(0, T)$ is worth $\mathsf{F}(1, T) - \mathsf{F}(0, T)$ per commodity unit at the end of the first trading day, $\mathsf{F}(2, T) - \mathsf{F}(0, T)$ per unit at the end of the second day, and so on. On the delivery day itself the value of the position will be $S_T - \mathsf{F}(0, T)$, just as for a forward contract initiated at $\mathsf{F}(0, T)$. Upon paying $\mathsf{F}(T, T) = S_T$ at delivery, the net cost to the long party is $S_T - [S_T - \mathsf{F}(0, T)] = \mathsf{F}(0, T)$ per unit, which was the futures price when the contract was first created. One can see, then, that the effect of marking to market is merely to distribute the net gain or loss over time, rather than allowing it all to come due at T. We consider in chapters 4 and 10 under what conditions and how this difference in timing affects the relative values of futures and forward positions and the relation between forward and futures prices.

The example just given pertained to futures positions that were held open to delivery, but in fact only a small proportion of positions are held this long. Instead, parties usually cancel their commitments to the clearing house by taking offsetting positions before expiration; that is, going long to cancel a short, and *vice versa*. For those using futures to hedge the risk of adverse price fluctuations, this eliminates the requirement to meet the exact quantity, quality, location, and time requirements specified in the contract. The futures position still provides a partial hedge for spot purchases and sales in local markets, to the extent that local spot prices in the commodity are correlated with those on which futures contracts are written.[4]

One should bear in mind that futures prices, like forward prices, are not themselves prices of traded assets. Instead, it is the futures position that is the asset, the futures price being merely the quantifiable variable that fluctuates so as to keep the values of new futures positions equal to zero.

1.1.3 *"Vanilla" Options*

Put and call options give those who hold them the rights to transact in underlying assets at specified prices during specified periods of time. There are many variations on this theme, but we consider in this section just the most basic instruments, often called "vanilla" options. One who owns (is long in) a vanilla put has the right to sell the underlying asset at a fixed

[4] Kolb and Overdahl (2006) describes in detail the operation of futures markets and the use of futures in hedging.

price, called the "strike price" or "exercise price", during or at the end of a specified period. If this option is exercised, the counterparty who is short the put has the obligation to take the other side of the trade; that is, to buy the asset at the strike price. Calls work in just the opposite way: one who is long a call has the option to buy at the strike price during or at the end of a specified period, while the counterparty must sell at this price if the holder chooses to exercise. Options not exercised by the end of the specified period are said to expire. Even vanilla puts and calls come in two flavors, having different conventions as to when they can be exercised. American options can be exercised at the discretion of the holder at any time on or before the expiration date, whereas European options can be exercised only at that specific time. The names have nothing to do with where these options are traded today.

The seller of an option is often referred to as the "writer", a term that originated before the advent of exchange trading. In those days all option contracts were negotiated between buyers and sellers or purchased from dealers, known as put and call brokers. Since opportunities for resale were limited, options were almost always held until they either expired or were exercised. Exchange trading of options was begun in 1973 by the Chicago Board Options Exchange (CBOE). Currently, exchanges in major financial centers around the world list European- and American-style options on a variety of underlying assets, including common stocks, stock indexes, and currencies. In the U.S. exchange-traded stock options are almost always American,[5] but there are European-style options on various stock indexes, such as the Dow Jones Industrials and the S&P 500. Index options, like index futures contracts, are settled in cash rather than through delivery of the underlying.[6] Exchange-traded options on spot currencies are also available, in both American and European style. Besides these spot options there are also exchange-traded futures options. A futures option gives the right to assume a position in a futures contract that expires at some later date than the option itself. One who exercises a futures put gets a short position in a futures contract plus cash equal to the difference between the strike price and the futures price. Exercising a futures call conveys a long position in the futures plus the difference between the futures price and

[5] "Flex" options available through the CBOE can be obtained in European form.

[6] One who exercises a call on an index receives from the seller in cash an amount proportional to the difference between the index value and the strike price. For puts, the exercise value is proportional to the difference between the strike price and the index.

the strike.[7] There are exchange-traded futures options on many underlying assets, including stock indexes, currencies, short- and long-term bonds, and various agricultural commodities.

As for futures contracts, the key features of most exchange-traded options—time to expiration, strike price, and quantity—are standardized in order to promote a liquid market. For example, stock options are typically for 100 shares and expire in six months or less.[8] Buying or selling an exchange-traded option (of any sort) works much like buying or selling an exchange-listed stock: an order placed through a broker is transmitted to the exchange, where it is either matched with another individual's order or accepted by the market maker.

Sales of options either reduce or close out existing long positions arising from previous purchases, or else they create short positions. Short positions may be covered or uncovered. A covered short is one that is backed by a position in the underlying (or in another derivative) that automatically enables one to meet the obligation imposed by the option holder's decision to exercise. For example, a short position in calls would be covered if backed by a long position in the underlying, which could be delivered if the call were exercised. By contrast, to meet the terms of an uncovered or "naked" short position in calls one must have the resources to purchase the underlying at the prevailing spot price. To reduce the chance of nonperformance, brokers demand deposits of cash or other securities as surety for naked short positions, just as they do for short positions in stocks. However, whether they are covered or not, short sales of options are not bound by the tick rules that restrict short sales of stocks on certain exchanges.[9] As with futures, short positions in options can be eliminated simply by buying options of the same type. As for futures also, option exchanges act as intermediaries to all transactions in order to reduce counterparty risk to traders.

[7]Note that, unlike the payoffs of spot options, payoffs of futures options are linked to futures prices and therefore not to prices of traded assets. We will see that this difference has important implications for pricing futures options.

[8]The CBOE's flex options, designed to appeal to institutional investors, do offer flexible features, but they must be traded in large quantities. There are also long-term options, called "LEAPS" with lives of up to 39 months.

[9]Securities and Exchange Commission rules in the U.S. prohibit the short sale of a stock at a price below that of the last preceding trade. A short sale at that price is allowed only if that price is itself above the last preceding different price.

Introducing the general notation used throughout for prices and contract features of standard options, we let T represent the expiration date; X, the strike price; and S_t, the value at t of the market-determined price to which the option's payoff is tied at expiration—that is, the underlying price. We use more specific notation for the underlying price in certain cases; for example, F_t for the underlying price of a futures option. Time-t prices of generic calls and puts are represented as $C(S_t, T-t)$ and $P(S_t, T-t)$, where $T-t$ is the remaining time until expiration. Other arguments are added when it is necessary to specify contract features or other parameters affecting value; for example, $C(S_t, T-t; X_1)$, $C(S_t, T-t; X_2)$ for calls with different strikes. Superscripts are used to identify specific types of options, as C^A, C^E for the American and European varieties. Taking $t = 0$ as the initiation date of an option contract, the initial values of puts and calls are thus $P(S_0, T)$ and $C(S_0, T)$, and terminal values—values at expiration if not previously exercised—are $P(S_T, 0)$ and $C(S_T, 0)$. As for futures contracts, our convention is that these represent values to one who has a long position in the option.

Terminal values of vanilla options can be stated explicitly in terms of the underlying price at T and the strike alone. Since exercising an option is not obligatory, a call (which gives the right to buy at X) would not (rationally) be exercised at expiration unless $S_T > X$. Otherwise, the option conveys nothing of value, and is said to be "out of the money". When $S_T > X$ an expiring call is "in the money" and, abstracting from transaction costs, has net value to the holder equal to $S_T - X$. Likewise, a put would not be exercised at expiration unless $S_T < X$ and would then have value $X - S_T$. Terminal values of puts and calls can therefore be expressed as

$$P(S_T, 0) = \max(X - S_T, 0) \equiv (X - S_T) \vee 0 \equiv (X - S_T)^+$$
$$C(S_T, 0) = \max(S_T - X, 0) \equiv (S_T - X) \vee 0 \equiv (S_T - X)^+.$$

Notice that when the underlying price is bounded below by zero and unbounded above (the usual case), then the terminal value of the put is bounded above by the strike price, whereas the call's value has no upper bound. One who is short a put can therefore lose no more than X, whereas one who is short an uncovered call has unlimited potential loss.

Figure 1.1 represents these elementary payoff functions at expiration. Unlike payoffs of forwards, they are clearly not linear functions of S_T. We will see in chapter 4 that this fact has far-reaching implications for pricing the options at dates before expiration; that is, for finding $P(S_t, T-t)$ and

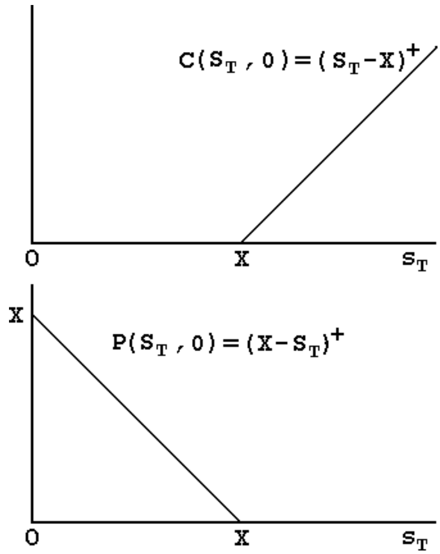

Fig. 1.1. Payoff functions of call and put options.

$C(S_t, T-t)$ for $t < T$. Pricing American options faces the additional complication that exercise can occur at any time before T. We begin in chapter 5 to develop specific techniques for pricing both American and European options, extending and developing these procedures in later chapters.

1.1.4 *Other Derivative Products*

Besides the many exchange-traded derivatives there are available over the counter—that is, through negotiated transactions with dealers and other private parties—an immense and rapidly growing array of structured derivative products. Serving institutional investors, firms involved in manufacturing, finance, and commerce, public enterprises, and governmental units, these products make it possible to hedge risks of price movements that would adversely affect cash flow and/or the values of assets. Of course, as several well-publicized incidents have demonstrated, derivatives also support purely speculative activity that can have disastrous financial consequences if there is insufficient internal oversight.

One form of specialized derivatives sold over the counter are the so-called "exotic" options, which can be grouped according to how they modify the features of vanilla puts and calls. Here is a sampling. We treat most of these in chapter 7.

1. Variations on the terminal payoff function. There are "digital" calls that pay a fixed sum when the underlying price at expiration is above the strike, and "threshold" calls that pay $S_T - X$, but only when $S_T > K > X$ for some threshold K. There are various path-dependent options whose payoffs depend on the entire path of price up to expiration; for examples, (i) "down-and-out" puts that pay $X - S_T$ provided S_t remains above some level $K < X$ for $0 \leq t \leq T$; (ii) "lookback" options whose payoffs depend on the extrema of price over $[0, T]$; and (iii) "Asian" options whose payoffs depend on the average price.
2. Variations on the underlying. There are options on other options. There are "basket" options with payoffs depending on the value of some portfolio; options on the minimum or maximum of two or more underlying prices; and "quanto" derivatives, whose payoffs in domestic currency are in fixed proportion to prices of assets denominated in a foreign currency, irrespective of changes in exchange rates.
3. Variations on what right the option confers. There are "chooser" options that let one choose at some future date whether the option is to be a put or a call.
4. Variations on the option's term. There are "forward-start" options that come to life at some future date, t^*, as at-the-money puts or calls (that is, with $X = S_{t^*}$) that expire at some $T > t^*$. There are "extendable" options that must be exercised if in the money at one or more discrete dates but are otherwise extended one or more times. There are "Bermudan" options that can be exercised at certain discrete times but do not have to be.

There is also a huge over-the-counter market in interest-rate derivatives, whose values depend primarily on prices of fixed-income assets. These include: (i) options on bonds; (ii) interest-rate "swaps", which are exchanges of payments on floating-rate loans for payments on fixed-rate loans; (iii) currency swaps, which are exchanges of loans in one currency for loans in another; (iv) options on swaps, called "swaptions"; (v) "caps" and "floors", which provide upper and lower limits to variable-rate loans; and (vi) options on caps and floors. Chapter 10 serves as an introduction to the large literature on these instruments.

1.2 An Overview of Derivatives Pricing

The theory of derivatives pricing is founded on two observations about the prices we find in markets for economic goods:

- Two identical commodities, offered for sale at the same time and place, usually sell for the same price.
- One cannot ordinarily obtain for free something that people value.

The first observation, minus the qualification *usually*, is often referred to in economics as the law of one price. The second observation—again minus the qualification—corresponds to the trite expressions, "You can't get something for nothing" and "There's no such thing as a free lunch". Of course, we recognize that both qualifiers are needed, since we have all seen exceptions to the unqualified statements. For example, the first condition can fail when not everyone is aware of what goods are available in the marketplace or not well informed about their characteristics; and the second can fail when one who owns some commodity either has some charitable motive or is simply unaware that the commodity has value to others. Of course, both conditions can fail when some coercive authority controls prices by fiat or limits individuals' ability to transact. However, few would disagree that the unqualified versions of these statements characterize situations to which things tend over time in free, competitive markets. As individuals learn through experience or from others, their actions to obtain more of the things they value will cause prices of goods, services, and assets to attain levels consistent with their marginal social values.

The failure of either of these conditions to hold in a freely functioning and frictionless market for assets presents an opportunity for arbitrage. Specifically, there is an arbitrage opportunity if one can either:

- Obtain a sure, immediate return in cash (or other good) by trading assets, or
- Get for free a claim on cash (or other good) that has a positive chance of paying off in the future.

In either case something of value—cash, good, or valuable claim—can be obtained for sure at no cost. Thus, when transacting is costless a failure of the law of one price confers the opportunity for a sure, immediate cash return. To earn it, one simply sells the more expensive asset and buys the cheaper one. Likewise, selling one asset and buying another that costs the same but adds potential cash payments conveys a valuable claim for no net outlay.

1.2.1 *Replication: Static and Dynamic*

In its simplest form pricing some financial asset by arbitrage works by (i) finding a way to replicate its payoffs exactly by assembling a portfolio of other traded assets whose prices are known, then (ii) invoking the law of one price. Arbitrage pricing would thus allow a financial firm or its client to judge the value of a derivative asset that is not already traded but that the firm might be asked to sell. The method's usefulness relies on two key conditions:

1. That markets function well enough to eliminate significant opportunities for arbitrage, and
2. That the introduction of the new asset does not appreciably change the value of the replicating portfolio.

If these conditions hold, then the law of one price equates the value of the asset to that of the replicating portfolio. What distinguishes derivatives from primary assets is that it *is* sometimes possible to replicate their payoffs and price them by arbitrage. In chapter 4 we will state formally the sufficient conditions on markets, but—depending on the derivative—we shall see that other conditions may be needed on the dynamics of the underlying price as well.

The use of arbitrage arguments for pricing financial assets can be traced at least to John B. Williams' classic 1938 monograph, *The Theory of Investment Value*. Applying what he called the "principle of conservation of investment value", Williams argued that the true worth of a firm is invariant under changes in its capital structure—that is, changes in its mix of debt and equity. Modigliani and Miller (1958) developed a rigorous proof of this proposition for firms in a frictionless economy without taxes and free of transaction costs and the legal costs associated with contracting and bankruptcy. In Miller and Modigliani (1961) they argued on similar grounds that a firm's dividend policy is also irrelevant to its total market value. All of the arguments rely on the law of one price to show that the total value of claims on a stream of earnings should not depend on how the receipts are allocated between dividends and interest. For example, individuals could in effect create their own desired debt/equity mix by combining the firm's stock with borrowing and lending on their own account, thereby replicating any preferred allocation.

Modigliani-Miller's recipe for combining stock and bonds to attain a particular degree of financial leverage is an example of "static" replication. It is static in the sense that no further trades are required once the replicating portfolio is put in place. We shall see in chapter 4 that static replication is also possible for certain derivatives. For example, one can mimic a long position in a forward agreement by buying the underlying asset now, borrowing the present value of the forward payment, and repaying the loan at delivery. Although just one transaction is involved, the portfolio's market value nevertheless equals that of the forward position at any time until expiration. However, static replication is not usually possible for more complicated derivatives whose payoffs are not linear functions of the underlying price. It remained for Black and Scholes (1973) and Merton (1973) to show how nonlinear derivatives such as European options can be valued by "dynamic" replication. Dynamic replication works by forming a self-financing portfolio of traded assets and trading back and forth between them in such a way that the portfolio is sure to have the same value as the derivative when the derivative expires. The term *self-financing* means that purchases of one asset are always financed by sales of others, so that new funds need never be added nor withdrawn. For example, we shall see in chapters 5 and 6 that the terminal payoff of a European stock option can be replicated with a self-financing portfolio of stock and bonds, assuming that the prices of these assets behave in certain ways. Again, by invoking the law of one price we can conclude that the arbitrage-free value of the derivative in a frictionless market must be that of the replicating portfolio.

1.2.2 *Approaches to Valuation when Replication is Possible*

This is all well and good, but how exactly does one find a dynamic replicating portfolio? Paradoxically, once the existence of a self-financing, replicating portfolio can be verified, it happens that the value of the derivative can be worked out directly without first finding the portfolio weights. Indeed, the solution will tell us what those weights must be. There are two general ways to work out the derivative's arbitrage-free value.

Pricing by Solving P.D.E.s

The first approach, following the path laid out by Black and Scholes (1973) and Merton (1973), is to infer from the replicating argument that the price of the derivative asset must follow a certain fundamental equation of motion. This fundamental equation is either a partial difference equation or a partial differential equation, depending on whether the underlying price is modeled in discrete time or in continuous time. In either case it describes how the arbitrage-free price of the derivative changes with respect to time and the various state variables, such as the prices of assets that comprise the replicating portfolio. The solution gives price as a weighted sum of the prices of the primary assets and thereby indicates precisely what the replicating portfolio is. Moreover, the solution specifies the portfolio weights as functions of time and the state variables, thus prescribing how the replicating portfolio is to be adjusted as the state variables evolve.

Risk-Neutral/Martingale Pricing

The second way to find a derivative's arbitrage-free price and replicating portfolio relies on the no-free-lunch property of arbitrage-free markets. Adopting Arrow's (1964) and Debreu's (1959) characterization of assets as bundles of state-contingent receipts, we can view the market value of an asset in an arbitrage-free market as simply the sum of the values of the contingent receipts it offers. In other words, it is the sum of products of the cash amounts received in various states and times multiplied by the market prices associated with those time-and-state-dependent payoffs. Each time-state price is, of course, the value of a unit cash receipt. It reflects both the likelihood of the state and the subjective trade-off between having cash to spend now versus having a unit payoff at the particular time and in the particular set of circumstances that the state represents.

These state prices have two important properties. First, state prices are positive if and only if state probabilities are positive. This is because the right to a positive chance of receiving a receipt at some date T always commands a positive price in an arbitrage-free market, whereas there can be no value to a payoff that is tied to a contingency that has no chance of occurring. Second, in an arbitrage-free market the sum of all the state prices for contingent payoffs at T must equal the price of a default-free, T-maturing, discount "unit" bond that returns one unit of currency at T in every state—i.e., a *riskless* bond. If we divide the state prices by the price of such a riskless bond, the resulting normalized state prices then have all the properties of probabilities; that is, they are nonnegative and sum to unity. Indeed, the normalized prices represent a new probability measure over states that is "equivalent" to the real one, in the technical sense that both measures assign positive values to the same sets of contingencies.

This equivalent measure affords another way to characterize an asset's arbitrage-free price. Summing products of payoffs and state prices is equivalent to summing products of payoffs and pseudo-probabilities and then multiplying the result by the price of the riskless bond. Of course, multiplying by the price of the bond is the same as discounting at the average rate of interest that the bond pays over its lifetime. Therefore, the asset's price can be thought of as the mathematical expectation of its payoffs in the pseudo-probability measure, discounted to the present at the riskless rate of interest. Since assets would actually be valued at their discounted expected payoffs in the true measure if people were risk neutral, it makes sense to refer to this pseudo-measure as the "risk-neutral" measure.

Once one has found the equivalent risk-neutral measure, there is a simple recipe for pricing a derivative security whose value at expiration (date T) is a known function of an underlying price: (i) find the mathematical expectation of the derivative's value at T in the risk-neutral measure and (ii) discount it to the present at the riskless rate. Derivatives with payoffs at more than one date can be decomposed into the dated claims, which can be priced individually in this way and then added together. Thus, we can calculate an arbitrage-free price for any derivative if we can find a new probability measure in which other assets, including its underlying, are priced as though they were riskless. The one concern is whether the risk-neutral measure and the price obtained from it are unique. As we shall see, the ability to replicate the derivative's payoff with traded assets is precisely what is needed for uniqueness.

This risk-neutral approach to pricing followed from the insights of Cox and Ross (1976), who noted that the Black-Scholes (1973) and Merton (1973) formulas for values of European puts and calls can be interpreted as discounted expected values of their terminal payoffs—$(X - S_T)^+$ or $(S_T - X)^+$—in a different measure. Later, Harrison and Kreps (1979) showed that suitably normalized prices of assets behave as martingales in the equivalent risk-neutral measure; that is, they are stochastic processes whose current values equal the conditional expectations of their values at any future date. The usual normalizing factor or numeraire is the current value of a savings account that grows continuously at the riskless rate of interest. From this perspective risk-neutral pricing of a derivative asset becomes "equivalent-martingale" pricing, and the recipe becomes:

- Choose a numeraire—a strictly positive price process by which to normalize all other relevant prices;
- Find a measure in which normalized prices of assets (and derivatives) are martingales;
- Calculate the derivative's current normalized price as the conditional expectation of its future value; and then
- Multiply by the current value of the numeraire to obtain the derivative's arbitrage-free price in currency units.

We shall fill in the details of this procedure in later chapters and apply the recipe repeatedly.

1.2.3 *Markets: Complete and Otherwise*

In the abstract, at least, martingale pricing seems simple enough, assuming that one can find the mysterious martingale measure. But even if we take for granted that it can be found, an obvious question is whether the measure and the prices derived from it are unique. If not, then a derivative could have more than one price that was consistent with the absence of arbitrage opportunities—hardly a satisfactory situation. As it turns out, uniqueness depends ultimately on the dynamics followed by the underlying assets. When they are such that the payoff of any contingent claim can be replicated with a portfolio of traded assets (dynamically or otherwise), then the market for these claims is said to be "complete". As shown by Harrison and Pliska (1981) and more rigorously by Delbaen and Schachermayer (1994), there is only one equivalent-martingale measure in a complete market—and one set of prices that denies opportunities for arbitrage.

Our first applications of martingale pricing are to models for underlying prices that do allow replication. Chapter 5 treats discrete models, in which the underlying price evolves in discrete time and has discrete outcome states. The binomial model considered there both develops our intuitive understanding of martingale pricing and provides a practical numerical method for pricing many complicated derivatives. We take up in chapter 6 the continuous-time/continuous-state dynamics that Black and Scholes used to price European options. Chapter 7, where we price various exotic options under Black-Scholes dynamics, highlights the complementarity between their p.d.e. approach to valuation and the modern martingale approach. Other complete-markets situations are encountered in chapter 10, where we take up the modeling of stochastic interest rates and the pricing of interest-sensitive derivatives.

1.2.4 *Derivatives Pricing in Incomplete Markets*

The hard fact is that certain models that allow for more realistic behavior of the dynamics of underlying assets do not allow all contingent claims to be replicated with traded assets—even dynamically. In the incomplete markets that such models generate, infinitely many measures exist that could potentially price the claims. Of these measures the market actually uses just one, but we cannot discover which one it uses without more information. What this means in practice is that derivatives cannot be priced by arbitrage arguments alone when markets are incomplete. Instead, prices will depend on free parameters that are usually interpreted as reflecting peoples' tastes for risk bearing. While these parameters are in principle discoverable, they are not readily observable.

Since the risk parameters are in fact embedded in the observed market prices of traded derivatives, one way of discovering them is to see what values of the parameters make predictions of the pricing model accord with observation. That is, if we can somehow find the equivalent-martingale prices that correspond to any given set of parameters—either by solving the p.d.e.s or by calculating the conditional expectations—then we can grope around in the parameter space, repricing at each point until we land on one at which there is a good fit with the prices quoted in the market. Chapters 8 and 9 provide the first look at these incomplete-markets models, for which the relevant parameters for pricing must be filtered in this way from the observed prices of traded derivatives.

The program ahead of us clearly involves some math. In the next two chapters we review the principal concepts of analysis, probability, and stochastic processes that will be applied in financial modeling, and we learn to use the special tools of stochastic calculus for treating processes that evolve in continuous time.