A Course on the Theory of Competitive Financial Markets

Costis Skiadas

To Robin

Preface

This text is the distillation of material I have taught to doctoral students at Northwestern University for over two decades. I have relied on lectures for broader context and on this text for a self-contained statement of a theory with classical roots in Walrasian competitive analysis, extended by Arrow and Debreu to include time and uncertainty and further developed in Finance to emphasize the role of arbitrage arguments and the tools of modern stochastic analysis. There is equal emphasis on sound economics and well-motivated methodology.

The progression of topics can be thought of as increasing in scope and decreasing in realism. Arbitrage arguments are presented first, followed by characterizations of optimality and competitive equilibrium. Arbitrage arguments utilize the assumption that the market does not allow incremental cash flows that are desirable in the narrow sense of arbitrage. Optimality is then introduced as a refinement of the no-arbitrage assumption, where the notion of a desirable cash flow is enlarged through preferences and the idea of an arbitrage is extended to allow for multi-agent transactions through profitable market-making opportunities. Restrictions on preferences are initially minimal, emphasizing the fact that competitive equilibrium notions are robust to preference structure, and are gradually strengthened in order to express ideas of how market prices and optimal consumption/portfolio choice relate to preferences for smoothing across time and states of the world. Auxiliary results on utility representations are collected in Appendix A, which includes results that, to my knowledge, have previously only been available in research paper form.

On the methodological side, a self-contained introduction to probabilistic methods starts with a rigorous treatment on a finite information tree and concludes with an introduction to the continuous-time theory, which omits several technical details but leverages the thorough understanding of the tools on a finite tree. In an approximate numerical sense, the continuous-time model is presented as a simplified special case of a high-frequency finite-information model. Tools like martingale representations, Girsanov change-of-measure arguments, the Ito calculus, forward and backward stochastic differential equations are hopefully demystified in this way, providing an entry point to a literature which is typically obfuscated by the requirements of set-theoretic rigor. The optimality and equilibrium theory emphasizes a unified

geometric viewpoint and convex analysis methods, making this course complementary to a macroeconomics course emphasizing dynamic programing methods.

The text can form the basis for either a fast-paced quarter-long course (with some material omitted) or a more relaxed semester-long course. The end-of-chapter exercises are an integral part of the book, with detailed solutions available to instructors. As for required background, the most important ingredient is graduate-level maturity in absorbing economic and mathematical ideas. Although the material is formally self-contained, a background on basic linear algebra is essential and some prior exposure to introductory probability theory and microeconomics is helpful. Prior to this course, students are asked to read Appendix B on convex analysis, with emphasis on a geometric understanding of the results rather than the details of the provided rigorous proofs.

This text is consistent in approach with my older book (Skiadas [2009]), but differs in some significant ways (as well as in numerous smaller improvements not listed here). The treatment has been streamlined around a central conceptual development and has been extended to include an introduction to continuous-time methodology, resulting is a shorter but more panoramic and tightly integrated book. The material is directly presented in a dynamic setting, as opposed to the more traditional order of first considering the static theory. The probabilistic foundations are pedagogically interweaved into the main material, as opposed to a disconnected appendix. Some of the older book's material that is not essential to the main narrative has been omitted or been relegated to the exercises, which are better integrated to the main text, classroom tested, and with available detailed solutions. Some of the peripheral theory on preferences has been significantly refined and extended and pulled into Appendix A. The overview of convex analysis in Appendix B has also been both tightened and extended (most notably, in the treatment of super-differentials and some subtler technical arguments on how convexity can substitute for compactness).

Evanston IL, February 2020

Contents

CHAPTER 1

Market and Arbitrage Pricing

An arbitrage is a trade that results in a positive incremental cash flow, that is, some inflow at some time in some contingency, but no possible outflow. This chapter introduces a formal notion of a market and lays the foundations for pricing arguments based on the assumption of the lack of arbitrage opportunities. Essential notation relating to trading strategies and associated budget equations is also introduced. Throughout this text, we use set-theoretic notation common in graduate-level mathematics and write $\equiv$ to mean "equal by definition."

1.1. Uncertainty and information

We begin with a formal representation of uncertainty and a common information stream that is available to market participants over a finite time horizon.

A **time** is an element of the set $\{0, 1, \ldots, T\}$, for a positive integer T that is fixed throughout. One of a finite number of possible states of the world, or contingencies, is realized by the **terminal time** T. These **states** are represented by the elements of a finite set Ω, the **state space.** We are not yet concerned with the likelihood of any one state occurring, but every contingency represented by a state in Ω is possible and every relevant contingency is represented by a state in Ω.

The subsets of Ω are called **events.** A **partition** of Ω is a set of mutually exclusive nonempty events whose union is Ω. Time-t information is represented by a **partition** $\mathcal{F}_t^0$ of Ω. All that is known at time t is what element of $\mathcal{F}_t^0$ contains the state realized at time T. At time 0 it is only known that the state to be revealed at time T is an element of Ω and therefore $\mathcal{F}_0^0 = \{\Omega\}$. At time T the state is revealed and therefore $\mathcal{F}_T^0 = \{\{\omega\} \mid \omega \in \Omega\}$.

Example 1.1.1. Information is generated by observing the outcome of a coin toss at each time $t = 1, 2, \ldots, T$. Let us encode heads with 1 and tails with -1. A state is a finite sequence $\omega = (\omega_1, \ldots, \omega_T)$, where $\omega_t \in \{1, -1\}$, and the state space is the cartesian product $\Omega \equiv \{1, -1\}^T$. At time $t > 0$ the first t coin toss outcomes $\bar{\omega}_1, \ldots, \bar{\omega}_t \in \{1, -1\}$ have been observed and it is therefore known that the state is an element of the event $\{\omega \in \Omega \mid \omega_1 = \bar{\omega}_1, \ldots, \omega_t = \bar{\omega}_t\}$. The partition $\mathcal{F}_t^0$ is the set of all these events as $(\bar{\omega}_1, \ldots, \bar{\omega}_t)$ ranges over $\{1, -1\}^t$. $\diamondsuit$

As in the preceding example, we assume perfect recall: If at some time the state is known to belong to a partition element, the same remains true at all subsequent times. More formally, we assume that if $u > t$, the partition $\mathcal{F}_u^0$ is a **refinement** of the partition $\mathcal{F}_t^0$, meaning that every event in $\mathcal{F}_t^0$ is the union of events in $\mathcal{F}_u^0$.

It is mathematically convenient to also define the sets

$$\mathcal{F}_t = \left\{F \mid F \text{ is a union of elements of } \mathcal{F}_t^0\right\}, \quad t = 0, \ldots, T. \tag{1.1.1}$$

An event F belongs to $\mathcal{F}_t$ if and only if at time t it is known whether F contains the state to be revealed at time T.

EXAMPLE 1.1.2. Let $\Omega \equiv \{1,2,3,4\}$ and $\mathcal{F}_1^0 \equiv \{\{1,2\},\{3\},\{4\}\}$. Then $\mathcal{F}_1 = \{\emptyset, \{1,2\},\{3\},\{4\},\{1,2,3\},\{1,2,4\},\{3,4\},\Omega\}$. Suppose state 1 is realized at time $T \equiv 2$. At time 1 it is known that the state is either 1 or 2. From that it can be inferred whether every one of the events in $\mathcal{F}_1$ contains 1 or not, and these are all the events about which such a claim can be made. $\diamondsuit$

Every $\mathcal{F}_t$ is an **algebra** of events, meaning that it contains $\emptyset$ and Ω, and it is closed relative to the formation of Boolean set operations: For all $A, B \in \mathcal{F}_t$, the **union** $A \cup B \equiv \{\omega \mid \omega \in A \text{ or } \omega \in B\}$, **intersection** $A \cap B \equiv \{\omega \mid \omega \in A \text{ and } \omega \in B\}$, and **set difference** $A \setminus B \equiv \{\omega \mid \omega \in A \text{ and } \omega \notin B\}$ are all elements of $\mathcal{F}_t$. In particular, for all $F \in \mathcal{F}_t$, the **complement** $F^c \equiv \Omega \setminus F$ is an element of $\mathcal{F}_t$. This definition of an algebra is of course redundant. For example, since $A \cap B = (A^c \cup B^c)^c$ and $A \setminus B = A \cap B^c$, an **algebra** (of events) is any nonempty set of events that is closed with respect to the formation of unions and complements. Besides providing convenient notation, algebras are key in formulating this text's theory in ways that can be interpreted in infinite state-space extensions, where algebras are not generated by partitions. Here we will take full advantage of the partition representation, even though most results will be stated in ways that are amenable to generalization.

The intersection of an arbitrary collection of algebras is also an algebra. (The reader can construct a simple example showing that the union of two algebras is not necessarily an algebra.) The algebra $\sigma(\mathcal{S})$ **generated** by a set of events $\mathcal{S}$ is the intersection of all algebras that include $\mathcal{S}$. It is straightforward to verify that $\mathcal{F}_t = \sigma(\mathcal{F}_t^0)$ for all t. Conversely, $\mathcal{F}_t^0$ can be recovered from $\mathcal{F}_t$ as the set of nonempty elements of $\mathcal{F}_t$ that do not have a nonempty proper subset in $\mathcal{F}_t$.

Note that $\mathcal{F}_u^0$ is a refinement of $\mathcal{F}_t^0$ if and only if $\mathcal{F}_t \subseteq \mathcal{F}_u$. This motivates the definition of a **filtration** as a time-indexed sequence $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_T$ of algebras of events, abbreviated to $\{\mathcal{F}_t\}$, such that $u \geq t$ implies $\mathcal{F}_u \supseteq \mathcal{F}_t$. We can therefore equivalently specify the information primitive of our model as a filtration $\{\mathcal{F}_t\}$ satisfying

$$\mathcal{F}_0 = \{\emptyset, \Omega\} \quad \text{and} \quad \mathcal{F}_T = 2^\Omega \text{ (the set of all subsets of } \Omega). \tag{1.1.2}$$

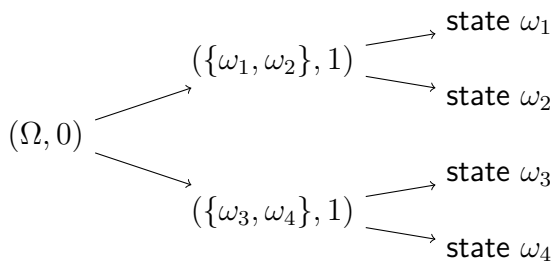


FIGURE 1.1.1. Information tree of Example 1.1.1 with $T = 2$. Each state ω_i can be identified with the terminal spot $(\{\omega_i\}, 2)$.

As illustrated in Figure 1.1.1, a filtration $\{\mathcal{F}_t\}$ can be thought of as an information tree, whose nodes correspond to what we call "spots." Formally, a **spot** of the filtration $\{\mathcal{F}_t\}$ is a pair (F, t) where $t \in \{0, \ldots, T\}$ and $F \in \mathcal{F}_t^0$. The root of the information tree corresponds to the **initial spot** $(\Omega, 0)$. A **terminal spot** takes the form $(\{\omega\}, T)$, where $\omega \in \Omega$, and can therefore be identified with the state ω as well as the unique path on the information tree from the initial spot to the given terminal spot. A nonterminal spot $(F, t-1)$, $t \in \{1, \ldots, T\}$, has **immediate successor spots** $(F_0, t), \ldots, (F_d, t)$, where $F_0, \ldots, F_d$ are the elements of $\mathcal{F}_t^0$ whose union is F. The spot $(F, t-1)$ can be thought of as the set of paths on the information tree from in the initial spot to every terminal spot corresponding to a state in F.

Economic quantities such as cash flows and prices will be represented by stochastic processes, which are time-indexed sequences of random variables that are consistent with the information structure just introduced. We now formalize these notions and introduce related notation.

A **random variable** is a function of the form $x : \Omega \to \mathbb{R}$. A **stochastic process,** or simply **process,** can equivalently be defined as a time-indexed sequence $(x_0, x_1, \ldots, x_T)$ of random variables or as a function of the form $x : \Omega \times \{0, 1, \ldots, T\} \to \mathbb{R}$, where $x_t(\omega) = x(\omega, t)$ for $\omega \in \Omega$ and $t \in \{0, \ldots, T\}$. On occasion we write $x(t)$ instead of x_t. The function $x(\omega, \cdot) : \{0, \ldots, T\} \to \mathbb{R}$, for any fixed $\omega \in \Omega$, is a **path** of the process x.

As is common in probability theory, we identify a scalar α and the process that is identically equal to α. For example, $x + \alpha$ denotes the process that takes the value $x(\omega, t) + \alpha$ at state ω and time t. Sums and products of processes are defined point-wise: $(x + yz)(\omega, t) \equiv x(\omega, t) + y(\omega, t)\, z(\omega, t)$. We write $x \leq y$ to mean $x(\omega, t) \leq y(\omega, t)$ for all (ω, t), and analogously for any other relation. A process is said to be **strictly positive** if it is valued in $(0, \infty)$.

Analogous conventions apply to random variables (which can after all be viewed as processes with $T = 0$). Random variables are often used to define events in terms of predicates, as in $\{\omega \in \Omega \mid x(\omega) \leq \alpha\}$. In such cases, we simplify the notation by eliminating the state variable, as in $\{x \leq \alpha\}$.

Another useful piece of notation is that of an **indicator function** 1_A of a set A, which takes the value 1 on A and 0 on the complement of A in the implied domain. For example, if A is an event, then 1_A is the random variable

$$1_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A; \\ 0, & \text{if } \omega \notin A. \end{cases}$$

If $A \subseteq \Omega \times \{0, \ldots, T\}$, 1_A denotes the process defined as above, but with (ω, t) in place of ω.

Suppose information is represented by a given underlying filtration $\{\mathcal{F}_t\}$ on Ω satisfying (1.1.2). If a process is to represent an observed quantity, it cannot reveal more information than implied by the postulated filtration. For example, if $T = 1$ and $\Omega = \{0, 1\}$, the process $x_0(\omega) = x_1(\omega) = \omega$ is not consistent with the information structure, because observation of the realization of x_0 at time zero reveals the state ω at time zero. To formalize this type of informational constraint, we first introduce the key notion of measurability with respect to an algebra.

A random variable x is said to be **measurable** with respect to an algebra $\mathcal{A}$, or **$\mathcal{A}$-measurable**, if $\{x \leq \alpha\} \in \mathcal{A}$ for every $\alpha \in \mathbb{R}$. If $\mathcal{A}$ is generated by the partition $\mathcal{A}^0 = \{A_1, \ldots, A_n\}$, then x is $\mathcal{A}$-measurable if and only if it can be expressed as $x = \sum_{i=1}^n \alpha_i 1_{A_i}$, where $\alpha_i \in \mathbb{R}$ is a constant value x takes on A_i.

The set of $\mathcal{A}$-measurable random variables, which we denote by $L(\mathcal{A})$, is a linear subspace of $\mathbb{R}^\Omega$ that is also closed relative to nonlinear combinations of its elements, in the following sense, where $f(x_1, \ldots, x_n)$ denotes the random variable that maps ω to $f(x_1(\omega), \ldots, x_n(\omega))$.

PROPOSITION 1.1.3. *If $x_1, \ldots, x_n \in L(\mathcal{A})$ then $f(x_1, \ldots, x_n) \in L(\mathcal{A})$ for all $f : \mathbb{R}^n \to \mathbb{R}$, and $\{(x_1, \ldots, x_n) \in S\} \in \mathcal{A}$ for all $S \subseteq \mathbb{R}^n$.*

PROOF. If each x_i is constant over every element of $\mathcal{A}^0$, then so is $f(x_1, \ldots, x_n)$. Letting $f(x_1, \ldots, x_n) = 2 \times 1_{\{(x_1, \ldots, x_n) \in S\}}$ shows that $\{(x_1, \ldots, x_n) \in S\} = \{f(x_1, \ldots, x_n) \leq 1\} \in \mathcal{A}$. □

We can now formalize the requirement that a process respects the given information structure with the notion of adaptedness. The process x is said to be **adapted** (to the underlying filtration) if $x_t \in L(\mathcal{F}_t)$ for every time t. Since $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_T = 2^\Omega$, the initial value of an adapted process is constant, while its terminal value can be any random variable.

The space of all adapted processes, which we denote by $\mathcal{L}$, can be identified with the Euclidean space $\mathbb{R}^{1+N}$, where $1+N$ is the total number of spots. To see how, consider any adapted process x. For any spot (F,t), the random variable x_t is constant over F, taking a value that we denote by $x(F,t)$ (that is, $x(\omega,t) = x(F,t)$ for all $\omega \in F$). If $\mathcal{F}_t^0 = \{F_1,\ldots,F_n\}$, then $x_t = \sum_{i=1}^n x(F_i,t)\,1_{F_i}$. One can therefore regard x as an assignment of a real number to every spot of the information tree. The set of strictly positive adapted processes is denoted by $\mathcal{L}_{++}$ and is analogously identified with $\mathbb{R}_{++}^{1+N}$.

Related to the notion of an adapted process is that of a **stopping time** defined as a function of the form

$$\tau : \Omega \to \{0,1,\ldots,T\} \cup \{\infty\},$$

provided that $\{\tau \leq t\} \in \mathcal{F}_t$ for every time t. The last restriction is equivalent to the adaptedness of the **indicator process** $1_{\{\tau \leq t\}}$, which takes the value zero prior to the (random) time τ, and the value one from time τ on (which on the event $\{\tau = \infty\}$ is never). A stopping time, or corresponding indicator process, announces the (first) arrival of an event which is consistent with the information stream encoded in the underlying filtration. For example, if x is an adapted process, then the first time that $x_t \geq 1$ defines a stopping time (with the value ∞ being assigned on the event that x remains valued below one). On the other hand, the first time that x reaches its path maximum is not generally a stopping time. For any process x and stopping time τ, the random variable x_τ or $x(\tau)$ is defined by letting $x_\tau(\omega) = x(\omega,\tau(\omega))$, with the convention $x(\omega,\infty) = 0$.

In applications, it is common to specify the filtration $\{\mathcal{F}_t\}$ as the information revealed by given processes representing observable quantities. To formally define this way of constructing the information stream, we first define a notion of information revealed by a given set of random variables.

The algebra **generated** by a set S of random variables is the intersection of all algebras relative to which every $x \in S$ is measurable, and is denoted by $\sigma(S)$. If $S = \{x_1,\ldots,x_n\}$, $\sigma(x_1,\ldots,x_n) \equiv \sigma(S)$ is the same as the algebra generated by the partition of all nonempty events of the form $\{(x_1,\ldots,x_n) = \alpha\}$, where $\alpha \in \mathbb{R}^n$. We interpret $\sigma(x_1,\ldots,x_n)$ as the information that can be inferred by observing the realization of the random variables $x_1,\ldots,x_n$. Any other variable whose realization is revealed by this information must be determined as a function of the realization of $(x_1,\ldots,x_n)$:

PROPOSITION 1.1.4. *Given any random variables $x_1,\ldots,x_n$, a random variable y is $\sigma(x_1,\ldots,x_n)$-measurable if and only if there exists a function $f : \mathbb{R}^n \to \mathbb{R}$ such that $y = f(x_1,\ldots,x_n)$.*

PROOF. That $f(x_1,\ldots,x_n)$ is $\sigma(x_1,\ldots,x_n)$-measurable follows from Proposition 1.1.3. Conversely, suppose y is $\sigma(x_1,\ldots,x_n)$-measurable

and let $\{(x_1(\omega), \ldots, x_n(\omega)) \mid \omega \in \Omega\} = \{\alpha_1, \ldots, \alpha_m\} \subseteq \mathbb{R}^n$. The algebra $\sigma(x_1, \ldots, x_n)$ is generated by the partition $\{A_1, \ldots, A_m\}$, where $A_i \equiv \{(x_1, \ldots, x_n) = \alpha_i\}$. Let $y = \sum_{j=1}^m \beta_i 1_{A_i}$, where β_i is the constant value of y on A_i. Selecting any function $f : \mathbb{R}^n \to \mathbb{R}$ such that $\beta_i = f(\alpha_i)$ results in $y = f(x_1, \ldots, x_n)$. □

A filtration is said to be **generated** by given processes $B^1, \ldots, B^d$, where each B_0^i is a constant, if

$$\mathcal{F}_t = \sigma\left(\{B_s^1, \ldots, B_s^d \mid s = 1, \ldots, t\}\right), \quad t = 1, \ldots, T.$$

Writing $B = \left(B^1, \ldots, B^d\right)'$, it follows from Proposition 1.1.4 that in this case a process x is adapted if and only if x_0 is constant and for every time $t > 0$, there exists some function $f(t, \cdot) : \mathbb{R}^{d \times t} \to \mathbb{R}$ such that

$$x(\omega, t) = f(t, B(\omega, 1), \ldots, B(\omega, t)), \quad \omega \in \Omega.$$

In other words, x_t is a function of the path of B up to time t.

1.2. Market and arbitrage

A financial market can be thought of as a set of net incremental cash flows that are generated by trading financial contracts such as bonds, stocks, futures, options and swaps. For example, the purchase of a share of a stock at some spot generates the cash flow consisting of minus the stock price at the given spot followed by the dividend stream resulting from holding the share on the information subtree rooted at the given spot. Effectively, a market facilitates the exchange of funds across spots, be it saving, borrowing, hedging or speculation, or some combination thereof that cannot be cleanly categorized. We implicitly assume that the terms of market exchanges are set competitively, meaning that there is a large number of market participants each of whom has negligible bargaining power, but who collectively influence market-clearing prices. A more formal definition of a market and related concepts follows.

Definition 1.2.1. A **cash flow** is any adapted process. A **market** is any linear subspace X of the space $\mathcal{L}$ of all cash flows. A cash flow x is said to be **traded** in X if $x \in X$. A cash flow c is an **arbitrage** if $0 \neq c \geq 0$. A market X is **arbitrage-free** if it contains no arbitrage cash flow.

All cash flows are specified in some implicit unit of account that is fixed throughout. A market X represents the set of all net incremental cash flows that are available to market participants, typically by following some trading strategy over time, a notion that is formally defined later in this chapter. From the perspective of time zero, an agent can use the market X to modify a cash flow c to $c + x$ for any $x \in X$. Of course, the agent will add a traded cash flow x to c only if

$c+x$ is preferred to c. An arbitrage cash flow x satisfies $x(F,t) \geq 0$ for every spot (F,t), and $x(F,t) > 0$ for some spot (F,t). An agent who prefers more to less prefers $c+x$ to c, for every c and every arbitrage cash flow x. If X contains an arbitrage, such an agent cannot find any c optimal given X. In a competitive equilibrium, to be defined formally in Chapter 3, agents follow optimal plans and therefore the market cannot contain arbitrage opportunities.

A market X is assumed to be a linear subspace reflecting the underlying assumption of perfectly competitive markets with no transaction costs or trading constraints. Traded cash flows can be reversed, scaled arbitrarily and combined. The discussion of market equilibrium in Chapter 3 provides an argument as to why market forces work to relax binding market constraints. Yet, frictions such as moral hazard, asymmetric information, search or information processing costs, concentrated market power and phenomena of market panics and runs can work to impede this process of increasingly complete markets and work against the definition of traded cash flows as a linear subspace.[1] This being a first course, we will mostly confine ourselves to the case of linear markets, which is why linearity is part of a market's definition. Besides its pedagogical value, the perfect case is of practical value as an approximation in thinking about highly liquid and competitive markets in standardized financial contracts as long as the limitations of such an approximation are well recognized.

Fixing a reference market X in the background, we make a technical distinction between traded cash flows, which are the elements of X, and marketed cash flows, which are cash flows that can be obtained in the market given sufficient time-zero cash. Recall that $1_{\Omega\times\{0\}}$ denotes the process that takes the value one at spot zero and the value zero everywhere else on the information tree.

DEFINITION 1.2.2. The cash flow c is **marketed** (in the market X at time zero) if $c - \alpha 1_{\Omega\times\{0\}} \in X$ for some $\alpha \in \mathbb{R}$. The market is **complete** if every cash flow is marketed.

A marketed cash flow c can be expressed as $c = \alpha 1_{\Omega\times\{0\}} + x$ for some traded cash flow x. The scalar α represents a time-zero price of c, which is unique if and only if $1_{\Omega\times\{0\}} \notin X$, a condition known as **the law of one price.** Clearly, an arbitrage-free market satisfies the law of one price but the converse is not generally true. For expositional simplicity, we will assume that the market is arbitrage-free, even where the law of one price would suffice.

[1] This text's conceptual framework and several of its arguments extend to allow for exogenous trading constraints, resulting in shapes of X that are not linear subspaces. From an economic viewpoint, the more interesting aspect of trading constraints, however, is their endogenous source in equilibrium.

DEFINITION 1.2.3. Assuming the market is arbitrage-free, the (time-zero) **present value** of a marketed cash flow c is the unique scalar α such that $c - \alpha 1_{\Omega\times\{0\}} \in X$.

Proceeding under the assumption that X is arbitrage-free, note that the set of marketed cash flows is the linear span of X and $1_{\Omega\times\{0\}}$, whose dimension is one more than that of X. The function that maps every marketed cash flow to its present value is linear and positive, where positivity means that the present value of every arbitrage cash flow is (strictly) positive. The set X is the kernel of this function, that is, the set of marketed cash flows of zero present value. If the market is complete, every cash flow is marketed and therefore has a uniquely defined present value. The dimension of X in this case is N, where $1 + N$ is the total number of spots on the information tree. In the following section we will show that if the market is not complete, then the present value function on the set of marketed cash flows can be extended to all of $\mathcal{L}$ while retaining linearity and positivity, which leads to some useful mathematical representations of the present-value function. Such an extension is not unique, however.

We have defined the market from the perspective of time zero. A market $X_{F,t}$ can also be defined analogously from the perspective of any other spot (F, t) as a subset of

$$\mathcal{L}_{F,t} \equiv \left\{x \in \mathcal{L} \mid x = x1_{F\times\{t,\dots,T\}}\right\},$$

which is the set of cash flows that can only take nonzero values on the subtree rooted at spot (F, t). Analogously to Definition 1.2.2, we have:

DEFINITION 1.2.4. A cash flow c is **marketed** at spot (F, t) in the market $X_{F,t}$ if $c1_{F\times\{t,\dots,T\}} - \alpha 1_{F\times\{t\}} \in X_{F,t}$ for some $\alpha \in \mathbb{R}$. The market $X_{F,t}$ is **complete** at (F, t) if every cash flow in $\mathcal{L}_{F,t}$ is marketed at (F, t) in $X_{F,t}$.

The following proposition introduces assumptions that allow us to construct $X_{F,t}$ in terms of the the time-zero market X.

PROPOSITION 1.2.5. *Suppose that the (time-zero) market X is arbitrage-free, and for some spot (F, t), the set $X_{F,t}$ satisfies:*

1. *(adaptedness) $X_{F,t} \subseteq \mathcal{L}_{F,t}$.*
2. *(dynamic consistency) $X_{F,t} \subseteq X$.*
3. *(liquidity) Every $x \in X$ is marketed at (F, t) in $X_{F,t}$.*

Then $X_{F,t} = X \cap \mathcal{L}_{F,t}$.

PROOF. By adaptedness and dynamic consistency, $X_{F,t} \subseteq X \cap \mathcal{L}_{F,t}$. To show the reverse inclusion, suppose that $x \in X \cap \mathcal{L}_{F,t}$. By liquidity, there exist $y \in X_{F,t}$ and $\alpha \in \mathbb{R}$ such that $x = x1_{F\times\{t,\dots,T\}} = \alpha 1_{F\times\{t\}} + y$. By dynamic consistency, $y \in X$ and therefore $\alpha 1_{F\times\{t\}} = x - y \in X$. Since X is arbitrage-free, $\alpha = 0$ and $x = y$, and therefore $x \in X_{F,t}$. □

The Proposition's assumptions have simple interpretations. The elements of $X_{F,t}$ represent cash flows available to a market participant at spot (F,t) through trading over the information subtree rooted at (F,t). Such cash flows are naturally viewed as elements of $\mathcal{L}_{F,t}$. The idea behind dynamic consistency is that at time zero a trader can have any cash flow x in $X_{F,t}$ by making a contingent plan to carry out the transactions that result in x if (F,t) materializes. The cash flow x is effectively also available to the agent at time zero, and must therefore be an element of the time-zero market X. Finally, liquidity, in the narrow technical sense used here, means that if at time zero the agent starts following a plan generating the cash flow $x \in X$, then at any future spot the agent can liquidate all positions and cancel all remaining cash flows. For example, suppose there is no uncertainty, $T = 2$ and at time zero the agent buys a bond for 99 (units of account) that pays 100 at time two, thus generating the cash flow $x = (-99, 0, 100)$. In a liquid market, the cash flow $(0, p, -100)$ is traded at time one for some price p, in other words, the same bond continues to be traded at some price.

A **dynamic market** specifies a market $X_{F,t}$ for every spot (F,t), with $X = X_{\Omega,0}$ being the time-zero market, and is **liquid** if it satisfies the third condition of Proposition 1.2.5. Under the assumptions of Proposition 1.2.5, it must be the case that $X_{F,t} = X \cap \mathcal{L}_{F,t}$ and therefore X specifies the entire liquid dynamic market. Not every time-zero market is consistent with a liquid dynamic market, however, since the liquidity requirement of Proposition 1.2.5 must be satisfied. This motivates the following definition.

Definition 1.2.6. The market X is **liquid** if for every spot (F,t), every $x \in X$ is marketed at spot (F,t) in the market $X \cap \mathcal{L}_{F,t}$.

The preceding definition abuses terminology in the interest of simplicity; liquidity is really a property of a dynamic market. It is entirely consistent to consider a dynamic market $\{X_{F,t}\}$, where $X_{\Omega,0} = X$ is a market satisfying the property of Definition 1.2.6, while for some non-zero spot (F,t), $X_{F,t}$ violates the liquidity property of Proposition 1.2.5, for example, $X_{F,t} = \{0\}$. Whenever we specify a *liquid market* X, however, it is implicitly assumed that at each spot (F,t), the market $X_{F,t} = X \cap \mathcal{L}_{F,t}$ is available, and therefore a corresponding liquid dynamic market $\{X_{F,t}\}$ is specified by X.

Liquidity requires that an initially marketed cash flow continues to be marketed, but it does not require that every cash flow is marketed. On the other hand, an arbitrage-free complete market is liquid (or, more precisely, implies a liquid dynamic market) as a consequence of the following observation.

Proposition 1.2.7. *Suppose X is an arbitrage-free complete market. Then $X \cap \mathcal{L}_{F,t}$ is complete at (F,t).*

PROOF. Consider any cash flow c. Using the assumption that X is complete, pick scalars α and β such that $-\alpha 1_{\Omega\times\{0\}} + c1_{F\times\{t,\dots,T\}} \in X$ and $-\beta 1_{\Omega\times\{0\}} + 1_{F\times\{t\}} \in X$. Since X is arbitrage-free, $\beta > 0$. Linearity of X then implies that $c1_{F\times\{t,\dots,T\}} - (\alpha/\beta)\, 1_{F\times\{t\}} \in X$. Therefore, c is marketed at (F,t) in $X \cap \mathcal{L}_{F,t}$. □

1.3. Trading and pricing of financial contracts

While market participants are ultimately interested in the incremental cash flows of the market X, they must undertake certain actions to generate these cash flows. In theory, every traded cash flow could be implemented as a buy-and-hold portfolio in contracts generating cash flows that form a linear basis of X. The problem with this approach is that if there are $1 + N$ spots, the dimension of a complete market is N, which typically rises exponentially as the number T of period increases. For example, if each spot has at least two immediate successors, then $N > 2^{T+1}$ (why?). Moderate values of T imply astronomically large values for N. Such a large number of competitively traded basic contracts is implausible even as a rough approximation of reality. A key insight is that a small number of contracts can implement a high-dimensional market provided these same contracts can be traded at every spot of the filtration. Postponing a more formal statement and proof of this claim, in this section we define contracts and we discuss their relationship to a market and some straightforward implications of the no-arbitrage assumption for contract pricing.

We use the term "contract" in a narrow formal sense to mean a dividend stream together with a price process indicating at what price the dividend stream can be traded at every spot.

DEFINITION 1.3.1. A **contract** is a pair (δ, V) of adapted processes satisfying the convention

$$\delta_0 \equiv 0 \quad \text{and} \quad \delta_T \equiv V_T. \tag{1.3.1}$$

The process δ is the contract's **dividend process** and V is the contract's **value process** or **cum-dividend price process.** The contract's **ex-dividend price process** is

$$S \equiv V - \delta.$$

We use the word "dividend" in a generalized sense to mean any cash flow that is the result of holding a long position in the contract up to time T and liquidating at time T. In applications, such payments may correspond to coupon payments and a face-value payment at maturity in the case of bonds, dividends prior to time T and the sale price cum dividend at time T in the case of a stock (whose actual dividend stream can extend beyond time T), net cash settlements in the case of a swap, and so on. A contract entitles the owner (or long position) to a single

dividend stream. An extension to include options, where the owner can select from a set of dividend streams, is discussed in Section 1.5.

The owner of a contract (δ, V) receives the dividend payment δ_t at time t. Contracts can be bought or sold at every spot. The contract's time-t value V_t represents, by convention, a cum-dividend price. If the contract is bought at time t, either the buyer pays the seller V_t and receives the time-t dividend δ_t, or the buyer pays the seller the ex-dividend price S_t and the time-t dividend goes to the seller. The dividend convention (1.3.1) is made for notational simplicity and entails no loss of generality within the scope of the model presented here. One can think of the condition $\delta_T = V_T$ as reflecting the implicit assumption that the contract is liquidated at the terminal date T, resulting in the payment $V_T = S_T + \delta_T$. It makes no difference within the model how the value V_T is split between S_T and δ_T or how V_0 is split between S_0 and δ_0; for simplicity, we set $S_T = 0$ and $\delta_0 = 0$.

DEFINITION 1.3.2. A contract (δ, V) is **traded** at spot (F, t) in the market X if

$$x = -V1_{F\times\{t\}} + \delta 1_{F\times\{t,\dots,T\}} \in X. \tag{1.3.2}$$

The cash flow x is **generated by buying** the contract at spot (F, t), while $-x$ is **generated by selling** the contract at the same spot. The contract (δ, V) is **traded** (in X) if it is traded at every spot.

We proceed taking as given a reference arbitrage-free market X. When we say that a contract is traded (at some spot), it is implied that the contract is traded in the reference market X. It is an immediate consequence of the definitions that a cash flow δ is marketed in X if and only if there is a contract (δ, V) that is traded at time zero, in which case V_0 is the present value of δ. Note that this simple statement relies critically on the convention $V_T = \delta_T$. The essence of the conclusion is that V_0 is the present value of the dividends paid out prior to the terminal date plus the present value of the contract's terminal (cum-dividend) value.

REMARK 1.3.3. The trades generating an arbitrage as a consequence of the violation of a claimed arbitrage pricing relationship are instructive in that they suggest potential ways in which the pricing relationship can be violated in realistic applications due to frictions left out of the formal model. For a simple example, suppose the contracts (δ, V) and (δ, V') are both traded, but $V_0 < V_0'$. The arbitrage $\left(V_0' - V_0\right) 1_{\Omega\times\{0\}}$ results by selling the second contract and buying the first one. Suppose the latter represents a security, like a US treasury bond, that can be used for the purpose of posting collateral, thus facilitating other trades. If collateral in this sense is scarce in equilibrium, the present value of the cash dividend stream δ may not present adequate compensation for parting with the security and the value V_0 may

well exceed the present value of δ. In the presence of a well-functioning competitive lease market in the security, V_0 equals the present value of the equilibrium lease payments resulting from lending out the security. The extent to which these lease payments exceed the dividends δ reflects the equilibrium value of holding an additional unit of inventory. $\diamondsuit$

Returning to our current idealized formal setup, the preceding observations can be applied from the perspective of any other spot, and therefore for every traded contract (δ, V), the value $V(F,t)$ is a function of the restriction of δ on $F \times \{t, \dots, T\}$, which is the present value of δ from the perspective of spot (F,t). Thus, in an arbitrage-free market, any two traded contracts (δ, V^1) and (δ, V^2) with a common dividend process δ (and therefore common terminal value $V^1_T \equiv \delta_T \equiv V^2_T$) must also have a common value process: $V^1 = V^2$. It is also worth noting that buying the contract (δ, V) at spot $(F, t-1)$ and selling it at each of its immediate successor spots generates the cash flow

$$x = -S1_{F\times\{t-1\}} + V1_{F\times\{t\}}, \tag{1.3.3}$$

where $S \equiv V - \delta$. Therefore, if the contracts (δ^1, V^1) and (δ^2, V^2) are traded and $V^1_t = V^2_t$ on some $F \in \mathcal{F}_{t-1}$, then $S^1_{t-1} = S^2_{t-1}$ on F. As a corollary, if (δ^1, V^1) and (δ^2, V^2) are traded in an arbitrage-free market and $V^1_t = V^2_t$ for all $t > 0$, then the two contracts are identical.

We have defined what it means for a contract to be traded in a given market. Conversely, trading in a given set of contracts implements a market, which can be succinctly defined as follows.

Definition 1.3.4. The market **implemented** by a set of contracts is the intersection (and hence smallest relative to inclusion) market in which all contracts in the given set are traded.

Suppose $\mathcal{C}$ is a set of contracts, let X be the market implemented by $\mathcal{C}$, and let X^0 be the set of all cash flows of the form (1.3.2) for every $(\delta, V) \in \mathcal{C}$ and spot (F,t). The set $\text{span}(X^0)$ of all finite linear combinations of elements of X^0 is a market that clearly includes X and in which every contract in $\mathcal{C}$ is traded. Therefore, $X = \text{span}(X^0)$. This construction also makes it clear that X is liquid (why?). An element of $\text{span}(X^0)$ can be thought of as a contingent plan to buy or sell contracts at various spots, in other words, a trading strategy. A contract that is not in $\mathcal{C}$ but is traded in the arbitrage-free market X is synthetic in $\mathcal{C}$, meaning that it is generated by a trading strategy in contracts in $\mathcal{C}$. Section 1.6 discusses trading strategies, synthetic contracts and associated budget equations more systematically. In the following chapter, we will see that the minimum number of contracts implementing a complete market is equal to the maximum number of immediate successor spots to each spot.

1.4. Present-value functions

A dual approach to arbitrage pricing utilizes the notion of a present-value function, formally defined below in terms of a reference market X that is taken as given throughout this section. Recall that a **linear functional** on a vector space is a real-valued linear function whose domain is the entire vector space. A linear functional is **positive** if it assigns a positive value to every vector x such that $0 \neq x \geq 0$.

DEFINITION 1.4.1. A (time-zero) **present-value function** (for the market X) is a positive linear functional Π on $\mathcal{L}$ such that $\Pi(x) \leq 0$ for all $x \in X$ and $\Pi(1_{\Omega\times\{0\}}) = 1$.

A present-value function Π specifies a time-zero value for every cash flow c, marketed or not, which is positive if c is an arbitrage cash flow. The essential restriction that Π is nonpositive on X is equivalent to $\Pi(x) = 0$ for all $x \in X$, since X is a linear subspace. (In extensions with trading constraints, X is no longer a linear subspace, but Definition 1.4.1 is still valid, and the present value of a traded cash flow can be strictly negative.) The requirement $\Pi(1_{\Omega\times\{0\}}) = 1$ is merely a normalization.

The properties of a present-value function Π combine to determine the value $\Pi(c)$ of a marketed cash flow c as the present value of c in the sense of Definition 1.2.3. Suppose the scalar α is such that $c - \alpha 1_{\Omega\times\{0\}} \in X$. The existence of Π implies that X is arbitrage-free (why?) and therefore α is unique. The fact that Π vanishes on X implies that $\Pi\left(c - \alpha 1_{\Omega\times\{0\}}\right) = 0$. The linearity of Π implies that $\Pi(c) = \alpha\Pi(1_{\Omega\times\{0\}})$. Finally, the normalization assumption results in $\Pi(c) = \alpha$. Another way of stating this conclusion is that if the contract (δ, V) is traded at time zero and Π is a present-value function, then $V_0 = \Pi(\delta)$.

If the market X is arbitrage-free and complete, letting $\Pi(c)$ equal the (unique) present value of c defines a function $\Pi : \mathcal{L} \to \mathbb{R}$ that is easily confirmed to be a present-value function. Therefore, every complete arbitrage-free market admits a unique present-value function. Conversely, every present-value function Π for X defines a unique complete market for which Π is a present value function; it is the kernel of Π, that is, the set $\{c \in \mathcal{L} \mid \Pi(c) = 0\}$. If X is complete, then the kernel of Π is equal to X. If X is incomplete, then the kernel of Π is a proper superset of X and can be thought of as an arbitrage-free completion of X.

The relationship between the market X and a present-value function Π has a geometric interpretation in the space $\mathbb{R}^{1+N}$, where $1+N$ is the total number of spots on the information tree. Recall that the set $\mathcal{L}$ of adapted processes can be identified with $\mathbb{R}^{1+N}$, since an adapted process x is an assignment of a scalar $x(F,t)$ to every spot (F,t). Using this identification, we endow $\mathcal{L}$ with the usual Euclidean inner product

in $\mathbb{R}^{1+N}$, denoted

$$x \cdot y \equiv \sum_{\text{all spots } (F,t)} x(F,t)\, y(F,t).$$

DEFINITION 1.4.2. A **state-price process** is any adapted process p with $p_0 > 0$ such that

$$\Pi(c) \equiv \frac{1}{p_0} p \cdot c, \quad c \in \mathcal{L},$$

defines a present-value function Π. In this case, we say that p **represents** Π. An **Arrow cash flow**[2] is a cash flow of the form $1_{F\times\{t\}}$, where (F,t) is a spot.

Every cash flow is a linear combination of the $1+N$ Arrow cash flows, reflecting the identification of $\mathcal{L}$ and $\mathbb{R}^{1+N}$. If Π is a present-value function and we define $p(F,t) = \Pi(1_{F\times\{t\}})$ for every spot (F,t), then $\Pi(c) = p \cdot c$ for all $c \in \mathcal{L}$ (why?). Therefore, every present-value function can be represented by a state-price process. (In mathematical terms, the Arrow cash flows are a linear basis of $\mathcal{L}$ and p is the Riesz representation of Π.) If p is a state-price process representing the present-value function Π, then $p(F,t) = p_0\Pi(1_{F\times\{t\}})$ for every spot (F,t), and therefore p is strictly positive, it is uniquely determined by Π up to a positive scaling factor, and it represents relative present value of Arrow cash flows: For all spots (F,t), (G,s),

$$\frac{p(F,t)}{p(G,s)} = \frac{\Pi(1_{F\times\{t\}})}{\Pi(1_{G\times\{s\}})}.$$

For a geometric interpretation of a state-price process p representing Π, note that, since $\Pi(x) = 0$ for all $x \in X$, p is orthogonal to X, and since Π is positive, p lies in $\mathbb{R}^{1+N}_{++}$ (the interior of the positive orthant of $\mathbb{R}^{1+N}$). The set of all state-price processes representing Π can be identified with a directed half line in $\mathbb{R}^{1+N}_{++}$. Inspection of Figure 1.4.1 suggests that such a direction exists if and only if

$$X \cap \mathbb{R}^{1+N}_{+} = \{0\},$$

which is another way of saying that X is arbitrage-free. If, as in Figure 1.4.1, X is complete and therefore N-dimensional, there is only one orthogonal-to-X direction within $\mathbb{R}^{1+N}$, and therefore the corresponding present-value function must be unique. If X is incomplete, the dimension of X is at most $N-1$, leaving at least one dimension

[2]The term, which is not entirely standard, is in recognition of Arrow [1953] (translated to English in Arrow [1963]), who together with Debreu [1959] provided the modern conceptual framework for incorporating uncertainty in classical competitive equilibrium theory. The more common term "Arrow-Debreu security" refers to a claim to an Arrow cash flow. The (relative) time-zero prices of Arrow-Debreu securities are also known as Arrow-Debreu prices, corresponding to the notion of a state-price process here.

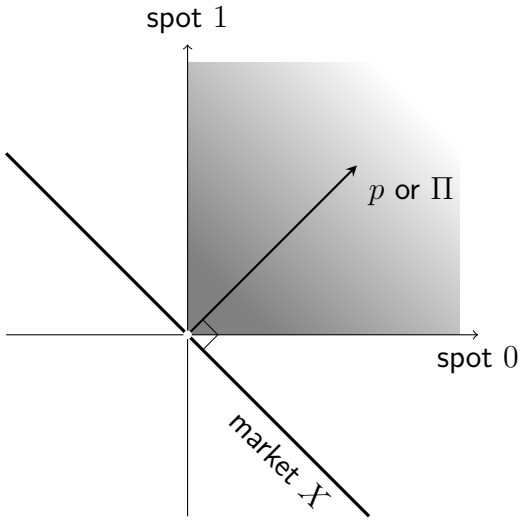


FIGURE 1.4.1. Example of a market and present-value function with $N = 2$. The shaded region (including the axes but not the origin) is the set of arbitrage cash flows. The market X does not cut into the shaded region if and only if the half line defined by the orthogonal vector p lies in the interior of the shaded region.

along which p can be rotated within $\mathbb{R}^{1+N}_{++}$ while maintaining orthogonality to X, which suggests that an incomplete market admits multiple present-value functions. For an example, extend Figure 1.4.1 by introducing a spot-2 orthogonal axis, while maintaining the market X as a line. Rotating the orthogonal-to-X vector p around X while staying within the interior of the positive orthant traces out all present-value functions relative to X. A rigorous proof of these informal insights follows.[3]

THEOREM 1.4.3. *A present-value function exists if and only if X is arbitrage-free; and is unique if and only if the market X is complete.*

PROOF. (Existence) If Π is a present-value function, then $\Pi(x) \leq 0$ for all $x \in X$. If c is an arbitrage, then $\Pi(c) > 0$ and therefore c cannot

[3]The equivalence of the lack of arbitrage opportunities and the existence of a present-value function is an example of the so-called theorems of the alternative in convex analysis, exposited in Chapter 1 of Stoer and Witzgall [1970]. In the current finite-dimensional financial market context, the result is due to Ross [1978]. It was extended to infinite-dimensional spaces by Yan [1980] and Kreps [1981]. In general, under infinitely many states, the result requires the exclusion of a stronger notion of arbitrage opportunities than mere positivity. A notable exception is the case of a market generated by finitely many assets in discrete time, as shown by Dalang et al. [1990] (with simplified proofs given by Schachermayer [1992] and Kabanov and Kramkov [1994]). Extension to models with continuous-time trading are reviewed by Delbaen and Schachermayer [2006].

be in X. For the converse, we will show that a state-price process exists assuming only that X is a closed convex cone that contains no arbitrage cash flow, a generality that will be useful later on, as discussed in Remark 1.4.4 below. We fix an enumeration $0, 1, \ldots, N$ of all spots, where $(\Omega, 0)$ is spot zero, and we write $x = (x_0, x_1, \ldots, x_N)$ for the element of $\mathbb{R}^{1+N}$ corresponding to the adapted process x. (The conflicting notation x_t, where t is a time, is not used in this proof.) We use the Euclidean inner product on $\mathcal{L}$, which takes the familiar form $x \cdot y = \sum_{n=0}^{N} x_n y_n$. A state-prices process p satisfies $p \cdot x \leq 0$ for all $x \in X$ and $p \cdot c > 0$ for every arbitrage c, and therefore separates the convex sets X and $\mathbb{R}^{1+N}_{+} \setminus \{0\}$. This suggests that the existence of p follows from the separating hyperplane theorem, except for the subtlety that $p \cdot c > 0$ for every arbitrage c only implies that $0 \neq p \geq 0$, not necessarily that $p \in \mathbb{R}^{1+N}_{++}$, as required of a state-price process. To overcome this difficulty, we instead separate the compact convex set $\Delta \equiv \{x \in \mathbb{R}^{1+N}_{+} \mid \sum_n x_n = 1\}$ from X. Since X is arbitrage-free, $X \cap \Delta = \emptyset$. By the Projection Theorem B.4.1, the function that maps each point of Δ to its projection on X is continuous. Since norms are continuous, the function that maps each point of Δ to its least distance from X is also continuous (as well as convex) and therefore achieves a minimum over Δ by Proposition B.2.6 (or Proposition B.3.3). There is, therefore, a pair $(\bar{x}, \bar{y}) \in X \times \Delta$ such that $\bar{x}$ is the projection of $\bar{y}$ on X and $\bar{y}$ is the projection of $\bar{x}$ on Δ. By the Projection Theorem B.4.1, $-p \equiv \bar{x} - \bar{y}$ supports X at $\bar{x}$ and therefore $p \cdot x \leq p \cdot \bar{x}$ for all $x \in X$. Since X is a cone, $p \cdot \bar{x} = 0$ and $p \cdot x \leq 0$ for all $x \in X$. Similarly, $p \equiv \bar{y} - \bar{x}$ supports Δ at $\bar{y}$, and therefore $p \cdot y \geq p \cdot \bar{y}$ for all $y \in \Delta$. Since $p \cdot \bar{x} = 0$, $p \cdot \bar{y} = p \cdot p > 0$. Therefore, $p \cdot y > 0$ for all $y \in \Delta$, which implies that p is strictly positive. A corresponding present-value function is defined by $\Pi(c) = p \cdot c$.

(Uniqueness) Suppose Π is a present-value function. Let M denote the set of marketed cash flows, spanned by the elements of X and the cash flow $1_{\Omega \times \{0\}}$, which we identify with the vector $\mathbf{1}^0 \equiv (1, 0, \ldots, 0)$ in $\mathbb{R}^{1+N}$. We have already shown that $\Pi(c)$ is uniquely determined for all $c \in M$. If X is complete, then $M = \mathbb{R}^{1+N}$ and Π is uniquely determined. Suppose instead that X is incomplete, and fix any $c \in \mathbb{R}^{1+N} \setminus M$. Taking an orthogonal projection, let $c = m + z$, where $m \in M$ and $z \neq 0$ is orthogonal to M and therefore also to X and $\mathbf{1}^0$. Let $p \in \mathbb{R}^{1+N}_{++}$ be the Riesz representation of Π, that is, $\Pi(c) = p \cdot c$ for all $c \in \mathbb{R}^{1+N}$. Consider any scalar $\alpha \neq 0$ such that $p^\alpha \equiv p + \alpha z \in \mathbb{R}^{1+N}_{++}$. Since z is orthogonal to X and $\mathbf{1}^0$, $p^\alpha \cdot x = 0$ for all $x \in X$ and $p^\alpha \cdot \mathbf{1}^0 = p^0 \cdot \mathbf{1}^0 = 1$. Therefore, $\Pi^\alpha(c) = p^\alpha \cdot c$ defines a continuum of distinct present-value functions as α ranges over all scalars such that $p^\alpha \in \mathbb{R}^{1+N}_{++}$. □

REMARK 1.4.4. The preceding proof shows the existence of a present-value function assuming only that X is a closed convex cone, not necessarily a linear subspace. This generality is utilized in our later discussion of dominant choice. It also allows an easy extension of Theorem 1.4.3 to a **constrained market** X, which we define as a closed convex set of cash flows such that $0 \in X$ and there exists $\varepsilon > 0$ such that for all $x \in X$, $0 < \|x\| < \varepsilon$ implies $(\varepsilon/\|x\|)\, x \in X$. In words, every trade that is small in the sense that its norm is no more than ε can be scaled up so that its norm is ε. This allows for larger position limits as well as short-sale constraints (since $x \in X$ need not imply $-x \in X$). Let us show that a *constrained market X is arbitrage-free if and only if a present-value function exists.* First note that X is arbitrage-free if and only if the cone $C \equiv \{sx \mid s \in \mathbb{R}_+, x \in X\}$ generated by X is arbitrage-free. Since C is a closed convex cone, the proof of Theorem 1.4.3 applies. That C is convex is immediate. Closure of C follows from the assumption that trades smaller than ε can be scaled up to have norm ε. The interested reader can prove this by confirming that $C \cap X = C \cap \left\{x \in \mathbb{R}^{1+N} \mid \|x\| \leq \varepsilon\right\}$. $\diamondsuit$

We have so far defined present-value functions from the perspective of time zero. A present-value function $\Pi_{F,t} : \mathcal{L}_{F,t} \to \mathbb{R}$ can be defined analogously from the perspective of every other spot (F,t). Suppose $X_{F,t} \subseteq \mathcal{L}_{F,t}$ is the set of traded cash flows from the perspective of spot (F,t), and the dynamic consistency assumption $X_{F,t} \subseteq X$ is satisfied (as in Proposition 1.2.5). Then the restriction of a time-zero present-value function Π to $\mathcal{L}_{F,t}$ is a positive linear functional on $\mathcal{L}_{F,t}$ satisfying $\Pi(x) \leq 0$ for all $x \in X_{F,t}$. After normalization so that $\Pi_{F,t}(1_{F\times\{t\}}) = 1$, we arrive to the following definition.

DEFINITION 1.4.5. The spot-(F,t) present-value function **induced** by the time-zero present value function Π is the function

$$\Pi_{F,t}(c) = \frac{\Pi\left(c1_{F\times\{t,\ldots,T\}}\right)}{\Pi\left(1_{F\times\{t\}}\right)}, \quad c \in \mathcal{L}. \tag{1.4.1}$$

Note that we have defined $\Pi_{F,t}$ over the entire domain $\mathcal{L}$, while only its restriction $\Pi_{F,t} : \mathcal{L}_{F,t} \to \mathbb{R}$ is meaningful as a spot-(F,t) present-value function. This is merely a convenience. The following proposition shows that for an arbitrage-free liquid market the preceding definition covers all spot-(F,t) present-value functions.

PROPOSITION 1.4.6. *Suppose the market X is arbitrage-free and liquid and $\Pi_{F,t}$ is a positive linear functional on $\mathcal{L}_{F,t}$ such that $\Pi_{F,t}(x) \leq 0$ for every $x \in X \cap \mathcal{L}_{F,t}$ and $\Pi_{F,t}(1_{F\times\{t\}}) = 1$. Then $\Pi_{F,t}$ is induced by some time-zero present value function.*

PROOF. Fix any time-zero present-value function $\tilde{\Pi}$ and define the function $\Pi : \mathcal{L} \to \mathbb{R}$ by letting $\Pi(c) = \tilde{\Pi}(\tilde{c})$, where $\tilde{c}$ is the cash flow

obtained from c after replacing the restriction of c on $F \times \{t, \ldots, T\}$ with a single payment at spot (F,t) equal to $\Pi_{F,t}(c)$, that is, $\tilde{c} \equiv c - c1_{F\times\{t,\ldots,T\}} + \Pi_{F,t}(c)\, 1_{F\times\{t\}}$. Then Π is a positive linear functional that satisfies $\Pi(1_{\Omega\times\{0\}}) = 1$. We now show that Π vanishes on X and is therefore a present-value function that induces $\Pi_{F,t}$. Given any $x \in X$, the liquidity assumption allows us to find $y \in X \cap \mathcal{L}_{F,t}$ such that $(x-y)\, 1_{F\times\{t+1,\ldots,T\}} = 0$ and therefore $\Pi_{F,t}(x-y)\, 1_{F\times\{t\}} = (x-y)\, 1_{F\times\{t,\ldots,T\}}$, which in turn implies $\Pi(x-y) = \tilde{\Pi}(x-y) = 0$. Moreover, since $y \in \mathcal{L}_{F,t}$ and $\Pi_{F,t}(y) = 0$, we also have $\Pi(y) = 0$. Therefore, $\Pi(x) = \Pi(x-y) + \Pi(y) = 0$. □

Unless otherwise indicated, our focus is on liquid markets and the notation $\Pi_{F,t}$ will refer to the conditioned version of Π, as defined by equation (1.4.1).

We conclude this section with some simple but essential observations on the relationship between present-value functions and contract pricing.

Definition 1.4.7. The positive linear functional $\Pi : \mathcal{L} \to \mathbb{R}$ is said to **price** the contract (δ, V) if for every spot (F,t),

$$V(F,t) = \Pi_{F,t}(\delta), \tag{1.4.2}$$

where $\Pi_{F,t}$ denotes the spot-(F,t) present-value function induced by Π.

Proposition 1.4.8. *(a) In every market, all present-value functions price all traded contracts.*

(b) A positive linear functional on $\mathcal{L}$ is a present-value function for the market implemented by a set of contracts if and only if it prices every contract in the given set.

Proof. (a) Set to zero the present value of the cash flow (1.3.3) resulting from buying a traded contract at a given spot.

(b) Let X^0 be defined as in last section's last paragraph. A positive linear functional vanishes on $\mathrm{span}(X^0)$ if and only if it vanishes on X^0 if and only if it prices every contract defining X^0. □

1.5. Options and dominant choice

An option generalizes our earlier notion of a contract by allowing the owner to choose any dividend process within a specified set of cash flows O. For example, suppose O represents the possible cash flows resulting from a finite opportunity set of projects available to a firm. Shareholder owners of the firm may disagree on what is the best project to undertake. Assuming shareholders only care about the cash flow generated by a project and every cash flow in O can be sold in a complete market, then shareholders who would disagree absent a market can agree on a project that is assigned the highest price by the market. Extending this idea, we will see that in the presence of a sufficiently

complete arbitrage-free market, a present-value maximizing choice is dominant in that it is optimal given the market for every shareholder. Moreover, assuming the market is liquid, a choice that is dominant at time zero remains dominant as uncertainty unfolds and therefore there is no incentive to deviate from an initially selected dominant choice. The existence of a dominant choice in turn leads to the pricing of a traded option. Another important application of these arguments is to standardized traded financial options such as American call and put options.

We assume throughout that a market X is available and that O is a nonempty set of cash flows.

DEFINITION 1.5.1. A cash flow $\delta^* \in O$ is **dominant** (in O given X) if for every $\delta \in O$, there exists some $x \in X$ such that $\delta^* + x \geq \delta$.

A dominant choice is optimal for every agent who is not averse to additional income at some spot (or can freely dispose of income). Given any set of agents I, suppose agent $i \in I$ finds δ^i optimal in O. If $\delta^* \in O$ is dominant, then there exist trades $x^i \in X$ such that $\delta^* + x^i \geq \delta^i$. Agent i is therefore at least as well off selecting δ^* instead of δ^i and at the same time entering the trade x^i. In this sense, all agents in I agree on the optimality of δ^*, even though the way they use the market to transform δ^* can differ.

THEOREM 1.5.2. *Suppose the market X is arbitrage-free and O is a nonempty set of cash flows. Then $\delta^* \in O$ is dominant in O if and only if $\Pi(\delta^*) = \max\{\Pi(\delta) \mid \delta \in O\}$ for every present-value function Π for X.*

PROOF. Suppose δ^* is dominant. For any $\delta \in O$, we can write $\delta^* + x \geq \delta$ for some $x \in X$, and therefore $\Pi(\delta^*) = \Pi(\delta^* + x) \geq \Pi(\delta)$ for every present-value function Π.

For the converse, it is instructive, although formally redundant, to first consider a simple argument for a complete market X. Suppose $\delta^* \in O$ maximizes the unique present-value function Π over O. Since X is complete, $\delta^* = \Pi(\delta^*)\mathbf{1}^0 + y^*$ and $\delta = \Pi(\delta)\mathbf{1}^0 + y$ for some $y^*, y \in X$. Letting $x = y - y^* \in X$, we have $\delta^* + x = \Pi(\delta^*)\mathbf{1}^0 + y \geq \Pi(\delta)\mathbf{1}^0 + y = \delta$, which proves the dominance of δ^*.

More generally, suppose X is arbitrage-free but not necessarily complete and $\delta^* \in O$ is *not* dominant. We will show that there exists some present-value function that δ^* does *not* maximize. As in the proof of Theorem 1.4.3, we identify $\mathcal{L}$ with $\mathbb{R}^{1+N}$. Since δ^* is not dominant, there exists $\delta \in O$ such that $\delta^* - \delta + x \notin \mathbb{R}^{1+N}_+$ for all $x \in X$. With $x^* \equiv \delta^* - \delta$, the set $X^* = \{x + \alpha x^* \mid x \in X,\ \alpha \in \mathbb{R}_+\}$ is a closed convex cone that contains no arbitrage. By Remark 1.4.4, there exists $p \in \mathbb{R}^{1+N}_{++}$ such that $p \cdot x \leq 0$ for all $x \in X^*$. Such a vector p is a state-price process that satisfies $p \cdot x^* \leq 0$. By construction, $x^* \notin X$.

Let $x^* = \bar{x} + z$, where $\bar{x} \in X$ and z is nonzero and orthogonal to X. Pick $\varepsilon > 0$ small enough so that $p^\varepsilon \equiv p - \varepsilon z$ is a state-price process such that $p^\varepsilon \cdot x^* = -\varepsilon z \cdot z < 0$. The present-value function Π defined by p^ε satisfies $\Pi(x^*) < 0$ and therefore $\Pi(\delta^*) < \Pi(\delta)$. □

COROLLARY 1.5.3. *Suppose the market is arbitrage-free, the set of cash flows O is compact (for example, finite), and every element of O is marketed. Then a dominant choice in O exists.*

If the option contains non-marketed cash flows, there may not be a dominant choice, even if O is compact. For a trivial example, suppose $T = 1$, there is no uncertainty, $O = \{(0,1),(1,0)\}$ and $X = \{0\}$. Neither element of O dominates the other. Every positive linear functional is a present-value function for the trivial market $X = \{0\}$, and each element of O maximizes some present value.

The preceding theorem characterizes dominance from the perspective of time zero. Suppose now that the option holder does not have to commit to an initial selection. Will there be an incentive to deviate from a previously made dominant selection in the face of new information? To address this issue we refine the definition of an option to incorporate the idea that the option holder can deviate from a previous cash flow choice at any time. As before, let O represent a set of cash flows available to the option holder at time zero. For every cash flow δ and spot (F,t), the set of cash flows in O that are equal to δ up to but not including spot (F,t) is

$$O_{F,t}(\delta) \equiv \left\{ \tilde{\delta} \in O \mid \tilde{\delta} = \delta \text{ on } F \times \{0,\dots,t-1\} \right\}, \tag{1.5.1}$$

Selecting $\delta \in O_{\Omega,0} = O$ at time zero and switching to $\tilde{\delta} \in O_{F,t}(\delta)$ at spot (F,t) is equivalent to selecting $\delta + (\tilde{\delta} - \delta)1_{F \times \{t,\dots,T\}}$ at time zero, which must therefore also belong to O if the option holder does not have to commit at time zero. This motivates the following formal definition.

DEFINITION 1.5.4. An **option** is a set $O \subseteq \mathcal{L}$ such that for every spot (F,t), $\delta \in O$ and $\tilde{\delta} \in O_{F,t}(\delta)$ implies $\delta + (\tilde{\delta} - \delta)1_{F \times \{t,\dots,T\}} \in O$.

EXAMPLE 1.5.5. (American call) An American call is the right but not the obligation to receive a specified traded security in exchange for a specified cash amount, known as the strike, at most once any time up to a given maturity date. We can formally view an American call as a special case of an option in the sense of Definition 1.5.4. Given a **maturity** $\bar{\tau} \in \{1,\dots,T-1\}$, let $\mathcal{T}$ denote the set of all stopping times that are valued in $\{0,\dots,\bar{\tau}\} \cup \{\infty\}$. For $\tau \in \mathcal{T}$, let $1_{[\tau]}$ denote the process whose value at (ω,t) is one if $\tau(\omega) = t$ and zero otherwise. An **American option** with payoff process $D \in \mathcal{L}$ is the option

$$O = \left\{ D1_{[\tau]} \mid \tau \in \mathcal{T} \right\}, \quad \text{where } D_\infty \equiv 0.$$

Selecting cash flow $D1_{[\tau]}$ is referred to as **exercising** the American option at time τ. An **American call** on some underlying traded contract with ex-dividend processes S is the American option with payoff process $D = S - K$, where the scalar K is the **strike.**

Suppose that the underlying contract, which we henceforth refer to as the **stock**, does *not* pay any dividends up to and including the maturity date $\bar{\tau}$. Suppose also that for every time prior to $\bar{\tau}$ there is a way to **save** K (units of account) and have K for sure at time $\bar{\tau}$ plus potentially some non-negative interest. (Section 1.7 defines more formally a money-market account that can implement such savings, provided the interest rate process is non-negative.) In this case, it is a dominant choice to *not* exercise the American call prior to maturity. To see why, suppose the option holder is considering exercising the option at some spot (F, t) with $t < \bar{\tau}$. The option holder is at least as well off keeping the option alive to maturity, shorting the stock and saving K. Exercising at spot (F, t) results in an inflow $S(F, t) - K$ and nothing thereafter. The alternative strategy also results in an inflow $S(F, t) - K$ at spot (F, t), but now the option holder still has the call option, a saved amount that is sufficient to pay the strike at maturity, and a short position on the underlying stock. At maturity, if the stock is worth more than K, the option holder can spend K from savings, exercise the call option and close out the short position. If on the other hand the stock price is less than K at maturity, the option holder ends up with $K - S_{\bar{\tau}}$ plus any additional interest on savings. Put together, relative to exercising at spot (F, t), the alternative strategy results in the additional time-$\bar{\tau}$ payoff of $(K - S_{\bar{\tau}})^+$ (which is the payoff of a European put option) plus any interest on the strike.

The argument fails if the stock pays sufficiently high dividends, since the benefit of collecting dividends can outweigh the costs of early exercise. This tradeoff between immediate payoff and the benefit of weighting in order to condition actions on future information, as well as possibly collect interest on costs associated with exercising the option, captures an essential intuition behind the optimal exercise of American options. We return to this example in the following chapter using a dual approach based on present-value maximization and Jensen's inequality. $\diamondsuit$

Fixing a reference option O and an arbitrage-free and liquid (but possibly incomplete) market X, we now extend the notion of dominance to apply from the perspective of any given spot (F, t), where it is implicit that dominance is relative to the spot-(F, t) market $X_{F,t} \equiv X \cap \mathcal{L}_{F,t}$.

Definition 1.5.6. The cash flow δ in O is **dominant at spot** (F, t) if for all $\tilde{\delta}$ in $O_{F,t}(\delta)$, there exists some $x \in X_{F,t}$ such that $\delta + x \geq \tilde{\delta}$ on $F \times \{t, \ldots, T\}$.

Theorem 1.5.7 can now be extended as follows. Recall that every (time-zero) present value function Π induces a spot-(F,t) present value function $\Pi_{F,t}$, defined by (1.4.1).

THEOREM 1.5.7. *Given an arbitrage-free liquid market, the following conditions are equivalent, for all $\delta^* \in O$.*

(1) *δ^* is dominant at every spot.*
(2) *δ^* is dominant (at spot zero).*
(3) *$\Pi(\delta^*) = \max_{\delta \in O} \Pi(\delta)$ for every present-value function Π.*
(4) *For every present-value function Π and spot (F,t),*

$$\Pi_{F,t}(\delta^*) = \max\left\{\Pi_{F,t}(\tilde{\delta}) \mid \tilde{\delta} \in O_{F,t}(\delta)\right\}. \tag{1.5.2}$$

PROOF. (1 $\implies$ 2) If δ^* is dominant, it is dominant at spot $(\Omega, 0)$.
(2 $\iff$ 3) See Theorem 1.5.2.
(3 $\implies$ 4) Suppose that $\Pi_{F,t}(\delta) > \Pi_{F,t}(\delta^*)$ for some present-value function Π, spot (F,t) and $\delta \in O_{F,t}(\delta^*)$. Let $\tilde{\delta} \equiv \delta^* + (\delta - \delta^*)\, 1_{F \times \{t,\ldots,T\}} \in O$. Using equation (1.4.1) with $c = \tilde{\delta} - \delta^*$, we conclude that $\Pi(\tilde{\delta}) > \Pi(\delta^*)$.
(4 $\implies$ 1) Since the market is assumed liquid, Proposition 1.4.6 implies that every spot-(F,t) present-value function $\Pi_{F,t}$ can be computed by (1.4.1) in terms of some present-value function Π. Therefore δ^* is dominant at (F,t) by the argument of Theorem 1.5.2 applied to $X_{F,t}$. □

Suppose now that an option to be bought or sold. The option buyer pays the option price, commonly referred to as the **premium**, to the option seller (or "writer"), who is then obligated to deliver the cash flow selected by the buyer. Assuming the existence of a marketed dominant choice at time zero, as well as a dominant choice at every other spot given any exercise history, we will show that the premium that is consistent with the absence of arbitrage opportunities is the present value of a dominant cash flow. The only subtlety in this argument is that a potential arbitrageur that writes an option must be able to hedge potentially suboptimal choices by the option buyer.

We continue taking as given the arbitrage-free market X and an option O with time-zero[4] premium p. We analyze the buying and selling of an option separately in order to account for the fact that only the owner decides how to exercise the option. If p is sufficiently low, then an arbitrage can be created by buying the option and selecting a dominant cash flow. A buyer of the option pays the premium p, can select any cash flow δ in O and can trade in the market X, thus

[4]To keep the notation simple, we consider the trading of the option O at time zero only. For every $\delta \in O$, the trading of the option $O_{F,t}(\delta)$ at spot (F,t) can be analyzed by applying the same arguments on the subtree rooted at (F,t).

generating a cash flow of the form

$$-p1_{\Omega\times\{0\}}+\delta+x,\quad \delta\in O,\quad x\in X. \tag{1.5.3}$$

Excluding such an arbitrage implies a lower bound on the option premium.

PROPOSITION 1.5.8. *Suppose the market is arbitrage-free and δ^* is a dominant cash flow in O that is marketed with present value p^*. Then $p\geq p^*$ if and only if there is no arbitrage of the form (1.5.3).*

PROOF. Choose $x^*\in X$ such that $\delta^*=p^*1_{\Omega\times\{0\}}+x^*$. If $p<p^*$, then the cash flow $-p1_{\Omega\times\{0\}}+\delta^*-x^*=(p^*-p)\,1_{\Omega\times\{0\}}$ is an arbitrage. Conversely, suppose $p\geq p^*$ and let $c\equiv -p1_{\Omega\times\{0\}}+\delta+x$ for any $\delta\in O$ and $x\in X$. The dominance of δ^* implies that $\delta^*+y\geq\delta$ for some $y\in X$. Therefore,

$$c\leq -p^*1_{\Omega\times\{0\}}+(\delta^*+y)+x=x^*+x+y\in X.$$

Since X is arbitrage-free, c is not an arbitrage. □

If the option premium is higher than the present value p^* of the dominant choice, one may wish to sell the option towards an arbitrage. A potential difficulty in this case is that the option buyer cannot be assumed to select the dominant cash flow δ^*. Given a slightly stronger version of the assumption of Proposition 1.5.8, however, we can still show that if $p>p^*$ then an arbitrage is possible that involves writing the option and hedging all possible choices by the option buyer. A key aspect of the arbitrageur's hedging strategy is that it can be implemented without knowledge of the option holder's future choices. We formalize this type of informational restriction as follows.

DEFINITION 1.5.9. An **O-adapted strategy** is a mapping h that assigns to each nonterminal spot (F,t) a function $h_{F,t}:O\to X_{F,t}$ such that $\delta=\tilde{\delta}$ on $F\times\{0,\dots t\}$ implies $h_{F,t}(\delta)=h_{F,t}(\tilde{\delta})$.

Given an O-adapted strategy h, we think of $h_{F,t}(\delta)$ as the incremental trade that the option writer must enter at spot (F,t) in order to hedge all possible future choices by the option buyer who has selected δ up to and including spot (F,t). To simplify the notation, for every $\delta\in O$ and time $t<T$, let the function $h_t(\delta):\Omega\to X$ be defined by

$$h_t(\delta)(\omega)\equiv h_{F,t}(\delta)\quad \text{for all } \omega\in F. \tag{1.5.4}$$

An option seller who receives the premium p at time zero must deliver whatever cash flow δ in O is selected by the option buyer. If the option seller follows the hedging strategy h, the resulting cash flow is

$$p1_{\Omega\times\{0\}}-\delta+\sum\nolimits_{t=0}^{T-1}h_t(\delta). \tag{1.5.5}$$

Excluding every arbitrage of this type implies an upper bound for the option premium, which together with the lower bound of Proposition 1.5.8 pins down the option premium as the present value of a dominant cash flow.

PROPOSITION 1.5.10. *Suppose the market is arbitrage-free and δ^* is a dominant cash flow in O that is marketed with present value p^*. Suppose further that for every $\delta \in O$ and spot (F,t), there exists a cash flow in $O_{F,t}(\delta)$ that is dominant at (F,t). Then $p \leq p^*$ if and only if there exists no O-adapted strategy h such that the cash flow (1.5.5) is an arbitrage for every $\delta \in O$.*

PROOF. The "only if" part can be shown similarly to the corresponding part of Proposition 1.5.8. Conversely, we assume $p > p^*$ and show the existence of an O-adapted strategy h such that the cash flow (1.5.5) is an arbitrage for all $\delta \in O$. Given $\delta \in O$, select $\delta^{F,t} \in O_{F,t}(\delta)$ to be dominant at spot (F,t) and let $\delta^t \equiv \sum_{i=1}^n \delta^{F_i,t} 1_{F_i \times \{0,\ldots,T\}}$, where $\{F_1,\ldots,F_n\}$ is the partition generating $\mathcal{F}_t$. The definition of an option implies that $\delta^t \in O_{F,t}(\delta)$ for every spot (F,t). Using the dominance of δ^*, select $x \in X$ such that $\delta^* + x \geq \delta^1$ and let $h_0(\delta) \equiv -p^* 1_{\Omega\times\{0\}} + \delta^* + x$. The value $h_0(\delta)$ depends on δ only through the value δ_0, since the choice of x has this property. At time zero the arbitrageur sells the option, receiving the premium p, buys the cash flow δ^* for a price p^* and enters the trade x, which together with δ^* dominates δ^1. After paying $\delta_0 = \delta_0^1$ to the option buyer, the arbitrageur faces the cash flow

$$c^0 \equiv p 1_{\Omega\times\{0\}} + h_0(\delta) - \delta 1_{\Omega\times\{0\}} \geq (p - p^*) 1_{\Omega\times\{0\}} + \delta^1 1_{\Omega\times\{1,\ldots,T\}}.$$

Proceeding inductively, suppose that after all transactions prior to time $t \in \{1,\ldots,T-1\}$, the arbitrageur faces an overall cash flow

$$c^{t-1} \geq (p - p^*) 1_{\Omega\times\{0\}} + \delta^t 1_{\Omega\times\{t,\ldots,T\}}.$$

Using the dominance of $\delta^{F,t}$ at spot (F,t), choose $h_{F,t}(\delta) \in X_{F,t}$ so that

$$\delta^t 1_{F\times\{t,\ldots..T\}} + h_{F,t}(\delta) \geq \delta^{t+1} 1_{F\times\{t,\ldots..T\}}.$$

The arbitrageur can choose $\delta^{t+1} 1_{F\times\{t,\ldots,T\}}$, and therefore $h_{F,t}(\delta)$, having observed only $\delta 1_{F\times\{0,\ldots,t\}}$. At time t the arbitrageur enters the trade $h_t(\delta)$, defined in (1.5.4), and pays out δ_t to the option holder, resulting in the new cash flow

$$c^t \equiv c^{t-1} + h_t(\delta) - \delta 1_{\Omega\times\{t\}} \geq (p - p^*) 1_{\Omega\times\{0\}} + \delta^{t+1} 1_{\Omega\times\{t+1,\ldots,T\}}.$$

At time T the arbitrageur pays out δ_T to the option holder, resulting in the overall arbitrage cash flow $c^{T-1} - \delta 1_{\Omega\times\{T\}} \geq (p - p^*) 1_{\Omega\times\{0\}}$. The recursive construction of the cash flows c^t implies that $c^{T-1} - \delta 1_{\Omega\times\{T\}}$ is equal to (1.5.5), and we have therefore produced an O-adapted strategy h such that (1.5.5) defines an arbitrage for all $\delta \in O$. □

1.6. Trading strategies

As we saw in Section 1.2, the market implemented by a given set of contracts can be described as the set of cash flows generated by trading strategies. A trading strategy essentially specifies a contingent trade for every spot of the information tree. This section provides a more systematic discussion of trading strategies and associated notation and terminology, which is tailored to the application of probabilistic methods introduced in the following chapter.

Throughout this section, we take as given the contracts

$$\left(\delta^1, V^1\right), \ldots, \left(\delta^J, V^J\right), \tag{1.6.1}$$

with corresponding ex-dividend price processes $S^j = V^j - \delta^j$. Trading strategies in these contracts are time-indexed sequences of portfolios representing positions over each period. **Period** $t \in \{1, \ldots, T\}$ is the time interval whose beginning is time $t-1$ and whose end is time t. A period-t portfolio $\left(\theta_t^1, \ldots, \theta_t^J\right)$ is formed at the beginning of period t, where θ_t^j represents a number of shares in contract j. By convention, we set $\theta_0^j = 0$. In general, it is important to keep track of what quantities are known at the beginning of the period and what quantities are only known at the end of the period. To emphasize this distinction, we define a process x to be **predictable** if x_t is $\mathcal{F}_{t-1}$-measurable, for every time $t > 0$, and x_0 is constant. Thus if x is adapted, we can only claim that x_t is known at the end of period t, while if x is predictable, we know that x_t is revealed at the beginning of period t. We let $\mathcal{P}$ denote the set of all predictable processes (relative to the given underlying filtration) and we set

$$\mathcal{P}_0 \equiv \{x \in \mathcal{P} \mid x_0 = 0\}.$$

The sequence $(\theta_0^j, \theta_1^j, \ldots, \theta_T^j)$ of positions in contract j specified by a trading strategy defines an element of $\mathcal{P}_0$.

Definition 1.6.1. A **trading strategy** in the contracts (1.6.1) is a J-dimensional row vector whose entries are elements of $\mathcal{P}_0$. The trading strategy $\left(\theta^1, \ldots, \theta^J\right)$ **generates** the cash flow x, where

$$x_t = \sum\nolimits_j \theta_t^j V_t^j - \theta_{t+1}^j S_t^j, \quad t < T; \quad x_T = \sum\nolimits_j \theta_T^j V_T^j. \tag{1.6.2}$$

In the **budget equation** (1.6.2), one can think of x_t as the net of the time-t value $\sum_j \theta_t^j V_t^j$ of the period-t portfolio and the time-t cost $\sum_j \theta_{t+1}^j S_t^j$ of forming the period-$(t+1)$ portfolio. Since $\theta_0 = 0$, the budget equation implies that $x_0 = -\sum_j \theta_1^j S_0^j$, which is the initial payment required to start the strategy. The final payment x_T equals the portfolio's time-T liquidation value.

In Section 1.2 we defined the set implemented by the contracts (1.6.1) as the smallest market in which every (δ^j, V^j) is traded.

PROPOSITION 1.6.2. *The market implemented by the contracts* (1.6.1) *is the set of all cash flows generated by all trading strategies in these contracts.*

PROOF. Let X be the market implemented by the contracts (1.6.1) and let X' be the set of all cash flows generated by all trading strategies in these contracts. We are to show that $X = X'$. In the closing paragraph of Section 1.2 we saw that the $X = \text{span}\,(X^0)$, where X^0 is the set of every $-V^j 1_{F\times\{t\}} + \delta^j 1_{F\times\{t,\ldots,T\}}$, where (F,t) is a spot and $j \in \{0,\ldots,J\}$. One can easily check that every cash flow in X^0 is generated by a traded strategy. Since X' is a linear subspace, this shows that $X \subseteq X'$. For the converse inclusion, we must show that the cash flow x generated by a trading strategy θ is a linear combination of cash flows in X^0. Note that $x = \sum_j x^j$, where $x_t^j = \theta_t^j V_t^j - \theta_{t+1}^j S_t^j$ for $t < T$ and $x_T^j = \theta_T^j V_T^j$. It is therefore, sufficient to show that $x^j \in \text{span}\,(X^0)$ for every j. This can be shown by induction in the number of trades of θ^j, defined as the number of spots where θ^j changes value. The details are left to the interested reader. □

The budget equation (1.6.2) is more conveniently expressed in terms of gain processes, using a process transform operator that directly corresponds to stochastic integrals in continuous-time versions of the theory and facilitates the application of martingale methods introduced in the following chapter. Prior to stating this form of the budget equation, we digress briefly to introduce some useful notation.

For any process x, the **lagged process** x_- is defined by

$$x_-(0) = x(0), \quad x_-(t) = x(t-1), \quad t = 1,\ldots,T.$$

Clearly, the process x is adapted if and only if x_- is predictable. The **increments process** of x is the process $\Delta x = x - x_-$, or equivalently,

$$\Delta x_0 = 0, \quad \Delta x_t = x_t - x_{t-1}, \quad t = 1,\ldots,T.$$

The **integral**[5] $x \bullet y$ of the process x with respect to the process y is the process

$$(x \bullet y)_0 = 0, \quad (x \bullet y)_t = \sum_{s=1}^{t} x_s \Delta y_s, \quad t = 1,\ldots,T.$$

We denote by $\mathbf{t}$ the process that counts time: $\mathbf{t}(t) = t$ for every time t. Therefore, for every process x,

$$(x \bullet \mathbf{t})_0 = 0, \quad (x \bullet \mathbf{t})_t = \sum_{s=1}^{t} x_s, \quad t = 1,\ldots,T.$$

The **gain process** G^j of contract (δ^j, V^j) is defined by

$$G_t^j = V_t^j + \sum_{s<t} \delta_s^j = S_t^j + \sum_{s\le t} \delta_s^j, \quad t = 0,\ldots,T.$$

[5]The bullet notation for an integral is more commonly found in the more advanced literature on stochastic analysis. See, for example, Jacod and Shiryaev [2003].

For times $t > s$, the increment $G^j_t - G^j_s$ represents the total gain (or loss if negative) resulting from purchasing the jth contract at time s and selling it at time t. If $\left(\theta^1, \dots, \theta^J\right)$ is a trading strategy, then $\theta^j \bullet G^j$ represents total gains from trading contract j up to time t.

PROPOSITION 1.6.3. *The trading strategy* $\left(\theta^1, \dots, \theta^J\right)$ *generates the cash flow* x *if and only if*

$$\sum\nolimits_j \theta^j V^j = -x_- \bullet \mathbf{t} + \sum\nolimits_j \theta^j \bullet G^j, \quad \sum\nolimits_j \theta^j_T V^j_T = x_T. \tag{1.6.3}$$

PROOF. Let $W = \sum_j \theta^j V^j$. Since $\Delta G^j_t = V^j_t - S^j_{t-1}$, the budget equation (1.6.2) can be written as

$$\Delta W_t = -x_{t-1} + \sum_j \theta^j_t \Delta G^j_t, \quad t \le T; \quad W_T = x_T,$$

which is equivalent to (1.6.3). □

Matrix notation helps us simplify expressions such as (1.6.3), so let us take a moment to introduce some associated notation. For any set $\mathcal{Z}$, we write $\mathcal{Z}^{m\times n}$ for the set of m-by-n matrices whose entries are elements of $\mathcal{Z}$. For example, trading strategies are elements of $\mathcal{P}_0^{1\times J}$. Unless explicitly specified otherwise (as we did for trading strategies), vectors of processes are assumed to be column vectors and we write $\mathcal{Z}^n$ rather than $\mathcal{Z}^{n\times 1}$. If $\mathcal{Z}$ is the set of all (adapted, predictable) processes, then $\mathcal{Z}^n$ is the set of **n-dimensional** (adapted, predictable) processes. If $\mathcal{Z}$ is a set of processes, we typically use superscripts to index vectors and matrices. For $x \in \mathcal{L}^{n\times m}$ and $y \in \mathcal{L}^{m\times l}$, the process $x \bullet y \in \mathcal{L}^{n\times l}$ is defined by the analog of the usual matrix multiplication formula:

$$(x \bullet y)^{ij} = \sum\nolimits_{k=1}^{m} x^{ik} \bullet y^{kj}.$$

With these conventions in place, we define the J-dimensional processes

$$\delta \equiv \begin{pmatrix} \delta^1 \\ \vdots \\ \delta^J \end{pmatrix} \quad \text{and} \quad V \equiv \begin{pmatrix} V^1 \\ \vdots \\ V^J \end{pmatrix}, \tag{1.6.4}$$

as well as

$$S \equiv V - \delta \quad \text{and} \quad G \equiv S + \delta \bullet \mathbf{t} = V + \delta_- \bullet \mathbf{t}. \tag{1.6.5}$$

The budget equation (1.6.3) can be restated more succinctly as

$$\theta V = -x_- \bullet \mathbf{t} + \theta \bullet G, \quad \theta_T V_T = x_T. \tag{1.6.6}$$

Other versions of the budget equation are obtained by a change of the unit of account (also known as a change of numeraire). Suppose the strictly positive adapted process π represents a unit conversion factor at every spot. A cash flow or price process x expressed in the

original unit of account becomes πx in the new unit of account. The gain process G in the new units becomes

$$G^{\pi} \equiv \pi V + (\pi\delta)_{-} \bullet \mathbf{t}.$$

Clearly, changing units should not affect the validity of a budget equation, resulting in the following version, whose proof is a simple exercise.

PROPOSITION 1.6.4. *For all $\pi \in \mathcal{L}_{++}$, the trading strategy θ generates the cash flow x if and only if*

$$\pi\theta V = -(\pi x)_{-} \bullet \mathbf{t} + \theta \bullet G^{\pi}, \quad \theta_T V_T = x_T. \tag{1.6.7}$$

Given the J contracts (1.6.1) and the vector notation (1.6.4), we write (δ, V) to refer to these contracts and we write $X(\delta, V)$ for the market that is implemented by trading in (δ, V). A trading strategy θ in (δ, V) defines a new, synthetic contract, which can be thought of as a share in a fund following strategy θ.

DEFINITION 1.6.5. The trading strategy $\theta \in \mathcal{P}_0^{1\times J}$, generating the cash flow x, defines the **synthetic contract** $(\delta^{\theta}, V^{\theta})$, where

$$(\delta_0^{\theta}, V_0^{\theta}) = (0, \theta_1 V_0), \quad (\delta_t^{\theta}, V_t^{\theta}) = (x_t, \theta_t V_t), \quad t = 1, \ldots, T.$$

A contract is **synthetic** in (δ, V) if it is of the form $(\delta^{\theta}, V^{\theta})$ for some trading strategy θ.

Note that the ex-dividend price process $S^{\theta} = V^{\theta} - \delta^{\theta}$ is given by

$$S_{t-1}^{\theta} = \theta_t S_{t-1}, \quad t = 1, \ldots, T; \quad S_T^{\theta} = 0, \tag{1.6.8}$$

and the gain process of the contract $(\delta^{\theta}, V^{\theta})$ is given by

$$G^{\theta} \equiv V^{\theta} + \delta_{-}^{\theta} \bullet \mathbf{t} = V_0^{\theta} + \theta \bullet G, \tag{1.6.9}$$

as can be seen by rearranging the budget equation (1.6.6).

Consider now the trading strategies $\theta_1, \ldots, \theta_m$ in the original contracts (1.6.1) and let $\alpha \in \mathcal{P}_0^m$ be a trading strategy in the synthetic contracts $(\delta^{\theta_i}, V^{\theta_i})$, $i = 1, \ldots, m$, generating the cash flow x. One can think of α as a trading strategy in m funds, where fund i follows trading strategy θ_i. The same cash flow x can be generated by the trading strategy $\theta = \sum_{i=1}^{m} \alpha_i \theta_i$ in the original contracts (δ, V). This argument justifies the following observations.

PROPOSITION 1.6.6. *The market implemented by any synthetic contracts in (δ, V) is a subset of the market $X(\delta, V)$ implemented by (δ, V). The market implemented by (δ, V) and any number of synthetic contracts in (δ, V) is $X(\delta, V)$.*

Dividend processes of synthetic contracts correspond to the cash flows that are marketed in $X(\delta, V)$:

PROPOSITION 1.6.7. *A cash flow c is marketed in $X(\delta, V)$ if and only if there exists a trading strategy θ in (δ, V) such that*

$$c_t = \delta_t^\theta, \quad t = 1, \ldots, T.$$

In particular, the market $X(\delta, V)$ is complete if and only if every cash flow c with $c_0 = 0$ is the dividend process of some contract that is synthetic in (δ, V).

PROOF. Recall that c is marketed in $X(\delta, V)$ if and only if there exists $x \in X(\delta, V)$ such that $c_t = x_t$ for $t > 0$. By the definition of $X(\delta, V)$ and a synthetic contract, the latter condition is equivalent to the existence of a trading strategy in (δ, V) such that $c_t = \delta_t^\theta$ for $t > 0$. □

Finally, we relate synthetic contracts to traded contracts. Note that in order to conclude that a traded contract is synthetic, the assumption that the market is arbitrage-free is essential.

PROPOSITION 1.6.8. *Every contract that is synthetic in (δ, V) is traded in $X(\delta, V)$. Conversely, if the market $X(\delta, V)$ is arbitrage-free, every contract that is traded in $X(\delta, V)$ is synthetic in (δ, V).*

PROOF. Buying a contract (δ^*, V^*) that is synthetic in (δ, V) at any spot generates a cash flow in $X(\delta^*, V^*)$, which, by Proposition 1.6.6, is a subset of $X(\delta, V)$. Therefore, a synthetic contract in (δ, V) is traded in $X(\delta, V)$.

Conversely, suppose that (δ^*, V^*) is traded in $X(\delta, V)$ and let x be the cash flow generated by buying (δ^*, V^*) at time zero. Since $x \in X(\delta, V)$, there exists a trading strategy θ in (δ, V) that generates x. Since the time-zero dividend of every contract is assumed to be zero by convention, it follows that $\delta^* = \delta^\theta$. Assuming the market is arbitrage-free, it must also be the case that $V^* = V^\theta$. □

1.7. Money market account and returns

A special type of contract we call a money-market account implements single-period default-free borrowing and lending: A unit of account invested at time $t-1$ pays $1 + r_t$ at time t, where the interest rate r_t is determined at time $t-1$ and is therefore $\mathcal{F}_{t-1}$-measurable. In practice, a loan can be made default-free by the posting of sufficient collateral, an important aspect of real-world markets that is not modeled here. The rate r_t is often referred to in the literature as the (period-t) risk-free rate. The rate r_t applies to a loan over a single period and is therefore a short-term interest rate and the predictable process r (with the convention $r_0 = 0$) is a short-term interest rate process, a term we abbreviate to short-rate process. More precisely, we adopt the following terminology.

DEFINITION 1.7.1. A **money-market account (MMA)** is a contract (δ^0, V^0), with ex-dividend price process $S^0 = V^0 - \delta^0$, such that for some $r \in \mathcal{P}_0$,

$$S^0_{t-1} = 1 \quad \text{and} \quad V^0_t = 1 + r_t, \quad t = 1, \dots, T.$$

The predictable process r is the account's (interest) **rate process.** Given a reference market X, a process $r \in \mathcal{P}_0$ is a **short-rate process** if r is the rate process of an MMA that is traded in X.

The following observations can be verified by the reader. We call two contracts **equivalent** if each is synthetic in the other.

PROPOSITION 1.7.2. *Suppose the market is arbitrage-free and r is a short-rate process. Then r is unique and $1 + r$ is strictly positive. Moreover, every traded contract (δ^0, V^0) whose value process V^0 is predictable and strictly positive is equivalent to an MMA and satisfies*

$$\frac{V^0_t}{S^0_{t-1}} = 1 + r_t, \quad S^0 \equiv V^0 - \delta^0, \quad t = 1, \dots, T. \tag{1.7.1}$$

In applications where the market is implemented by a given finite set of contracts, it is common to assume that one of these contracts is an MMA. In all such applications, we label the contracts generating the market as

$$\left(\delta^0, V^0\right), \left(\delta^1, V^1\right), \dots, \left(\delta^J, V^J\right), \tag{1.7.2}$$

where contract zero is the MMA, with rate process r and gain process G^0. Since $\delta^0_0 = r_0 = 0$, we have

$$V^0 = 1 + r, \quad \delta^0_- = r_-, \quad \delta^0_T = V_T, \quad G^0 = 1 + r \bullet \mathbf{t}. \tag{1.7.3}$$

Of course, all of last section's results apply to this case after a simple relabeling of the contracts, which affects the form of the budget equation. We adopt the matrix notation (1.6.4) and (1.6.5), where (δ, V) and S and G refer to contracts $1, \dots, J$ and exclude contract zero. A trading strategy in the $1 + J$ contracts (1.7.2) is denoted as

$$\left(\theta^0, \theta\right) \in \mathcal{P} \times \mathcal{P}^{1 \times J},$$

where θ^0_t represents the ex-dividend value in the MMA at the beginning of period $t > 0$, and θ is a trading strategy in the remaining J contracts. Adapting the budget equation (1.6.6) to this notation, it follows that the trading strategy (θ^0, θ) generates cash flow x if and only if

$$\theta^0 V^0 + \theta V = \left(\theta^0 r - x_-\right) \bullet \mathbf{t} + \theta \bullet G, \quad \theta^0_T V^0_T + \theta_T V_T = x_T. \tag{1.7.4}$$

In applications where portfolios values can be assumed to stay positive it is common to focus on returns rather than prices. The **return process** R^j associated with contract j is defined by

$$R^j_0 = 1, \quad R^j_t \equiv \frac{V^j_t}{S^j_{t-1}}, \quad t = 1, \dots, T, \tag{1.7.5}$$

provided that every S^j_{t-1} is nowhere zero, which we assume for the remainder of this section. We refer to R^j_t as the period-t return of contract j. Note that if r is the market's short-rate process, then

$$R^0 \equiv 1 + r.$$

The return process R^θ of a trading strategy(θ^0, θ) is defined as the return process of the corresponding synthetic contract, which we denote by $\left(\delta^\theta, V^\theta\right)$, instead of the more cumbersome $(\delta^{\left(\theta^0,\theta\right)}, V^{\left(\theta^0,\theta\right)})$. Letting $S^\theta \equiv V^\theta - \delta^\theta$, the return process R^θ is well defined provided the denominator in the following definition is nowhere zero:

$$R^\theta_0 \equiv 1, \quad R^\theta_t \equiv \frac{V^\theta_t}{S^\theta_{t-1}}, \quad t = 1, \ldots, T.$$

The **excess return** process of the trading strategy (θ^0, θ) is the difference $R^\theta - R^0$.

Portfolio returns can be more parsimoniously represented in terms of returns and portfolio weights, provided of course all relevant returns are well defined. Consider any trading strategy (θ^0, θ) such that the time-$(t-1)$ ex-dividend portfolio value

$$S^\theta_{t-1} = \theta^0_t + \theta_t S_{t-1}$$

is nowhere zero. Associated with (θ^0, θ) is a row vector

$$\psi = \left(\psi^1, \ldots, \psi^J\right) \in \mathcal{P}_0^{1\times J},$$

defined by

$$\psi^j_0 = 0, \quad \psi^j_t \equiv \frac{\theta^j_t S^j_{t-1}}{S^\theta_{t-1}}, \quad t = 1, \ldots, T. \tag{1.7.6}$$

At the beginning period t, which is time $t-1$, ψ^j_t represents the proportion of the portfolio's ex-dividend value S^θ_{t-1} that is allocated to contract j. The vector ψ_t omits the proportion allocated to the MMA, which can be computed as $\psi^0_t \equiv 1 - \sum_{j=1}^J \psi^j_t$. We will refer to ψ_t, which can be any element of $L\left(\mathcal{F}_{t-1}\right)^{1\times J}$, as a (period-$t$) **portfolio allocation**, and to ψ, which can be any element of $\mathcal{P}_0^{1\times J}$, as a **portfolio allocation policy.** The period-t return R^θ_t can be computed entirely in terms of ψ_t, which we therefore also denote, abusing notation, by R^ψ_t:

$$R^\theta_t \equiv R^\psi_t \equiv R^0_t + \sum_{j=1}^J \psi^j_t \left(R^j_t - R^0_t\right). \tag{1.7.7}$$

In the following chapter we will discuss notions of optimal portfolio allocations.

1.8. Exercises

Exercise 1 This exercise reviews some simple arbitrage arguments. Assume that there is a single period ($T = 1$) and therefore the filtration consists of spot zero and N time-one spots. In an arbitrage-free market X, you can trade a **stock**, which is a contract with time-zero price S (a scalar) and time-one payoff V (a random variable or element of $\mathbb{R}^N$), and a money-market account (MMA) whose single-period rate is r (a scalar). Buying the stock at time zero generates the cash flow $(-S, V) \in X$ and selling (or shorting) the stock generates the cash flow $(S, -V) \in X$. Investing a unit of account in the MMA generates the cash flow $(-1, 1 + r) \in X$ and borrowing a unit of account from the MMA generates the cash flow $(1, -(1 + r)) \in X$. Note that since the market is arbitrage-free, $1 + r > 0$.

(a) (Forward pricing) A **forward contract** for delivery of the stock at time one is a contract whose time-zero price is by definition zero and whose time-one payoff takes the form $V - K$ for a scalar K, which is the contract's **delivery price.** The **long** contract position generates the cash flow $(0, V - K)$ and the **short** contract position generates the cash flow $(0, K - V)$. The forward contract is **traded** in X if these cash flows are in X, in which case K defines the stock's **forward price** F. Notice that the delivery price K is part of the contract definition. The unique value of K such that $(0, V - K) \in X$ defines F. How can you create a synthetic forward using the stock and the MMA? What is the implied relationship between S and F assuming the forward contract is traded? What is an explicit arbitrage (assuming the forward is traded) if the claimed relationship between S and F is violated? How do your answers change if the stock cannot be shorted,[6] that is, if $(-S, V) \in X$ but X is only a cone, not a linear subspace, and $(S, -V) \notin X$.

(b) (Put-call parity) Entering a long forward contract generates a positive payoff on the event $\{V > K\}$ and a negative payoff on the event $\{V < K\}$. A trader entering a long forward position has the right to receive the positive payoff and the obligation to pay the negative payoff. A (European) **call option** (on the stock), or just **call**, with **strike** K is defined by the same payoff but without the obligation. The call's payoff is therefore $(V - K)^+$. Starting with a short forward contract and removing the obligation part results in the definition of the payoff $(K - V)^+$ of a (European) **put option** (on the stock), or just **put**, with **strike** K. A call or put option, with payoff V^o, is traded in X if $(-S^o, V^o)$ for a necessarily unique scalar S^o, which defines the option's **price** or **premium**. The cash flow $(-S^o, V^o)$ is generated by **buying**

[6] An arbitrageur that has a positive inventory of shares of the stock can incrementally sell some and therefore $(S, -V) \in X$. A short-sale constraint binds if the arbitrageur has no share inventory and cannot borrow shares. The economics of the implied lease market dictate the cost of shorting the stock.

(or going **long**) the option and the cash flow $(S^o, -V^o)$ is generated by **selling** (or going **short** or **writing**) the option. Given a commitment to pay V (resp. receive V) at time one, buying the call (resp. put) is a form of insurance, paying a premium S^o at time zero in return for the right to pay K rather than V on the event $\{V > K\}$ (resp. receive K rather than V on the event $\{V < K\}$).

Suppose the call and the put, both with strike K, are traded, with respective premia S^c and S^p. How can you use them to synthetically create the forward contract of part (a) with delivery price K? Assuming this forward contract is traded, and therefore $F = K$ is the forward price of the stock, what is the implied relationship (known as put-call parity) between $S^c - S^p$ and F? Suppose the relationship is violated. What is an explicit arbitrage using the options, the stock and the MMA (but not directly a forward)? How do your answers change if the stock cannot be shorted?

(c) (Simple binomial pricing) Specialize the setting further by assuming there are only two states ($N = 2$). Assume that the stock return V/S takes the values $1 + u$ and $1 + d$, where $u > d > -1$. The market X is implemented by trading in the stock and the MMA. What are necessary and sufficient conditions on the parameters r, u and d for X to be arbitrage-free? Proceeding under the assumption that X is arbitrage-free, show that X is complete and compute the corresponding present-value function. What is the premium S^c of a (European) call according to this present-value function? Compute a replicating portfolio, that is, a portfolio in the stock and the MMA whose time-one payoff is the same as that of the call option. Finally, confirm that the time-zero value of the replicating portfolio is consistent with your earlier call premium calculation.

Exercise 2 Suppose the reference market X is arbitrage-free and complete. Explain why for every cash flow c, there is a unique scalar $\Pi(c)$ such that $c - \Pi(c) 1_{\Omega\times\{0\}} \in X$. Then show that the function $\Pi : \mathcal{L} \to \mathbb{R}$ so defined is a present-value function.

Exercise 3 Suppose the reference market is arbitrage-free. Show that a cash flow is marketed if and only if it is assigned the same value by every present-value function. As a corollary, show that a cash flow is traded if and only if it is assigned the value zero by every present-value function.

Exercise 4 Fix a reference complete market with corresponding present-value function Π, which induces a spot-(F, t) present-value function $\Pi_{F,t}$ for every spot (F, t) (see Definition 1.4.5).

(a) Show that if Π prices the contract (δ, V) (see Definition 1.4.7) then V uniquely solves the backward recursion

$$V(F,t) = \delta(F,t) + \Pi_{F,t}\left(V 1_{F\times\{t+1\}}\right), \quad V = \delta_T.$$

(b) Consider an American option with payout process D as defined in Example 3.6.7, whose notation we use here. For every stopping time τ, let $\delta^{\tau} \equiv D1_{[\tau]}$ and define the adapted process V^{τ} by letting $V^{\tau}(F,t) \equiv \Pi_{F,t}(\delta^{\tau})$ for every spot (F,t). (Note that, by construction, Π prices the contract $(\delta^{\tau}, V^{\tau})$ and part (a) applies.) Define the adapted process V^{*} by letting, for every spot (F,t),

$$V^{*}(F,t) \equiv \max\left\{V^{\tau}(F,t) \mid \tau \in \mathcal{T}\right\}.$$

Show that V^{*} uniquely solves the backward recursion

$$V^{*}(F,t) = \max\left\{D(F,t), \Pi_{F,t}\left(V^{*}1_{F\times\{t+1\}}\right)\right\}, \quad V_{T}^{*} = D_{T}^{+}.$$

Label each spot (F,t) **red** if $V^{*}(F,t) = D(F,t)$ and **green** otherwise. A state ω corresponds to a path from spot zero to a terminal spot. Define $\tau^{*}(\omega)$ to be the time of the first red spot along the path ω, with the convention $\tau^{*}(\omega) = \infty$ if ω is a sequence of green spots only. Show that the stopping time τ^{*} so defined is dominant (in the sense that $D1_{[\tau^{*}]}$ is a dominant cash flow choice for the option).

Exercise 5 (a) Show that

$$x \bullet (y \bullet z) = (xy) \bullet z, \quad x, y, z \in \mathcal{L}.$$

(This is sometimes called the associate property of stochastic integrals and applies in more general stochastic settings.)

(b) Given contracts $(\delta, V) \in \mathcal{L}^{J\times 2}$, the trading strategies $\theta_1, \ldots, \theta_m$ define corresponding synthetic contracts $\left(\delta^{\theta_i}, V^{\theta_i}\right)$, $i = 1, \ldots, m$. Let $\alpha = (\alpha^1, \ldots, \alpha^m)$ be a trading strategy in the m synthetic contracts, generating cash flow x. Use the corresponding budget equation in the form

$$\sum\nolimits_{i=1}^{m} \alpha_i V^{\theta_i} = -x_{-} \bullet \mathbf{t} + \sum\nolimits_{i=1}^{m} \alpha_i \bullet G^{\theta_i}, \quad \sum\nolimits_{i=1}^{m} \alpha_{iT} V_{T}^{\theta_i} = x_T,$$

and part (a) to verify that the cash flow x is also generated by the trading strategy $\theta \equiv \sum_{i=1}^{m} \alpha_i \theta_i$ in the original contracts (δ, V).

Exercise 6 Give an example of contracts $(\delta, V) \in \mathcal{L}^{J\times 2}$ and a contract (δ^{*}, V^{*}) that is *not* a synthetic contract in (δ, V), yet it is traded in $X(\delta, V)$.

Exercise 7 (Binomial replication) This exercise extends Exercise 1c to the multi-period case, using the notation $U \equiv 1+u$ and $D \equiv 1+d$. Suppose $\Omega = \{0,1\}^{T}$ and the filtration $\{\mathcal{F}_t\}$ is generated by the process b, where

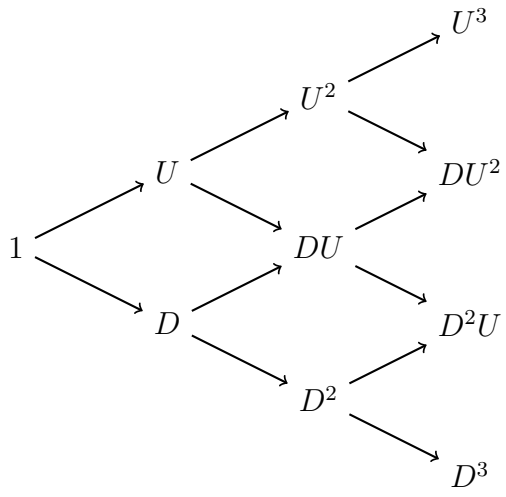
$$b_0 \equiv 0, \quad b_t(\omega) \equiv \omega_t, \quad \omega = (\omega_1, \ldots, \omega_T) \in \Omega, \quad t \in \{1, \ldots, T\}.$$

The process Z is specified by a given initial value $Z_0 \in (0, \infty)$ and the recursion

$$\frac{Z_t}{Z_{t-1}} \equiv b_t U + (1 - b_t) D, \quad t = 1, \ldots, T,$$

for given constants $U > D > 0$. Note that the process Z also generates the filtration $\{\mathcal{F}_t\}$, since $(Z_1, \ldots, Z_t)$ and $(b_1, \ldots, b_t)$ are mutually uniquely determined path by path.

(a) The **recombining tree** is the graph whose **nodes** are all possible values of Z_t/Z_0, and whose arrows connect a value z off Z_{t-1}/Z_0 to the corresponding possible values zU and zD of Z_t/Z_0. For example, here is the recombining tree for $T = 3$:



Note that spots correspond to the paths on the recombining tree. For example, there are three ways to go from the time-zero node to node DU^2, corresponding to the fact that there are three spots that coalesce to form the node. While the spot "remembers" how we got there (there is only one path to each spot), the node "forgets" (there can be many paths to each node). How many nodes are there and how many spots? How do these numbers increase with T? What is their order of magnitude for $T = 100$?

(b) Assume that the market is arbitrage-free and is implemented by two contracts: an MMA (δ^0, V^0) with a constant rate process $r > -1$, and a **stock** (δ, V) with ex-dividend price process $S \equiv V - \delta$, specified in terms of a constant **dividend yield** $y > -1$ by

$$S_{t-1} \equiv Z_{t-1} \quad \text{and} \quad V_t \equiv (1+y)\, Z_t, \quad t = 1, \ldots T. \tag{1.8.1}$$

As always, the convention is $\delta_0 \equiv 0$ and $\delta_T \equiv V_T$. Explain (without too much formalism) why

$$U\,(1+y) > 1 + r > D\,(1+y)\,.$$

(c) Consider a contract (δ^*, V^*) that pays no dividends prior to T, that is, $\delta^*_- = 0$. Assume that the contract's price can be expressed as $V^*_t = f_t\,(Z_t)$ for some functions $f_t : \mathcal{N}_t \to (0, \infty)$. Let (θ^0, θ) be

a trading strategy in the contracts (δ^0, V^0) and (δ, V), defining the synthetic contract $(\delta^\theta, V^\theta)$. Show that if $(\delta^*, V^*) = (\delta^\theta, V^\theta)$, then

$$\theta_t^0 = g_t\left(Z_{t-1}\right) \quad \text{and} \quad \theta_t = h_t\left(Z_{t-1}\right), \quad t = 1, \ldots, T,$$

for functions $g_t, h_t : \mathcal{N}_{t-1} \to \mathbb{R}$, for which you should be able to give explicit formulas in terms of f_t and the model parameters. Finally, provide a recursive algorithm for computing f_t.

CHAPTER 2

Probabilistic Methods in Arbitrage Pricing

The arbitrage-pricing theory of Chapter 1 postulates an exhaustive set of possible states but makes no use of any probabilities over these states. This chapter introduces probabilistic representations of valuation rules consistent with a given arbitrage-free market, which are useful in developing theoretical and computational methodology and essential in empirical applications.

2.1. Probability basics

In this section we review some essential probabilistic concepts and notation. To last chapter's primitives of a finite state space Ω and a filtration on this state space, we add a reference **probability** (measure) on the subsets of Ω, that is, a function $P : 2^\Omega \to [0,1]$ such that $P(\Omega) = 1$ and for all events A and B,

$$A \cap B = \emptyset \quad \Longrightarrow \quad P(A \cup B) = P(A) + P(B).$$

We assume that P has **full support:** $P(A) > 0$ if $A \neq \emptyset$.

The **expectation** or **mean** of a random variable x (under P) is

$$\mathbb{E}[x] \equiv \sum_{\omega \in \Omega} x(\omega) P(\{\omega\}) = \sum_{\alpha \in \{x(\omega) | \omega \in \Omega\}} \alpha P(\{x = \alpha\}).$$

The function $\mathbb{E} : \mathbb{R}^\Omega \to \mathbb{R}$ so defined is the **expectation operator** relative to P; it is the unique linear functional on $\mathbb{R}^\Omega$ that is positive ($\mathbb{E}[x] > 0$ if $0 \neq x \geq 0$) and satisfies $\mathbb{E}[1_A] = P(A)$ for every event A. We often omit excessive parentheses, as in $\mathbb{E}x = \mathbb{E}[x]$ and $P[x \leq \alpha] = P(\{x \leq \alpha\})$. The demeaned version of a random variable x is denoted $\hat{x} \equiv x - \mathbb{E}x$. The **covariance** of two random variables x, y is the scalar

$$\operatorname{cov}[x,y] \equiv \mathbb{E}[\hat{x}\hat{y}] = \mathbb{E}[xy] - \mathbb{E}[x]\,\mathbb{E}[y]. \tag{2.1.1}$$

We can then define the **variance** $\operatorname{var}[x] \equiv \operatorname{cov}[x,x]$, the **standard deviation** $\operatorname{stdev}[x] \equiv \sqrt{\operatorname{var}[x]}$, and provided x and y have positive variance, the **correlation coefficient**

$$\operatorname{corr}[x,y] \equiv \frac{\operatorname{cov}[x,y]}{\operatorname{stdev}[x]\operatorname{stdev}[y]}.$$

The random variables x and y are **uncorrelated** if $\operatorname{cov}[x,y] = 0$. Two uncorrelated random variables can be nontrivially determined by the same random source. For example, if $\Omega = \{-1, 0, 1\}$ and each state

is assigned the same probability, then the identity random variable $x(\omega) = \omega$ is uncorrelated with its square $x^2(\omega) = |\omega|$. If $f(x)$ and $g(y)$ are uncorrelated for all functions $f, g : \mathbb{R} \to \mathbb{R}$, then the random variables x, y are said to be **stochastically independent**, or just **independent** where there is no risk of confusion with other types of independence (like linear independence). By virtue of Proposition 1.1.4, the independence of x and y is really a property of the algebras $\sigma(x)$ and $\sigma(y)$. Define two algebras $\mathcal{A}$ and $\mathcal{B}$ to be (stochastically) **independent** if every random variable in $L(\mathcal{A})$ is uncorrelated with every random variable in $L(\mathcal{B})$. Then x and y are independent if and only if $\sigma(x)$ and $\sigma(y)$ are independent.

PROPOSITION 2.1.1. *For all algebras $\mathcal{A}$ and $\mathcal{B}$ and corresponding partitions $\mathcal{A}^0$ and $\mathcal{B}^0$, the following are equivalent conditions.*

(1) *$\mathcal{A}$ and $\mathcal{B}$ are independent.*
(2) *$P(A \cap B) = P(A) P(B)$ for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$.*
(3) *$P(A \cap B) = P(A) P(B)$ for all $A \in \mathcal{A}^0$ and $B \in \mathcal{B}^0$.*

PROOF. Since the covariance operator is linear in each of its arguments and every random variable is a linear combination of indicator functions, $\mathcal{A}$ and $\mathcal{B}$ are independent if and only if 1_A and 1_B are uncorrelated for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$ (resp. $A \in \mathcal{A}^0$ and $B \in \mathcal{B}^0$). Since $\operatorname{cov}(1_A, 1_B) = P(A \cap B) - P(A) P(B)$, this proves the equivalence of the first and second (resp. third) conditions. □

More generally, the random variables $x_1, \ldots, x_n$ are said to be (stochastically) **independent** if for each $i \in \{1, \ldots, n\}$, the algebras $\sigma(x_i)$ and $\sigma(\{x_j \mid j \neq i\})$ are independent.[1]

PROPOSITION 2.1.2. *For all random variables $x_1, \ldots, x_n$, the following are equivalent conditions.*

(1) *$x_1, \ldots, x_n$ are stochastically independent.*
(2) *$\sigma(x_1, \ldots, x_{i-1})$ and $\sigma(x_i)$ are independent for $i = 2, \ldots, n$.*
(3) *$\mathbb{E} \prod_{i=1}^n f_i(x_i) = \prod_{i=1}^n \mathbb{E} f_i(x_i)$ for all $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$.*
(4) *$P\left[\bigcap_{i=1}^n \{x_i \in S_i\}\right] = \prod_{i=1}^n P[x_i \in S_i]$ for all $S_1, \ldots, S_n \subseteq \mathbb{R}$.*

PROOF. That (1) implies (2) is immediate. We show (3) assuming (2) by induction in n. For $n = 2$, the claim is valid by definition. Suppose it is true for $n - 1$ variables. For all $f_i : \mathbb{R} \to \mathbb{R}$, (2) implies that $\prod_{i=1}^{n-1} f_i(x_i)$ and $f_n(x_n)$ are uncorrelated and therefore the expectation of their product equals the product of their expectation, which together with the inductive hypothesis gives (3). To show that

[1] This is not the same as pairwise independence. For example, suppose x_1 and x_2 are independent random variables, each taking the values $+1$ and -1 with equal probability. (As in Example 1.1.1 with $T = 2$, $x_i(\omega) = \omega_i$ and equal probability for each state.) Then the random variables $x_1, x_2, x_1 x_2$ are not independent even though any two of them are independent.

(3) implies (4), set $f_i(x_i) = 1_{\{x_i \in S_i\}}$. Finally, we show (1) assuming (4) holds. The partition generating $\sigma(x_1)$ is the set of all nonempty events of the form $A = \{x_1 = \alpha_1\}$, $\alpha_1 \in \mathbb{R}$, and the partition generating $\sigma(x_2, \ldots, x_n)$ is the set of the nonempty events of the form $B = \{x_2 = \alpha_2, \ldots, x_n = \alpha_n\}$, $\alpha_2, \ldots, \alpha_n \in \mathbb{R}$. For such A and B, (4) implies $P(A \cap B) = \prod_{i=1}^{n} P[x_i = \alpha_i]$ and $P(B) = \prod_{i=2}^{n} P[x_i = \alpha_i]$. Therefore $P(A \cap B) = P[A] P[B]$. By Proposition 2.1.1, this proves that $\sigma(x_1)$ is stochastically independent of $\sigma(x_2, \ldots, x_n)$. The same argument applies for any permutation of the $x_1, \ldots, x_n$, completing the proof. □

Uncorrelatedness and stochastic independence can equivalently be thought of as orthogonality conditions in the space $\hat{L} = \left\{\hat{x} \mid x \in \mathbb{R}^{\Omega}\right\}$ of zero-mean random variables, with the covariance inner product $\langle \hat{x} \mid \hat{y} \rangle = \operatorname{cov}[x, y]$ and induced norm $\|\hat{x}\| = \sqrt{\langle \hat{x} \mid \hat{y} \rangle} = \operatorname{stdev}[x]$. The random variables x, y are uncorrelated if and only if $\hat{x}$ and $\hat{y}$ are orthogonal in $\hat{L}$. Similarly, the algebras $\mathcal{A}$ and $\mathcal{B}$ are independent if and only if $\hat{L}(\mathcal{A}) \equiv \{\hat{x} \mid x \in L(\mathcal{A})\}$ and $\hat{L}(\mathcal{B})$ are orthogonal in $\hat{L}$. The Cauchy-Schwarz inequality in this context states that $|\operatorname{corr}[x, y]| \leq 1$ with equality holding if and only if $\hat{x} = \alpha \hat{y}$ or $\hat{y} = \alpha \hat{x}$ for some scalar α.

The conditional probability $P(\cdot \mid B) : 2^{\Omega} \to [0, 1]$ given a positive probability event B is defined by **Bayes' rule:**

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{2.1.2}$$

The expectation operator corresponding to $P(\cdot \mid B)$ is denoted by $\mathbb{E}[\cdot \mid B]$ and is easily seen to satisfy

$$\mathbb{E}[x \mid B] = \frac{\mathbb{E}[x 1_B]}{P(B)}, \quad \text{for all } x \in \mathbb{R}^{\Omega}. \tag{2.1.3}$$

The **conditional expectation operator** (under P) given the algebra $\mathcal{B}$ generated by the partition $\{B_1, \ldots, B_n\}$ is defined by

$$\mathbb{E}[x \mid \mathcal{B}] = \sum_{i=1}^{n} \mathbb{E}[x \mid B_i] 1_{B_i}, \quad \text{for all } x \in \mathbb{R}^{\Omega}. \tag{2.1.4}$$

The following proposition shows that the random variable $\mathbb{E}[x \mid \mathcal{B}]$ is the best estimate of x given information $\mathcal{B}$, in the sense of minimizing expected squared error.

PROPOSITION 2.1.3. *For all algebras $\mathcal{B}$, and random variables x, y, the following are equivalent conditions, provided y is $\mathcal{B}$-measurable.*

(1) $y = \mathbb{E}[x \mid \mathcal{B}]$.
(2) $\mathbb{E}[(x - y)^2] \leq \mathbb{E}[(x - z)^2]$ *for all* $z \in L(\mathcal{B})$.
(3) $\mathbb{E}[yz] = \mathbb{E}[xz]$ *for all* $z \in L(\mathcal{B})$.
(4) $\mathbb{E}y = \mathbb{E}x$ *and* $\operatorname{corr}[x - y, z] = 0$ *for all* $z \in L(\mathcal{B})$.
(5) $\mathbb{E}[y 1_B] = \mathbb{E}[x 1_B]$ *for all* $B \in \mathcal{B}$.

PROOF. Condition (2) is a norm-minimization problem in the vector space of random variables with the inner product $\langle x \mid y \rangle = \mathbb{E}[xy]$. The $y \in L(\mathcal{B})$ satisfying condition (2) is the projection of x onto $L(\mathcal{B})$. By the orthogonal projection theorem (Corollary B.4.2), such a $y \in L(\mathcal{B})$ is equivalently characterized by the orthogonality condition $\langle x \mid y - z \rangle = 0$ for all $z \in L(\mathcal{B})$, which is condition (3). The equivalence of (3) and (4) is immediate from the definitions. The equivalence of (3) and (5) is also straightforward, given that every $z \in L(\mathcal{B})$ is a linear combination of indicator functions of events in $\mathcal{B}$. Finally, let $\mathcal{B}^0 \equiv \{B_1, \ldots, B_n\}$ denote the partition generating $\mathcal{B}$. Since the indicator of any $B \in \mathcal{B}$ is the sum of random variables of the form 1_{B_i}, condition (5) is equivalent to its version resulting after replacing $\mathcal{B}$ with $\mathcal{B}^0$, whose equivalence to condition (1) follows easily from the definitions. □

REMARK 2.1.4. In infinite state space extensions of the theory, definition (2.1.4) is not generally meaningful and a version of the last proposition's orthogonality condition forms the basis for the usual textbook definition of a conditional expectation. The uniqueness claim in the last proposition relies on the full-support assumption on P. In general, any two random variables y and y' that are conditional expectations of x given $\mathcal{B}$ must satisfy $P[y = y'] = 1$ but can take arbitrary values on any $B \in \mathcal{B}$ such that $P(B) = 0$. We will have no need for this generality here, but the technicality becomes unavoidable in infinite state space extensions of the theory.

COROLLARY 2.1.5. *The algebras $\mathcal{A}$ and $\mathcal{B}$ are stochastically independent if and only if $\mathbb{E}[x \mid \mathcal{B}] = \mathbb{E}[x]$ for all $x \in L(\mathcal{A})$.*

PROOF. Recall that $\mathcal{A}$ and $\mathcal{B}$ are independent if and only if $\hat{L}(\mathcal{A})$ and $\hat{L}(\mathcal{B})$ are orthogonal in $\hat{L}$ under the covariance inner product, which is in turn equivalent to the requirement that the projection of every $\hat{x} \in \hat{L}(\mathcal{A})$ onto $\hat{L}(\mathcal{B})$ is zero. By Proposition 2.1.3, the last condition can be restated as $\mathbb{E}[\hat{x} \mid \mathcal{B}] = 0$ for all $\hat{x} \in \hat{L}(\mathcal{A})$. □

The important **law of iterated expectations** states that

$$\mathcal{B} \subseteq \mathcal{A} \quad \Longrightarrow \quad \mathbb{E}[\mathbb{E}[x \mid \mathcal{A}] \mid \mathcal{B}] = \mathbb{E}[x \mid \mathcal{B}], \tag{2.1.5}$$

for every random variable x and algebras $\mathcal{A}$ and $\mathcal{B}$. Indeed if $\mathcal{B} \subseteq \mathcal{A}$, $L(\mathcal{B})$ is a linear subspace of $L(\mathcal{A})$, and therefore projecting x on $L(\mathcal{B})$ is equivalent to first projecting x on $L(\mathcal{A})$ and then further projecting on $L(\mathcal{B})$, which translates to $\mathbb{E}[x \mid \mathcal{B}] = \mathbb{E}[\mathbb{E}[x \mid \mathcal{A}] \mid \mathcal{B}]$.

Another important property of conditional expectations, stated below, is an immediate consequence of identities (2.1.3) and (2.1.4) in the current context. The following less direct proof is instructive, however, in illustrating how the orthogonality characterization of expectations can be used in ways that also apply in infinite state-space extensions of the theory.

PROPOSITION 2.1.6. *For every random x and algebra $\mathcal{B}$, if $b \in L(\mathcal{B})$, then $\mathbb{E}[bx \mid \mathcal{B}] = b\mathbb{E}[x \mid \mathcal{B}]$.*

PROOF. Fix any $b \in L(\mathcal{B})$ and let $y = b\mathbb{E}[x \mid \mathcal{B}]$. We use the characterization of condition 2 of Proposition 2.1.3 twice, first to justify the middle equality in

$$\mathbb{E}[(bx)z] = \mathbb{E}[x(bz)] = \mathbb{E}[\mathbb{E}[x \mid \mathcal{B}](bz)] = \mathbb{E}[yz], \quad \text{for all } z \in L(\mathcal{B}),$$

and then to conclude that since $y \in L(\mathcal{B})$, the above condition implies $y = \mathbb{E}[bx \mid \mathcal{B}]$. □

The **conditional expectation** of a random variable x **given** the **random variables** $y_1, \ldots, y_n$ is defined by

$$\mathbb{E}[x \mid y_1, \ldots, y_n] = \mathbb{E}[x \mid \sigma(y_1, \ldots, y_n)],$$

which is a function of $(y_1, \ldots, y_n)$ in the sense of Proposition 1.1.4. Thus $\mathbb{E}[x \mid y_1, \ldots, y_n]$ is the projection of x onto $L(\sigma(y_1, \ldots, y_n))$, which is the linear space of all random variables of the form $f(y_1, \ldots, y_n)$ for arbitrary $f : \mathbb{R}^n \to \mathbb{R}$. This contrasts with a simple linear regression,[2] where x is projected on the smaller subspace of all random variables of the same form but with f restricted to be linear.

As in the last chapter, throughout this chapter we take as given an underlying filtration $(\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_T)$ on Ω, abbreviated to $\{\mathcal{F}_t\}$, with $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_T = 2^\Omega$. We write $L_t \equiv L(\mathcal{F}_t)$ for the set of all $\mathcal{F}_t$-measurable random variables, $\mathbb{E}_t[x]$ or $\mathbb{E}_t x$ for the conditional expectation $\mathbb{E}[x \mid \mathcal{F}_t]$, and cov_t for the conditional covariance given $\mathcal{F}_t$, defined as in (2.1.1) but with $\mathbb{E}_t$ in place of $\mathbb{E}$. Conditional variances, standard deviations and correlation coefficients given $\mathcal{F}_t$ are defined and denoted analogously.

A **martingale** (under P) is any adapted process M such that

$$M_t = \mathbb{E}_t M_s \quad \text{for all } s > t.$$

By the law of iterated expectations (2.1.5), an adapted process M is a martingale if and only if $\mathbb{E}_{t-1}[\Delta M_t] = 0$ for all $t > 0$ if and only if $M_t = \mathbb{E}_t[M_T]$ for all t. We let $\mathcal{M}$ denote the set of all martingales and $\mathcal{M}_0 \equiv \{M \in \mathcal{M} \mid M_0 = 0\}$, which is the set of all **zero-mean martingales** since a martingale M is in $\mathcal{M}_0$ if and only if $\mathbb{E}M_t = 0$ for all t.

The following simple observation has far-reaching implications for finance interpretations as well as martingale theory in general, including the extension of the theory of stochastic integrals in more general settings.

[2] A prominent case where the two projections coincide is when $(x, y_1, \ldots, y_n)$ has a Gaussian distribution, the definition of which requires an infinite state space.

PROPOSITION 2.1.7. *For every martingale M and predictable process ϕ, the integral $\phi \bullet M$ is a martingale.*[3]

PROOF. If $\phi_t \in L_{t-1}$, then $\mathbb{E}_{t-1}\left[\Delta\left(\phi \bullet M\right)_t\right] = \mathbb{E}_{t-1}\left[\phi_t \Delta M_t\right] = \phi_t \mathbb{E}_{t-1}\left[\Delta M_t\right] = 0$. □

EXAMPLE 2.1.8. In the context of Example 1.1.1, assume that P represents a fair coin: $P\left(\{\omega\}\right) = 2^{-T}$ for all $\omega \in \Omega$. A gambler betting on a coin toss gains a dollar if the outcome is heads and loses a dollar otherwise. Contingent on state ω, a gambler betting on the first t coin tosses gains (or loses if negative) $B_t\left(\omega\right) = \omega_1 + \omega_2 + \cdots + \omega_t$. Letting $B_0 = 0$, this defines an adapted process B whose increments $\Delta B_1, \ldots, \Delta B_T$ are stochastically independent and generate the underlying filtration. Since the algebra $\mathcal{F}_{t-1} = \sigma\left(\Delta B_1, \ldots, \Delta B_{t-1}\right)$ is independent of (the algebra generated by) ΔB_t, $\mathbb{E}_{t-1}\Delta B_t = \mathbb{E}\Delta B_t = 0$ and therefore $B \in \mathcal{M}_0$. A gambler who starts betting at time t and quits at a later time u can expect zero total gains: $\mathbb{E}_t\left[B_u - B_t\right] = 0$. A natural question is whether the gambler could beat the odds by following some clever strategy, wagering $\phi_t \in L_t$ dollars on the tth coin toss at time $t-1$. With the convention $\phi_0 = 0$, such a strategy defines an element ϕ of $\mathcal{P}_0$ and the process $M = \phi \bullet B$ defines the corresponding cumulative gains. By Proposition 2.1.7, M is also a zero-mean martingale. Therefore, a gambler following strategy ϕ from time t up to a later time u must again expected zero total gains: $\mathbb{E}_t\left[M_u - M_t\right] = 0$. In this sense, the gambler cannot beat the odds. If on the other hand the gambler were allowed to bet indefinitely, then the following doubling strategy would beat the odds: Bet one dollar, if heads stop, otherwise bet two dollars plus one on a second coin toss, if heads stop, otherwise bet four dollars plus one on a third coin toss, and so on. If allowed to play indefinitely, eventually heads is bound to show up resulting in a gain of one dollar. Of course, losses can be staggering in the meantime and any finite borrowing limit is enough to rule out the strategy. Here this type of doubling strategy is ruled out by the finite number of periods, but it is something one has to technically deal with in infinite-horizon or continuous-time models. ♢

Every adapted process x has a unique **Doob decomposition**:

$$x = x_0 + x^p + M, \quad x^p \in \mathcal{P}_0, \quad M \in \mathcal{M}_0. \tag{2.1.6}$$

To see why, note that (2.1.6) implies $\Delta x = \Delta x^p + \Delta M$ and therefore,

$$\Delta x^p_t = \mathbb{E}_{t-1}\Delta x_t, \quad t = 1, \ldots, T.$$

Conversely, if x^p is recursively defined by the last equation and $x^p_0 = 0$, then $x^p \in \mathcal{P}_0$ and $x - x^p$ is a martingale. The process x^p is known as the **compensator** of x.

[3]Note that it is not enough to assume that ϕ is adapted for $\phi \bullet M$ to be a martingale, a fact that is the main motivation behind the introduction of the notion of a predictable process.

2.2. Beta pricing and frontier returns

In Section 1.7 we defined excess returns as the difference between the returns of a traded contract and that of a traded (default-free) money market account (MMA). In this section we show that in an arbitrage-free market expected excess returns are proportional to the return's covariance with a traded return that is characterized by the property of minimizing variance given its expected value. The resulting expression for expected returns is known as a beta pricing equation.

Beta pricing is a characterization of single-period traded cash flows; it is not about the specific contracts implementing the market and a version of the theory can be formulated with or without a traded MMA. For expositional simplicity, however, we adopt the setting of Section 1.7. We take as given an arbitrage-free market that is implemented by the $1 + J$ contracts (1.7.2), where $S^j_{t-1} \neq 0$ everywhere (meaning at all states) and therefore returns are well-defined in (1.7.5). Contract zero is an MMA with rate process $r \in \mathcal{P}_0$. The corresponding period-t return is

$$R^0_t = 1 + r_t \in L_{t-1}.$$

A period-t portfolio allocation $\psi_t = \left(\psi^1_t, \ldots, \psi^J_t\right) \in L^{1\times J}_{t-1}$ results in the portfolio return

$$R^\psi_t \equiv R^0_t + \sum_{j=1}^{J} \psi^j_t \left(R^j_t - R^0_t\right).$$

The set of period-t **traded returns** is

$$\mathcal{R}_t \equiv \left\{ R^\psi_t \mid \psi_t \in L^{1\times J}_{t-1} \right\}.$$

A key observation is that returns can be mixed: For all $R^1_t, R^2_t \in \mathcal{R}_t$ and $\alpha_t \in L_{t-1}$, $(1 - \alpha_t) R^1_t + \alpha_t R^2_t \in \mathcal{R}_t$. To avoid trivialities, we assume throughout that for every time t, there exists some $R_t \in \mathcal{R}_t$ such that $\mathrm{var}_{t-1}[R_t] > 0$ everywhere.

Definition 2.2.1. The traded return $R^*_t \in \mathcal{R}_t$ is a (minimum variance) **frontier** return if for all $R_t \in \mathcal{R}_t$,

$$\{\mathbb{E}_{t-1} R_t = \mathbb{E}_{t-1} R^*_t\} \subseteq \{\mathrm{var}_{t-1}[R_t] \geq \mathrm{var}_{t-1}[R^*_t]\}.$$

The property of being a frontier return can also be defined spot by spot: $R^*_t \in \mathcal{R}_t$ is a **frontier** return at spot $(F, t-1)$ if for all $R_t \in \mathcal{R}_t$,

$$\mathbb{E}[R_t \mid F] = \mathbb{E}[R^*_t \mid F] \implies \mathrm{var}[R_t \mid F] \geq \mathrm{var}[R^*_t \mid F].$$

Since for all $F \in \mathcal{F}_{t-1}$, $R_t, R^*_t \in \mathcal{R}_t$ implies $R_t 1_F + R^*_t 1_{F^c} \in \mathcal{R}_t$, it follows that R^*_t is a frontier return if and only if it is a frontier return at every time-$(t-1)$ spot. Similarly, the following characterization of frontier returns, as well as the entire discussion of the remainder of this section, applies separately spot by spot.

LEMMA 2.2.2. *$R_t^* \in \mathcal{R}_t$ is a frontier return if and only if for all $R_t \in \mathcal{R}_t$, $\{\mathbb{E}_{t-1}[R_t - R_t^*] = 0\} \subseteq \{\mathrm{cov}_{t-1}[R_t - R_t^*, R_t^*] = 0\}$.*

PROOF. Fix any spot $(F, t-1)$ and let (F_i, t), $i = 0, 1, \ldots, d$, denote its immediate successor spots. On $\mathbb{R}^{1+d}$, we use the inner product

$$\langle x \mid y \rangle \equiv \sum_{i=0}^{d} x_i y_i P[F_i \mid F].$$

Let $\mathcal{R}_{F,t} \equiv \{(R(F_0, t), \ldots, R(F_d, t)) \mid R \in \mathcal{R}_t\}$ and define the linear manifold M of all $z \in \mathcal{R}_{F,t}$ such that $\sum_i (z_i - z_i^*) P[F_i \mid F] = 0$, where $z^* \equiv (R^*(F_0, t), \ldots, R^*(F_d, t))$. The fact that R_t^* is a frontier return implies that $\|z\| \geq \|z^*\|$ for all $z \in M$. In other words, z^* is the orthogonal projection of the zero vector onto M. By the orthogonal projection theorem (Corollary B.4.2), z^* is characterized by the orthogonality condition $\langle z - z^* \mid z^* \rangle = 0$ for all $z \in M$, which can be restated as, for all $R_t \in \mathcal{R}_t$, $\mathbb{E}[R_t - R_t^* \mid F] = 0$ implies $\mathbb{E}[(R_t - R_t^*) R_t^* \mid F] = 0$. The lemma's claim follows. □

We can use the above lemma to determine all **frontier allocations**, that is, all portfolio allocations generating frontier returns. Let

$$1 + \mu_t^i \equiv \mathbb{E}_{t-1} R_t^i \quad \text{and} \quad \Sigma_t^{ij} \equiv \mathrm{cov}_{t-1}\left[R^i, R^j\right], \quad i, j = 1, \ldots, J,$$

and define the column vector $\mu_t \equiv \left(\mu_t^1, \ldots, \mu_t^J\right)'$ and the $J \times J$ symmetric matrix $\Sigma_t \equiv [\Sigma_t^{ij}]$, which is positive semidefinite (at every time-$(t-1)$ spot). For simplicity, we assume that Σ_t is full rank and therefore positive definite. The assumption of a full-rank Σ_t is easily seen to be equivalent to the assumption that the contracts are **everywhere non-redundant**, meaning that conditionally on any spot $(F, t-1)$, there is no $j \in \{0, 1, \ldots, J\}$ such that the return R_t^i on F can be generated by an allocation in the remaining contracts. By Lemma 2.2.2, an allocation ψ_t^* generates a frontier return R_t^* if and only if for every period-t allocation ψ_t,

$$(\psi - \psi_t^*)(\mu_t - r_t) = 0 \quad \text{implies} \quad (\psi - \psi_t^*) \Sigma_t \psi_t^{*\prime} = 0.$$

In other words, whenever $\psi_t - \psi_t^*$ is orthogonal to $\mu_t - r_t$, it is also orthogonal to $\Sigma_t \psi_t^{*\prime}$ (all conditionally on each time-$(t-1)$ spot). Therefore, $\mu_t - r_t$ is collinear to $\Sigma_t \psi_t^{*\prime}$ in the sense that there exists $\alpha_t \in L_{t-1}$ such that $\alpha_t(\mu_t - r_t) = \Sigma_t \psi_t^{*\prime}$. Rearranging, we can parametrically describe all frontier allocations by

$$\psi_t^* = \alpha_t (\mu_t - r_t)' \Sigma_t^{-1}, \quad \alpha_t \in L_{t-1}. \tag{2.2.1}$$

Note that if ψ_t^* is any nonzero frontier allocation, then every other frontier allocation takes the form $\alpha_t \psi_t^*$ for some $\alpha_t \in L_{t-1}$, a fact known as **two-fund separation**. The term reflects the idea that every frontier return can be achieved by allocating a value proportion α_t in a fund allocated according to ψ_t^* and the rest in a fund that is an MMA.

As we will see shortly, two-fund separation is valid even without the assumption that the contracts are everywhere non-redundant.

The following is this section's central result, showing that the frontier returns other than the MMA return are exactly the returns relative to which beta-pricing is possible.

PROPOSITION 2.2.3. *For all* $R_t^* \in \mathcal{R}_t$ *such that* $\operatorname{var}_{t-1}[R_t^*] > 0$ *everywhere, the following two conditions are equivalent:*

(1) R_t^* *is a frontier return.*
(2) *For all* $R_t \in \mathcal{R}_t$,

$$\mathbb{E}_{t-1}R_t - R_t^0 = \frac{\operatorname{cov}_{t-1}[R_t^*, R_t]}{\operatorname{var}_{t-1}[R_t^*]}\left(\mathbb{E}_{t-1}R_t^* - R_t^0\right) \tag{2.2.2}$$

and for some $R_t \in \mathcal{R}_t$, $\mathbb{E}_{t-1}R_t \neq R_t^0$ *everywhere.*

PROOF. ($1 \implies 2$) Suppose $R_t^* \in \mathcal{R}_t$ is a frontier return and therefore $0 = \operatorname{var}_{t-1}[R_t^0] \geq \operatorname{var}_{t-1}[R_t^*]$ on the event $\{\mathbb{E}_{t-1}R_t^* = R_t^0\}$. Since, $\operatorname{var}_{t-1}[R_t^*] > 0$ everywhere, the event $\{\mathbb{E}_{t-1}R_t^* = R_t^0\}$ is empty.

Given any $R_t \in \mathcal{R}_t$, define $\tilde{R}_t \in \mathcal{R}_t$ by letting

$$\tilde{R}_t \equiv R_t^* + R_t - R_t^0 \text{ on } \left\{\mathbb{E}_{t-1}R_t = R_t^0\right\}, \text{ and}$$

$$\tilde{R}_t \equiv R_t^0 + \frac{\mathbb{E}_{t-1}R_t^* - R_t^0}{\mathbb{E}_{t-1}R_t - R_t^0}(R_t - R_t^0) \text{ on } \left\{\mathbb{E}_{t-1}R_t \neq R_t^0\right\}.$$

By construction, $\mathbb{E}_{t-1}\left[\tilde{R}_t - R_t^*\right] = 0$ and therefore, by Lemma 2.2.2, $\operatorname{cov}_{t-1}[\tilde{R}_t - R_t^*, R_t^*] = 0$, which expands to equation (2.2.2).

($2 \implies 1$) Conversely, suppose the second condition holds, which clearly implies that $\mathbb{E}_{t-1}R_t^* \neq R_t^0$ everywhere. We show that R_t^* is a frontier return by verifying the orthogonality condition of Lemma 2.2.2. Consider any $R_t \in \mathcal{R}_t$ such that $\mathbb{E}_{t-1}[R_t - R_t^*] = 0$. Applying the beta pricing equation and canceling out the term $\mathbb{E}_{t-1}R_t^* - R_t^0$ on each side, we conclude that $\operatorname{cov}_{t-1}[R_t - R_t^*, R_t^*] = 0$. □

The so-called beta-pricing equation (2.2.2) has been of considerable interest in empirical work, since it suggests that the slope coefficient of a linear regression (commonly denoted by β) can be used to explain expected excess returns. In practice, we can only identify a frontier return with error. We therefore have to consider the beta pricing equation relative to some proxy return $R_t^p = R_t^* + \varepsilon_t$, where the error ε_t is judged to be small. It is an instructive exercise to show that an arbitrarily small value of $\mathbb{E}_{t-1}\varepsilon_t^2$ is consistent with the existence of a traded return whose beta with respect to R_t^p is arbitrarily different from its beta with respect to R^*. The basic idea is that leverage (that is, borrowing from the MMA to invest in a risky portfolio) can arbitrarily amplify any discrepancy between the two betas. This pitfall can be avoided if we instead focus on excess returns normalized by their standard deviation.

The time-$(t-1)$ **Sharpe ratio** of a period-t return R_t is the ratio

$$\mathcal{S}_{t-1}[R_t] \equiv \frac{\mathbb{E}_{t-1}R_t - R_t^0}{\text{stdev}_{t-1}[R_t]}, \tag{2.2.3}$$

provided the denominator is nowhere zero. We adopt the convention that $\mathcal{S}_{t-1}[R_t] \equiv 0$ on the event $\{\text{var}_{t-1}[R_t] = 0\}$. The beta-pricing equation (2.2.2) can be restated in terms of Sharpe ratios as

$$\mathcal{S}_{t-1}[R_t] = \text{corr}_{t-1}[R_t^*, R_t]\mathcal{S}_{t-1}[R_t^*]. \tag{2.2.4}$$

This version of the beta-pricing equation is robust to replacing R_t^* with a highly correlated proxy R_t^p, in the following sense.[4]

PROPOSITION 2.2.4. *For all $R_t^*, R_t^p, R_t \in \mathcal{R}_t$, equation (2.2.4) implies its approximate version:*

$$\mathcal{S}_{t-1}[R_t] = \text{corr}_{t-1}[R_t^p, R_t]\mathcal{S}_{t-1}[R_t^p] + \varepsilon_{t-1}, \tag{2.2.5}$$

where

$$|\varepsilon_{t-1}| \leq 2\,|\mathcal{S}_{t-1}[R_t^*]|\,|1 - \text{corr}_{t-1}[R_t^*, R_t^p]|.$$

PROOF. We adopt the notation of the proof of Lemma 2.2.2, since the argument relates to a single step of the information tree following a given spot $(F, t-1)$. Analogously to z^*, let $z \equiv (R(F_0,t), \ldots, R(F_d,t))$ and $z^p \equiv (R^p(F_0,t), \ldots, R^p(F_d,t))$. We write $\mathcal{S}[z]$ for the value of $\mathcal{S}_{t-1}[R_t]$ on F, and analogously for $\mathcal{S}[z^*]$ and $\mathcal{S}[z^p]$. The vectors z, z^*, z^p can be thought of as random variables on $\{0, 1, \ldots, d\}$. For any such random variable x, we use the notation

$$\tilde{x} \equiv \frac{x - \sum_i x_i P[F_i \mid F]}{\text{stdev}[x]},$$

where the standard deviation is relative to the probability assigning mass $P[F_i \mid F]$ to state i. Note that $\langle \tilde{x} \mid \tilde{x} \rangle = 1$ and $\langle \tilde{x} \mid \tilde{y} \rangle = \text{corr}[x, y]$. Let $1 - \delta \equiv \langle \tilde{z}^* \mid \tilde{z}^p \rangle$, which is the value of $\text{corr}_{t-1}[R_t^*, R_t^p]$ on F, and let $\varepsilon \equiv \mathcal{S}[z] - \langle \tilde{z}^p \mid \tilde{z} \rangle \mathcal{S}[R^p]$, which is the value of ε_{t-1} on F, as defined by (2.2.5). Condition (2.2.4) implies that $\mathcal{S}[z] = \langle \tilde{z}^* \mid \tilde{z} \rangle \mathcal{S}[z^*]$ and $\mathcal{S}[z^p] = (1-\delta)\mathcal{S}[z^*]$. Therefore,

$$\begin{aligned} \frac{\varepsilon^2}{\mathcal{S}[z^*]^2} &= \langle \tilde{z}^* - (1-\delta)\tilde{z}^p \mid \tilde{z} \rangle^2 \\ &\leq \langle \tilde{z}^* - (1-\delta)\tilde{z}^p \mid \tilde{z}^* - (1-\delta)\tilde{z}^p \rangle^2, \end{aligned}$$

where the last inequality follows by the Cauchy-Schwarz inequality. Expanding the last factor, we find it equals $(2\delta - \delta^2)^2$. Therefore $\varepsilon^2 \leq \mathcal{S}[z^*]^2 (2\delta)^2$, which is the claimed bound. □

[4] While this section's material is standard textbook material, to my knowledge, Proposition 2.2.4 first appeared in Skiadas [2009].

The set of frontier traded returns other than the MMA return are the set of the traded returns of maximum absolute Sharpe ratio. We state this claim more formally below, using the convention that indeterminate Sharpe ratios are assigned the value zero.

PROPOSITION 2.2.5. *$R_t^* \in \mathcal{R}_t$ is a frontier return if and only if $|\mathcal{S}_{t-1}[R_t^*]| \geq |\mathcal{S}_{t-1}[R_t]|$ for all $R_t \in \mathcal{R}_t$.*

PROOF. The "only if" part is a corollary of Proposition 2.2.3, as can easily be seen by taking absolute values on both sides of equation (2.2.4) and using the fact that the absolute correlation is less than one. (Alternatively, one can show the claim more directly from the definition of frontier returns.) The converse is immediate from the definitions. □

Sharpe ratios are invariant to positions in the MMA. For any $\alpha_t \in L_{t-1}$, a mix of α_t of a portfolio allocation ψ_t and $1-\alpha_t$ in the MMA results in a return that has the same absolute Sharpe ratio as R_t^ψ:

$$\left|\mathcal{S}_{t-1}[R_t^0 + \alpha_t(R_t^\psi - R_t^0)]\right| = \left|\mathcal{S}_{t-1}[R_t^\psi]\right|. \tag{2.2.6}$$

Moreover, as we vary α_t we trace out all (conditional) mean-standard deviation pairs $(\alpha_t \mathbb{E}_{t-1} R_t^\psi, |\alpha_t| \operatorname{stdev}_{t-1}[R_t^\psi])$ of traded returns that are consistent with the given absolute Sharpe ratio value. In particular, if $R^\psi = R^*$ is a frontier return other than the risk-free return R_t^0, varying α_t traces out all traded returns with maximum absolute Sharpe ratio, that is, all frontier returns. We therefore obtain once again the two-fund separation result introduced earlier, but without the non-redundancy assumption. If the contracts are assumed to be everywhere non-redundant, we can use expression (2.2.1) to compute the maximum squared Sharpe ratio as

$$\mathcal{S}_{t-1}[R_t^*]^2 = (\mu_t - r_t)' \Sigma_t^{-1} (\mu_t - r_t). \tag{2.2.7}$$

The returns corresponding to frontier returns with positive Sharpe ratio, or equivalently positive (conditional) expected returns, are known as (conditionally) **mean-variance efficient**, since besides minimizing variance given expected return, they also achieve a maximum expected return given the variance of the return (all conditionally beginning-of-period information). Since the seminal contribution of Markowitz [1952], mean-variance efficiency has played a prominent role as a criterion for portfolio choice. The criterion is clearly limited in that it is myopic (only focuses on single-period returns) and uses variance as a measure of portfolio risk. A more sophisticated theory of portfolio choice, albeit with its own limitations, is developed in Chapter 3. Mean-variance efficiency resurfaces as a building block of optimal portfolios in the more sophisticated theory for a useful class of return dynamics in high trading frequency. So the concept is more robust than

this section's discussion suggests. Its empirical implementation, however, has its own serious limitations, which are beyond the scope of this text. Ultimately, the theory of mean-variance efficiency mainly serves as a parsimonious model that highlights the benefits of diversification.

2.3. State-price densities

A state price process was defined in Section 1.2 as a representation of a present-value function. A state price density process is essentially the same object, but with its values expressed as a density relative to a given reference probability. This simple construct opens the door for the use of probabilistic methods.

We continue to take as given a market X in the usual stochastic setting, consisting of a filtration $\{\mathcal{F}_t\}$ on the finite state space Ω, where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_T = 2^\Omega$, and a full-support probability P on $\mathcal{F}_T$. A state-price density can be defined as an adapted process π such that a state-price process p is well-defined by letting $p(F,t) = \pi(F,t)\,P(F)$ for every spot (F,t). The following equivalent definition bypasses reference to a state-price process and extends directly to infinite state-space settings.

Definition 2.3.1. A **state-price density process** or **SPD** (relative to the probability P) is any π in $\mathcal{L}_{++}$ such that a present-value function $\Pi : \mathcal{L} \to \mathbb{R}$ is well-defined by

$$\Pi(c) = \frac{1}{\pi_0}\mathbb{E}\left[\sum\nolimits_{t=0}^{T} \pi_t c_t\right], \quad c \in \mathcal{L}. \tag{2.3.1}$$

In this case, π is said to **represent** Π.

Proposition 2.3.2. *Every present-value function admits an SPD representation, which is unique up to positive scaling.*

Proof. Given a present-value function Π, let π denote its (unique) Riesz representation in $\mathcal{L}$ with the inner product $\langle x \mid y\rangle = \mathbb{E}\sum_{t=0}^{T} x_t y_t$. The positivity of Π implies that $\pi \in \mathcal{L}_{++}$, and therefore π is an SPD representing Π. Conversely, if π is an SPD representing Π, then π/π_0 is the Riesz representation of Π. □

A present-value function Π, which is defined from the perspective of time zero, induces a spot-(F,t) present-value function $\Pi_{F,t}$ (Definition 1.4.5) for every other spot (F,t). If the SPD π represents Π, then

$$\Pi_{F,t}(c) = \frac{1}{\pi(F,t)}\mathbb{E}\left[\sum_{u=t}^{T} \pi_u c_u \mid F\right]. \tag{2.3.2}$$

This link between conditional valuation and conditional expectation turns out to be methodologically quite useful.

In Proposition 1.4.8 we saw that if Π is a present-value function and (δ, V) is a traded contract, then Π prices the contract in the sense that

$$V(F,t) = \Pi_{F,t}(\delta) \quad \text{for every spot } (F,t).$$

If π is an SPD representing Π, then the same pricing condition can be expressed as

$$V_t = \frac{1}{\pi_t}\mathbb{E}_t\left[\sum\nolimits_{u=t}^{T}\pi_u\delta_u\right], \quad t = 0,\dots,T-1. \tag{2.3.3}$$

In this case we say that π **prices** the contract (δ, V). The following variant of the argument used in Proposition 1.4.8 shows the pricing condition (2.3.3) directly in a way that extends to infinite sate-space settings, where expression (2.3.2) is not generally meaningful. Given any time t and $F \in \mathcal{F}_t$, set to zero the present value of the cash flow (1.3.2) generated by buying the contract at time t on the event F and holding it to time T to find

$$\mathbb{E}[\pi_t V_t 1_F] = \mathbb{E}\left[\left(\sum\nolimits_{u=t}^{T}\pi_u\delta_u\right)1_F\right].$$

Equation (2.3.3) follows after applying Proposition 2.1.3.

The second part of Proposition 1.4.8 can also be restated in terms of SPDs:

PROPOSITION 2.3.3. *A process $\pi \in \mathcal{L}_{++}$ is an SPD for the market implemented by a set of contracts if and only if it prices every contract in the given set.*

The following link of pricing to martingales plays a central methodological role, especially in technically more advanced incarnations of the theory.

PROPOSITION 2.3.4. *A process $\pi \in \mathcal{L}_{++}$ prices the contract (δ, V) if and only if $G^\pi \equiv \pi V + (\pi\delta)_- \bullet \mathbf{t}$ is a martingale.*

PROOF. Multiply equation (2.3.3) by π_t and add $\sum_{u=0}^{t-1}\pi_u\delta_u$ on both sides to find $G_t^\pi = \mathbb{E}_t[G_T^\pi]$. This calculation can be reversed. □

REMARK 2.3.5. The preceding martingale condition characterizes an SPD for a market implemented by given contracts. Suppose that the market X is implemented by contracts (δ^j, V^j), $j = 1,\dots,J$, and let $\delta = \left(\delta^1,\dots,\delta^J\right)'$ and $V = \left(V^1,\dots,V^J\right)'$. Write G for the corresponding column vector of gain processes. Then, by the last two propositions, $\pi \in \mathcal{L}_{++}$ is an SPD if and only if G^π is a martingale, meaning that each $G^{j\pi} = \pi V^j + (\pi\delta^j)_- \bullet \mathbf{t}$ is a martingale. For the synthetic contract generated by a trading strategy θ, we have $G^{\theta\pi} = \pi_0 V_0^\theta + \theta \bullet G^\pi$, which is a martingale if G^π is a martingale, since θ is predictable. This observation merely reflects the fact that an SPD prices every traded contract, and every synthetic contract is traded. ♢

Yet another useful way of expressing the fact that an SPD π prices the contract (δ, V) is in the recursive form

$$S_{t-1} = \mathbb{E}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}} V_t\right], \quad t = 1, \ldots, T, \tag{2.3.4}$$

where $S \equiv V - \delta$. The equivalence of this condition to (2.3.3) follows from the law of iterated expectations (2.1.4). Alternatively, equation (2.3.4) can be rearranged to $\mathbb{E}_{t-1}[\Delta G_t^{\pi}] = 0$, which is one of the equivalent ways of stating the martingale property of G^{π}.

If $S_{t-1} \neq 0$ everywhere, then the return $R_t = V_t / S_{t-1}$ is well-defined and recursion (2.3.4) implies

$$1 = \mathbb{E}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}} R_t\right], \quad t = 1, \ldots, T. \tag{2.3.5}$$

Conversely, the validity of the pricing restriction (2.3.5) for every return R of a traded contract implies that π prices an arbitrary traded contract (δ, S), provided that at every nonterminal spot there exists at least some traded contract with non-zero ex-dividend price. If the event $F \equiv \{S_{t-1} = 0\}$ is empty, then clearly (2.3.5) with $R_t = V_t / S_{t-1}$ implies (2.3.4). What if F is nonempty? Equation (2.3.5) cannot be applied directly to the contract (δ, V). Instead, apply (2.3.5) twice—once to any traded contract with a well-defined return on F, and once to the return of a portfolio that holds this same contract as well as the contract (δ, V) whose ex-dividend price vanishes on F. Solving the resulting pair of equations gives the desired pricing condition (2.3.4).

Applying the pricing equation (2.3.5) to a traded MMA with rate process r results in an expression that gives r as a function of an SPD:

$$\frac{1}{1 + r_t} = \mathbb{E}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}}\right], \quad t = 1, \ldots, T. \tag{2.3.6}$$

This equation can be rearranged to

$$r_t = -\mathbb{E}_{t-1}\left[\frac{\Delta \pi_t}{\pi_{t-1}}\right] + \varepsilon_t, \quad \text{where } \varepsilon_t \equiv \frac{r_t^2}{1 + r_t}.$$

If a period in the model represents a sufficiently short time interval, then ε_t is numerically negligible and the period-t risk-free interest rate r_t is approximately equal to minus the expected growth rate of the SPD π_t over the period, conditionally on the beginning-of-period information. In this chapter's final section we will see that this approximation becomes an exact relationship in continuous-time, where r_t represents a continuously compounded interest rate.

Suppose now that π is an SPD, an MMA is traded and therefore the short-rate process r is given in terms of π as just described. The contract (δ, V) is said to be priced risk neutrally if $S_{t-1} = \mathbb{E}_{t-1} V_t / (1 + r_t)$, a relationship that is valid for an MMA. For a general contract, we expect that the ex-dividend price S_t must be adjusted relative to the

risk neutral valuation to reflect the uncertainty of the end-of-period payoff V_t. The appropriate risk adjustment can be written precisely by rearranging equation (2.3.4) to

$$S_{t-1} - \frac{\mathbb{E}_{t-1} V_t}{1 + r_t} = \operatorname{cov}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}}, V_t\right] = \operatorname{cov}_{t-1}\left[\frac{\Delta\pi_t}{\pi_{t-1}}, V_t\right]. \tag{2.3.7}$$

In words, the risk adjustment relative to the risk-neutral valuation is equal to the covariance of the end-of-period value with the SPD growth rate conditionally on the beginning-of-period information.

Assuming the return $R_t = V_t/S_{t-1}$ is well defined, the preceding relationship is often expressed as a restriction on the excess return $R_t - R_t^0$, where $R^0 \equiv 1 + r$. Using the definition of covariance and the pricing restrictions (2.3.5) and (2.3.6), we find

$$\mathbb{E}_{t-1} R_t - R_t^0 = -R_t^0 \operatorname{cov}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}}, R_t\right]. \tag{2.3.8}$$

To make the connection to last section's beta-pricing pricing notion, we need to project π_t/π_{t-1} to the space of traded returns. In the context of Section 2.2, we can write $\pi_t/\pi_{t-1} = R_t^* + \epsilon_t$, where R_t^* is a period-t traded return and $\mathbb{E}_{t-1}[\epsilon_t R_t] = 0$ for every period-t traded return R_t (and therefore $\mathbb{E}_{t-1}\epsilon_t = 0$). The existence of a (unique) such decomposition follows by a simple projection argument conditionally at each time-$(t-1)$ spot, using the inner product of the proof of Lemma 2.2.2. Substituting into equation (2.3.8) and dividing the resulting equation by its special case with R^* in place of R, we obtain the beta pricing equation

$$\mathbb{E}_{t-1} R_t - R_t^0 = \frac{\operatorname{cov}_{t-1}[R_t^*, R_t]}{\operatorname{var}_{t-1}[R_t^*]} \left(\mathbb{E}_{t-1} R_t^* - R_t^0\right).$$

The return R_t^* maximizes correlation with π_t/π_{t-1} conditionally on time-$(t-1)$ information among all period-t traded returns; a claim that reduces to the Cauchy-Schwarz inequality in the geometry used to define R_t^* and ϵ_t as the components of an orthogonal projection of π_t/π_{t-1}.

Another noteworthy consequence of (2.3.8) is the inequality

$$\mathcal{S}_{t-1}[R_t]^2 \leq (1 + r_t)^2 \operatorname{var}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}}\right]. \tag{2.3.9}$$

To show it, write the cov_{t-1} of equation (2.3.8) in terms of $\operatorname{corr}_{t-1}$ and use the fact that the latter is valued in $[-1, 1]$. Given the short rate r_t, the conditional variance of an SPD growth rate places an upper bound on the maximum conditional squared Sharpe ratio of a traded return (an expression for which was derived at the end of the last section). This fact is known as the **Hansen-Jagannathan bound**.[5]

[5]Named after the contribution of Hansen and Jagannathan [1991].

2.4. Equivalent martingale measures

Equivalent martingale measures are another methodologically useful way of representing present-value functions. Prior to introducing the concept, let us fix, for the entire section, a reference arbitrage-free market X and underlying full-support probability P. Suppose that π is an SPD relative to these primitives, as defined in the last section. Suppose also, for now, that an MMA is traded in X, with corresponding short rate process r. Consider any nonterminal spot $(F, t-1)$, with immediate successor spots (F_i, t), $i = 0, \ldots, d$, and define the positive scalars $Q(F_i \mid F)$ by

$$\frac{\pi(F_i, t)}{\pi(F, t-1)} = \frac{1}{1 + r(F, t)} \frac{Q(F_i \mid F)}{P(F_i \mid F)}, \quad i = 0, \ldots, d. \tag{2.4.1}$$

Pricing of the MMA, as in equation (2.3.6), gives

$$\frac{1}{1 + r(F, t)} = \sum_{i=0}^{d} \frac{\pi(F_i, t)}{\pi(F, t-1)} P(F_i \mid F), \tag{2.4.2}$$

and therefore $\sum_{i=0}^{d} Q(F_t \mid F) = 1$. In other words, $Q(F_i \mid F)$ can be thought of as a positive transition probability from spot $(F, t-1)$ to spot (F_i, t). For every path on the information tree, which corresponds to a state revealed at the terminal date T, these transitions probabilities can be multiplied through, thus defining a full-support probability Q, which is the Equivalent Martingale Measure (EMM) corresponding to π. In fact, this construction applies without the assumption that an MMA is traded by taking equation (2.4.2) to be the definition of $r(F, t)$.

We write $\mathbb{E}^Q$ for the expectation operator relative to Q, simplifying the conditional expectation notation $\mathbb{E}^Q[x \mid \mathcal{F}_t]$ to $\mathbb{E}^Q_t[x]$ or $\mathbb{E}^Q_t x$. (Omitting the superscript implies the default reference probability: $\mathbb{E} = \mathbb{E}^P$.) Identity (2.4.1) can be used to price any traded contract (δ, V), with $S = V - \delta$, resulting in

$$S_{t-1} = \frac{\mathbb{E}^Q_{t-1} V_t}{1 + r_t}, \quad t = 1, \ldots, T. \tag{2.4.3}$$

This pricing relationship is often referred to as **risk-neutral pricing**,[6] although the term is misleading in terms of economic content; the risk associated with the payoff V_t is priced in equation (2.4.3) just as it would by the SPD π. In fact, Q can be thought of as representing a pure (conditional) price of risk over a single period in the sense that if at spot $(F, t-1)$ one enters a forward contract to exchange a time-t fixed payment $f_{i,t}$ for a contingent unit of account that is paid if and

[6] The idea of risk-neutral pricing already appears in Arrow [1971] and Drèze [1971], and it is exploited in option pricing by Cox and Ross [1976]. The term "equivalent martingale measure" is due to Harrison and Kreps [1979].

only if spot (F_i, t) materializes, then $f_{i,t} = Q(F_i \mid F)$. In other words, $Q(F_i \mid F)$ is the one-period ahead spot-$(F, t-1)$-forward price of a unit payment at spot (F_i, t). The following example shows the role of the conditional expectation $\mathbb{E}_{t-1}^{Q}$ as a one-period-ahead forward pricing operator more formally.

EXAMPLE 2.4.1. Apply the pricing equation (2.4.3) for an assumed traded contract $(\delta, V) = (D - f, D - f)$, where D is an adapted process and f is a predictable process, with $f_0 = D_0 = 0$. For every $t > 0$, we interpret D_t as the time-t value of some asset and f_t as the time-$(t-1)$ forward price of the asset for time-t delivery. Buying the contract at time $t-1$ and selling it at time t results in the net cash flow x, where $x_u = 0$ for $u \neq t$ and $x_t = D_t - f_t$. From the perspective of time $t-1$, the trade is equivalent to entering a forward contract for time-t delivery of the asset at the forward price f_t, whose value is determined at time $t-1$. The pricing equation (2.4.3) in this case reduces to $f_t = \mathbb{E}_{t-1}^{Q} D_t, t = 1, \ldots, T$. Note that this argument does not apply in general if we extend the delivery of the asset to two or more periods, since the interest r can vary stochastically from period to period.

The development so far relies critically on the finite information tree and the arbitrary use of a state-price density. We now introduce an equivalent but more direct EMM definition that extends readily to infinite state-space versions of the theory. $\mathcal{Q}$ denotes the set of all full-support probability measures on 2^{Ω}.

DEFINITION 2.4.2. A **discount process** is a predictable strictly positive process ρ such that $\rho_0 = 1$. An **equivalent martingale measure**[7] **(EMM)** is a probability $Q \in \mathcal{Q}$ relative to which there exists a predictable SPD. An **EMM-discount pair** is a $(Q, \rho) \in \mathcal{Q} \times \mathcal{P}$ such that ρ is both a discount process and an SPD relative to Q. The present-value function **represented** by this pair is defined by $\Pi(c) = \mathbb{E}^{Q}\left[\sum_{t=0}^{T} \rho_t c_t\right]$, $c \in \mathcal{L}$.

Last section's results on pricing using state-price densities all apply to an EMM-discount pair (Q, ρ), with simplifications resulting from the fact that ρ is predictable. For example, if an MMA with rate process r is traded, then the pricing equation (2.3.6) with P replaced by Q and

[7]The terms "equivalent" and "martingale measure" have their origin in probability theory. A probability (measure) Q is said to be **equivalent** to P if $Q(A) = 0 \iff P(A) = 0$ for every event A. In our setting, for any $P \in \mathcal{Q}$, the set of all equivalent-to-P probabilities on 2^{Ω} is $\mathcal{Q}$. A **martingale measure** is a probability relative to which a given set of processes is a set of martingales. In the current context, such a set is that of the properly discounted gain processes of all traded contracts, as discussed at the end of this section.

π replaced by ρ results in $1/(1+r_t) = \rho_t/\rho_{t-1}$ and therefore

$$r = -\frac{\Delta\rho}{\rho} \quad \text{and} \quad \rho_t = \prod_{u=0}^{t} \frac{1}{1+r_u}, \quad t = 1, \ldots, T. \tag{2.4.4}$$

In general, given discount process ρ, we refer to the $r \in \mathcal{P}_0$ defined by the first equation in (2.4.4) as the **rate process implied** by ρ, and given $r \in \mathcal{P}_0$, we refer to the ρ defined by $\rho_0 = 1$ and the second equation in (2.4.4) as the **discount process implied** by r. The preceding discussion shows

PROPOSITION 2.4.3. *If r is the market's short-rate process and (Q, ρ) is an EMM-discount pair, then r is the rate process implied by ρ.*

Whether an MMA is traded or not, an EMM-discount pair (Q, ρ) prices a traded contract according to (2.4.3), where r is the rate process implied by ρ, as can be seen by formally replacing P by Q and π by ρ in equation (2.3.4). Last section's other pricing relationships can similarly be stated in terms of an EMM-discount pair.

Let us now establish a correspondence between SPDs relative to P and EMM-discount pairs, defined by the requirement that they represent the same present-value function. Not surprisingly, this correspondence is the same as that of equations (2.4.1) and (2.4.2), thus establishing the equivalence of the EMM notion of Definition 2.4.2 and that of this section's opening discussion. As preparation, we review two generally useful probabilistic constructs, starting with a change-of-measure formula.

Associated with every probability $Q \in \mathcal{Q}$ is the **density**[8] dQ/dP, which is the strictly positive unit-mean random variable

$$\frac{dQ}{dP}(\omega) = \frac{Q(\{\omega\})}{P(\{\omega\})}, \quad \omega \in \Omega,$$

and the **conditional density process**

$$\xi_t = \mathbb{E}_t\left[\frac{dQ}{dP}\right], \quad t = 0, 1, \ldots, T. \tag{2.4.5}$$

Since $\mathcal{F}_T = 2^\Omega$, $\xi_T = dQ/dP$ and therefore ξ determines Q and conversely. By the law of iterated expectations (2.1.5), ξ is a unit-mean strictly positive martingale (under P).

LEMMA 2.4.4. *For every $Q \in \mathcal{Q}$ and adapted process x,*

$$\mathbb{E}^Q[x_t] = \mathbb{E}[\xi_t x_t], \quad t = 0, \ldots, T. \tag{2.4.6}$$

PROOF. For every random variable z, we have

$$\mathbb{E}^Q z = \sum_{\omega\in\Omega} z(\omega)\, Q(\{\omega\}) = \sum_{\omega\in\Omega} \frac{Q(\{\omega\})}{P(\{\omega\})} z(\omega)\, P(\{\omega\}) = \mathbb{E}\left[\frac{dQ}{dP} z\right].$$

[8] Also known as the **Radon-Nikodym derivative** of Q with respect to P.

The law of iterated expectations and the fact that $x_t \in L_t$ can now be used to argue that

$$\mathbb{E}^Q x_t = \mathbb{E}\mathbb{E}_t\left[\frac{dQ}{dP}x_t\right] = \mathbb{E}\left[\mathbb{E}_t\left[\frac{dQ}{dP}\right]x_t\right] = \mathbb{E}\left[\xi_t x_t\right].$$

□

REMARK 2.4.5. For every spot (F,t), the last lemma with $x_t = 1_F$ implies that $Q(F) = \mathbb{E}^Q 1_F = \mathbb{E}[\xi_t 1_F] = \xi(F,t)P(F)$. Therefore, the value $\xi(F,t)$ that ξ_t takes on the event F is the likelihood ratio $Q(F)/P(F)$, which can be thought of as the ratio of the Q-probability of the path on the information tree leading from time zero to spot (F,t) to the P-probability of the same path. The random variable $\xi_T = dQ/dP$ has the same interpretation for terminal spots, which are of the form $(\{\omega\},t)$, $\omega \in \Omega$.

REMARK 2.4.6. We saw that a $Q \in \mathcal{Q}$ defines the density dQ/dP, which is strictly positive and satisfies $\mathbb{E}[dQ/dP] = 1$. Conversely, a strictly positive random variable Z such that $\mathbb{E}Z = 1$ defines a unique $Q \in \mathcal{Q}$ such that $Z = dQ/dP$; it is given by $Q(F) = \mathbb{E}[Z1_F]$ for every event A.

The second probabilistic construction we need to relate SPDs to EMMs is the following multiplicative version of the Doob decomposition. $\mathcal{P}_{++}$ denotes the set of all strictly positive predictable processes, and $\mathcal{M}$ denotes the set of all martingales under P.

LEMMA 2.4.7. *Every strictly positive adapted process π admits a unique decomposition of the form*

$$\pi = \pi_0 \rho \xi, \quad \rho \in \mathcal{P}_{++}, \quad \xi \in \mathcal{M}, \quad \xi_0 = 1. \tag{2.4.7}$$

PROOF. Given $\pi \in \mathcal{L}_{++}$, let $\rho \in \mathcal{P}_{++}$ be defined recursively by

$$\rho_0 = 1 \quad \text{and} \quad \frac{\rho_t}{\rho_{t-1}} = \frac{\mathbb{E}_{t-1}\pi_t}{\pi_{t-1}}, \quad t = 1,\ldots,T. \tag{2.4.8}$$

Define the process ξ so that $\pi = \pi_0\rho\xi$ and therefore $\xi_0 = 1$. The second equality in (2.4.8) is equivalent to the martingale property $\xi_{t-1} = \mathbb{E}_{t-1}\xi_t$. This proves the existence of decomposition (2.4.7) and also its uniqueness, since the martingale property of ξ implies that ρ must be given by (2.4.8). □

The last two lemmas together reveal the relationship between an SPD π representing a present-value function Π and a corresponding EMM-discount pair (Q,ρ) also representing Π: If $dQ/dP = \xi_T$ and ρ and ξ are as in decomposition (2.4.7), then for every cash flow c,

$$\Pi(c) = \frac{1}{\pi_0}\mathbb{E}\left[\sum\nolimits_{t=0}^{T}\pi_t c_t\right] = \mathbb{E}\left[\sum\nolimits_{t=0}^{T}\rho_t\xi_t c_t\right] = \mathbb{E}^Q\left[\sum\nolimits_{t=0}^{T}\rho_t c_t\right].$$

Here is a more detailed statement of this conclusion.

PROPOSITION 2.4.8. *Suppose π is an SPD representing the present-value function Π. Let ρ be defined by (2.4.8), and let*

$$Q(F) = \mathbb{E}[\xi_T 1_F], \ F \in \mathcal{F}_T, \quad \text{where} \quad \xi_T \equiv \frac{1}{\rho_T}\frac{\pi_T}{\pi_0}. \tag{2.4.9}$$

Then (Q, ρ) is an EMM-discount pair representing Π. Conversely, suppose (Q, ρ) is an EMM-discount pair representing the present-value function Π. Let ξ denote the conditional density process of Q relative to P, and let π_0 be an arbitrary positive scalar. Then $\pi = \pi_0\rho\xi$ is an SPD representing Π.

COROLLARY 2.4.9. *For every present-value function Π, there exists a unique Q in $\mathcal{Q}$ such that (Q, ρ) is an EMM-discount pair representing Π for some (necessarily unique) discount process ρ.*

The relationship $\pi = \pi_0\rho\xi$ of the preceding proposition connects the EMM notion of Definition 2.4.2 to the earlier construction of equations (2.4.1) and (2.4.2). To see how, consider any nonterminal spot $(F, t-1)$ with immediate successor spots (F_i, t), $i = 0, \ldots, d$. Bayes' rule and Remark 2.4.5 imply

$$Q(F_i \mid F) = \frac{Q(F_i)}{Q(F)} = \frac{\xi(F_i, t)}{\xi(F, t-1)}, \quad i = 0, \ldots, d. \tag{2.4.10}$$

Therefore, if r is the rate process implied by ρ, then equation (2.4.1) applied to all nonterminal spots is the same as $\pi/\pi_- = (\rho\xi)/(\rho\xi)_-$, which is an equivalent recursive expression of the condition $\pi = \pi_0\rho\xi$.

At the heart of the connection between pricing in terms of an SPD and pricing in terms of an EMM is a probabilistic change-of-measure formula, such as Lemma 2.4.4. We now review two other versions of the change-of-measure formula, each offering another insight on the relationship between pricing in terms of SPDs and EMMs.

We have encountered the pricing of risk in terms of covariances in equation (2.3.7), and in terms of an expectation under an EMM in equation (2.4.3). The connection between the two is made directly by the following conditional change-of-measure formula, which on a finite information tree is an immediate consequence of equation (2.4.10). The result applies in more general stochastic settings, however, where an expression like (2.4.10) is not meaningful. The following alternative argument applies to such settings and offers a chance to practice formal properties of conditional expectations.

LEMMA 2.4.10. *For every $Q \in \mathcal{Q}$, $V \in \mathcal{L}$ and time $t > 0$,*

$$\mathbb{E}^Q_{t-1} V_t = \mathbb{E}_{t-1}\left[\frac{\xi_t}{\xi_{t-1}} V_t\right] = \mathbb{E}_{t-1} V_t + \operatorname{cov}_{t-1}\left[\frac{\xi_t}{\xi_{t-1}}, V_t\right], \tag{2.4.11}$$

where ξ is the conditional density process of Q relative to P.

PROOF. Let $y = \mathbb{E}_{t-1}[\xi_t V_t]/\xi_{t-1}$. For all $z \in L_{t-1}$, Lemma 2.4.4, the conditional expectation properties of Proposition 2.1.6 and the fact that $yz \in L_{t-1}$ imply

$$\begin{aligned}\mathbb{E}^Q[yz] &= \mathbb{E}[\xi_t yz] = \mathbb{E}[\mathbb{E}_{t-1}[\xi_t yz]] = \mathbb{E}[\mathbb{E}_{t-1}[\xi_t]\, yz] = \mathbb{E}[\xi_{t-1} yz] \\ &= \mathbb{E}[\mathbb{E}_{t-1}[\xi_t V_t]\, z] = \mathbb{E}[\mathbb{E}_{t-1}[\xi_t V_t z]] = \mathbb{E}[\xi_t V_t z] = \mathbb{E}^Q[V_t z].\end{aligned}$$

By Proposition 2.1.3, it follows that $y = \mathbb{E}^Q_{t-1} V_t$. The second equation in (2.4.11) follows from the definition of the conditional covariance and the fact that $\mathbb{E}_{t-1}\xi_t/\xi_{t-1} = 1$. □

Now insert the SPD factorization $\pi = \pi_0 \rho \xi$, where ρ is a discount process with implied rate process r and ξ is a strictly positive unit-mean martingale, in the pricing equation (2.3.7) to find

$$S_{t-1} = \frac{1}{1+r_t}\left(\mathbb{E}_{t-1} V_t + \operatorname{cov}_{t-1}\left[\frac{\xi_t}{\xi_{t-1}}, V_t\right]\right). \tag{2.4.12}$$

This is the same as equation (2.4.3) if Q is the probability defined by $\xi_T = dQ/dP$, since the term in parentheses is equal to $\mathbb{E}^Q_{t-1} V_t$ by the Lemma just shown.

Yet another way in which state-pricing is expressed is through the martingale property of properly discounted gain processes. The connection between EMMs and SPDs in those terms is made through a probabilistic result on the martingale property after a change of measure reviewed below. For simplicity, let us assume that the market X is implemented by $1+J$ contracts, where contract zero is an MMA, just as described in Section 1.7, whose notation we adopt here. Consider any $\pi \in \mathcal{L}_{++}$ with decomposition (2.4.7), where ξ is the conditional density process of $Q \in \mathcal{Q}$. As an SPD, π prices the MMA if and only if the discount process ρ and the short rate process r are related by (2.4.4), a condition we henceforth assume. By Proposition 2.3.3, π is an SPD if and only if it prices the remaining J contracts with underlying probability P, if and only if ρ prices the same J contracts with underlying probability Q. Therefore, by Proposition 2.3.4, π is an SPD if and only if $G^\pi \equiv \pi V + (\pi\delta)_- \bullet \mathbf{t}$ is a martingale (under P); and Q is an EMM if and only if[9] $G^\rho \equiv \rho V + (\rho\delta)_- \bullet \mathbf{t}$ is a **Q-martingale**; that is, a martingale under Q. The fact that π is an SPD if and only if Q is an EMM corresponds to the purely probabilistic fact that G^π is a P-martingale if and only if G^ρ is a Q-martingale, which can be shown directly as a corollary of Lemma 2.4.10.

[9] It is this equivalence that justifies the term "martingale measure" for Q. The "equivalent" qualification comes from probability theory, where two probabilities are said to be **equivalent** if they vanish on exactly the same events. Since P is assumed to have full support, $\mathcal{Q}$ is the set of probabilities that are equivalent to P. In infinite state-space settings, the definition of full support does not apply as given for the finite case. For example, under the uniform distribution on the unit interval, every finite or countable event has zero probability.

EXAMPLE 2.4.11. (American call) This example illustrates how some basic facts from martingale theory can provide a dual argument showing the American call late exercise result of Example 1.5.5. Throughout this example, we restrict the time horizon to be the American call maturity $\bar{\tau}$ (rather than T), and we let $\bar{\mathcal{T}}$ denote the set of every stopping time valued in $\{0,\ldots,\bar{\tau}\}$. From Section 2.1, recall that every adapted process has a unique Doob decomposition, whose predictable part defines its compensator. A **submartingale** is an adapted process x with an **increasing** compensator x^p, meaning $\Delta x_t^p \geq 0$ or, equivalently, $\mathbb{E}_{t-1}\Delta x_t \geq 0$ (at every state and time $t>0$). We will use two basic facts about a submartingale x.

Fact 1: *$\tau \in \bar{\mathcal{T}}$ implies $\mathbb{E}x_{\bar{\tau}} \geq \mathbb{E}x_{\tau}$.*

PROOF. Define the martingale $M \equiv x - x^p$ and given any $\tau \in \bar{\mathcal{T}}$, the process $\phi_t = 1_{\{\tau<t\leq\bar{\tau}\}}$ (with $1_{\emptyset}\equiv 0$). Since ϕ is predictable, $\phi \bullet M$ is a zero-mean martingale and $0 = \mathbb{E}\left(\phi \bullet M\right)_T = \mathbb{E}\left[M_{\bar{\tau}} - M_{\tau}\right]$. Therefore $\mathbb{E}M_{\bar{\tau}} = \mathbb{E}M_{\tau}$ and, since x^p is increasing, $\mathbb{E}x_{\bar{\tau}} \geq \mathbb{E}\left[M_{\bar{\tau}} + x_{\tau}^p\right] = \mathbb{E}x_{\tau}$. □

Fact 2: *Suppose the function $f:\mathbb{R}\to\mathbb{R}$ is increasing and convex. Then $f(x)$ is also a submartingale.*

PROOF. The submartingale property $\mathbb{E}_{t-1}\Delta x_t \geq 0$ and the assumption that f is increasing imply $f\left(\mathbb{E}_{t-1}x_t\right) \geq f\left(x_{t-1}\right)$. The claim is therefore a corollary of **Jensen's inequality:** $\mathbb{E}_{t-1}f\left(x_t\right) \geq f\left(\mathbb{E}_{t-1}x_t\right)$. Using only the convexity of f, note that the right derivative of f at $\mathbb{E}_{t-1}x_t$ defines $d_{t-1}\in L_{t-1}$ as the slope of a supporting line of the graph of f at $\mathbb{E}_{t-1}x_t$ and therefore $f\left(x_t\right) \geq f\left(\mathbb{E}_{t-1}x_t\right) + d_{t-1}\left(x_t - \mathbb{E}_{t-1}x_t\right)$. Applying $\mathbb{E}_{t-1}$ on both sides results in Jensen's inequality. □

Consider now the American call of Example 1.5.5 with the underlying stock paying no dividends up to $\bar{\tau}$. Assume also that there is a traded MMA with a non-negative interest rate process, which means that the corresponding (predictable) discount process ρ is decreasing ($s>t$ implies $\rho_s \leq \rho_t$). For every EMM Q, we will show that $\bar{\tau}$ maximizes the corresponding present value:

$$\mathbb{E}^Q\left[\rho_{\bar{\tau}}\left(S_{\bar{\tau}}-K\right)^+\right] = \max_{\tau\in\bar{\mathcal{T}}}\mathbb{E}^Q\left[\rho_{\tau}\left(S_{\tau}-K\right)^+\right].$$

By Theorem 1.5.2, this proves that *not* exercising the call prior to maturity is a dominant choice. Fixing the reference EMM Q, let us select the underlying probability P to equal Q. The no-dividends assumption implies that ρS is a martingale (up to the assumed time-horizon $\bar{\tau}$), and since ρK is a predictable decreasing process, $\rho S - \rho K$ is a submartingale. The function that takes the positive part is convex and increasing. By Fact 2, $x \equiv \rho\left(S-K\right)^+$ is a submartingale, and by Fact 1, $\mathbb{E}x_{\bar{\tau}} \geq \mathbb{E}x_{\tau}$. ♢

2.5. Predictable representations

In linear algebra it is often convenient to introduce a linear basis and refer to vectors by their representation relative to the reference basis. In what is essentially the same idea applied at each node of the information tree, in this section, we introduce processes representing risk sources that span all uncertainty. We show how every adapted process can be represented in terms of its exposure to these risk sources, known as volatility, and we discuss how the market prices volatility. The resulting scaffolding is especially useful as we pass from high-frequency models to the continuous-time limit. Unless otherwise indicated, expectations, variances, covariances and the martingale property are all relative to a given underlying full-support probability P. Recall that $\mathcal{M}_0$ denotes the set of zero-mean martingales and $\mathcal{P}_0$ denotes the set of predictable processes that take the value zero at time zero.

The essential idea behind predictable representations is easiest to see in the one-period case, with a single time-zero spot and $1+d$ time-one spots. A random variable in this case is any vector in $\mathbb{R}^{1+d}$. Let the column vector $b = \left(b^1, \ldots, b^d\right)'$ list the elements of an orthonormal basis of the linear subspace of all zero-mean random variables with the covariance inner product. By construction, the b^i are uncorrelated to each other and they each have zero mean and unit variance. In this simple case, an $M \in \mathcal{M}_0$ can be represented as $M_1 = \Delta M_1 = \sigma b$ for some row vector $\sigma \in \mathbb{R}^{1\times d}$, necessarily given by $\sigma = \mathbb{E}\left[\Delta M_1 b'\right]$.

The same construction can be applied over the single period following any non-terminal spot on the filtration $\{\mathcal{F}_t\}$, for any time horizon T. For simplicity, we assume that every nonterminal spot has exactly $1+d$ immediate successor spots. We call such a filtration **uniform** with **spanning number** $1+d$. Focusing on any nonterminal spot $(F, t-1)$ and its immediate successor spots (F_i, t), $i = 0, \ldots, d$, we apply the earlier single-period construction conditionally on F. Adding the subscript F, t, we end up with the vector $b_{F,t} = (b^1_{F,t}, \ldots, b^d_{F,t})'$, where the $b^i_{F,t}$ are random variables taking the value zero outside F and such that $\mathbb{E}[\, b^i_{F,t} \mid F\,] = 0$, $\mathbb{E}[\,(b^i_{F,t})^2 \mid F\,] = 1$, and $\mathbb{E}[\, b^i_{F,t} b^j_{F,t} \mid F\,] = 0$ for $i \neq j$. For every $M \in \mathcal{M}_0$, $\Delta M_t = \sigma_{F,t} b_{F,t}$ on the event F, where $\sigma_{F,t} = \mathbb{E}[\,\Delta M_t b'_{F,t} \mid F\,]$. The bases $b_{F,t}$ are conveniently stitched together by defining the process $B = (B^1, \ldots, B^d)'$, where $B^i_0 = 0$ and $\Delta B^i_t 1_F = b^i_{F,t}$ for every nonterminal spot $(F, t-1)$.

The fact that the $b^i_{F,t}$ have conditionally zero mean is equivalent to

$$B \in \mathcal{M}_0^d.$$

The conditional second moment conditions satisfied by the $b^i_{F,t}$ are equivalent to the condition that B is **dynamically orthonormal:**

$$\mathbb{E}_{t-1}\left[\Delta B^i_t \Delta B^j_t\right] = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \qquad t = 1, \ldots, T. \tag{2.5.1}$$

The spot-by-spot representations of the increments of a martingale M become

$$M = \sigma \bullet B, \quad \sigma \in \mathcal{P}_0^{1\times d}, \tag{2.5.2}$$

where necessarily $\sigma_t = \mathbb{E}_{t-1}\left[\Delta M_t \Delta B_t'\right]$. Expression (2.5.2) is known as the **predictable martingale representation** of M (relative to B), and B is said to have the **martingale representation property**. Combining the latter with the Doob decomposition of an adapted process, as defined in Section 2.1, we have the following representation result.

PROPOSITION 2.5.1. *The underlying filtration is uniform with spanning number $1+d$ if and only if it is generated by a $B \in \mathcal{M}_0^d$ satisfying the conditional orthonormality condition (2.5.1). In this case every adapted process x can be uniquely represented in the form*

$$x = x_0 + \mu \bullet \mathbf{t} + \sigma \bullet B, \quad \mu \in \mathcal{P}_0, \quad \sigma \in \mathcal{P}_0^{1\times d}. \tag{2.5.3}$$

The processes μ and σ can be computed in terms of x by

$$\mu_t = \mathbb{E}_{t-1}\left[\Delta x_t\right] \quad \textit{and} \quad \sigma_t = \mathbb{E}_{t-1}\left[\Delta x_t \Delta B_t'\right], \quad t = 1, \ldots, T. \tag{2.5.4}$$

PROOF. We have already shown the existence of a conditionally orthonormal $B \in \mathcal{M}_0^d$ with the martingale representation property. Representation (2.5.3) results from the Doob decomposition of x with $\mu = \Delta x^p$. To show that B generates the underlying filtration, we must show that for every time t, the realization of $B_1, \ldots, B_t$ reveals the realized time-t spot. We argue by induction. For $t = 0$, the claim is vacuously true. Suppose it is true for $t-1$ and ω is the realized state, corresponding to spots $(F, t-1)$ and (G, t), where $\omega \in G \subseteq F$. The inductive hypothesis means that $(F, t-1)$ is the unique spot that is consistent with $B_s(\omega)$ for $s < t$. Let x be the adapted process that takes the value zero everywhere except for $x(G, t) = 1$. Representation (2.5.3) of x means that we can write $\Delta x_t = \mu_t + \sigma_t \Delta B_t$, where μ_t and σ_t take a constant value on F, revealed by $B_s(\omega)$ for $s < t$. If in addition we know $B(G, t)$, then the value $x(G, t)$ is also revealed, and hence the identity of the spot (G, t). This completes the inductive proof that B generates the underlying filtration. Finally, the necessity of the condition that each nonterminal spot has exactly $d+1$ immediate successor spots follows by a dimensionality argument. Consider any spot $(F, t-1)$ with immediate successors $(G_0, t), \ldots, (G_k, t)$. By representation (2.5.3), the set of $\mathcal{F}_t$-measurable random variables that take the value zero outside F and have zero conditional expectation given F is a d-dimensional linear space. The same linear space can be identified with the set of every vector $(x(G_0, t), \ldots, x(G_k, t))$ satisfying $\sum_i x(G_i, t) P[G_i \mid F] = 0$, which is k-dimensional. Therefore $k = d$. □

We henceforth fix a reference $B \in \mathcal{M}_0^d$ having the properties of the preceding proposition. We call such a martingale a **dynamically orthonormal basis** of $\mathcal{M}_0$ and the representation (2.5.3) as the **predictable representation** or the **dynamics** of x (relative to B). The processes μ and σ are, respectively, the **drift** and **volatility** of x. Clearly, an adapted process is a martingale if and only its drift is zero.

Consider now a state price density π and corresponding decomposition $\pi = \pi_0 \rho \xi$, where ρ is a (predictable) discount process and ξ is the conditional density process of the EMM Q associated with π. We are interested in the predictable representation of π, how it relates to the dynamics of ρ and ξ, and finally how it can be used to price a traded contract. We saw in the last section that if there is a traded MMA then the short rate process is $r = -\Delta\rho/\rho$. Let $r \in \mathcal{P}_0$ be defined by this equation, whether an MMA is traded or not. Simple algebra shows that

$$\frac{\Delta\pi}{\pi_-} = \frac{\rho}{\rho_-}\left(\frac{\Delta\rho}{\rho} + \frac{\Delta\xi}{\xi_-}\right) = -\frac{1}{1+r}\left(r - \frac{\Delta\xi}{\xi_-}\right).$$

The last term leads us to the process $(1/\xi_-) \bullet \xi$, which is a zero-mean martingale (why?). Define the process η by the corresponding unique predictable representation

$$\frac{1}{\xi_-} \bullet \xi = -\eta' \bullet B, \quad \eta \in \mathcal{P}_0^d. \tag{2.5.5}$$

The state-price dynamics can then be expressed as

$$\frac{\Delta\pi}{\pi_-} = -\frac{1}{1+r}\left(r + \eta' \Delta B\right). \tag{2.5.6}$$

By Proposition 2.5.1 and Lemma 2.4.10,

$$\eta_t = -\mathbb{E}_{t-1}\left[\frac{\xi_t}{\xi_{t-1}}\Delta B_t\right] = -\mathbb{E}_{t-1}^Q\left[\Delta B_t\right]. \tag{2.5.7}$$

Equivalently,

$$B + \eta \bullet \mathbf{t} \text{ is a } Q\text{-martingale.} \tag{2.5.8}$$

This conclusion is an instance of what in a more general setting is known as the *Girsanov-Lenglart theorem*.

Conversely, the process η determines the probability Q, since representation (2.5.5) can be restated as the recursion $\xi/\xi_- = 1 - \eta'\Delta B$, which can be multiplied through, starting with $\xi_0 = 1$, to obtain

$$\xi_t = \prod_{s=0}^{t}\left(1 - \eta_s' \Delta B_s\right), \quad t = 0, 1, \ldots, T. \tag{2.5.9}$$

In particular, η determines $\xi_T = dQ/dP$ and hence Q. Not every $\eta \in \mathcal{P}_0^d$ defines a $Q \in \mathcal{Q}$ in this manner, however, since ξ must be

strictly positive. In fact, the preceding correspondence between η and Q defines a bijection between $\mathcal{Q}$ and the set

$$\mathcal{H} = \left\{\eta \in \mathcal{P}_0^d : 1 - \eta' \Delta B \in \mathcal{L}_{++}\right\}. \tag{2.5.10}$$

In equations (2.4.4) of the last section, we saw that every rate process $r \in \mathcal{P}_0$ implies a unique discount process ρ (Definition 2.4.2) and vice versa. The preceding construction shows that every $Q \in \mathcal{Q}$ defines a unique $\eta \in \mathcal{H}$ by letting $\eta_t = -\mathbb{E}_{t-1}^Q[\Delta B_t]$, and conversely, every $\eta \in \mathcal{H}$ defines a $Q \in \mathcal{Q}$ where $dQ/dP = \xi_T$ is defined by (2.5.9). With $Q \in \mathcal{Q}$ and $\eta \in \mathcal{H}$ so related, we refer to η as the **predictable representation** of Q, and to Q as the probability **represented** by η. We also say that the pair (Q, ρ) **implies** a pair $(\eta, r) \in \mathcal{H} \times \mathcal{P}_0$ and vice versa.

DEFINITION 2.5.2. The process $\eta \in \mathcal{H}$ is a **market price of risk** processes if it represents an EMM Q, in which case, we call η the market price of risk process **implied** by the EMM Q.

In the last two sections, we saw a number of equivalent expressions of what it means for an SPD π or corresponding EMM-discount pair (Q, ρ) to price a given contract (δ, V). One of those is that $G^\rho = \rho V + (\rho\delta)_- \bullet \mathbf{t}$ is a Q-martingale. Let us now formulate the same condition in terms of the implied $(\eta, r) \in \mathcal{H} \times \mathcal{P}_0$ and the gain process predictable representation

$$G \equiv V + \delta_- \bullet \mathbf{t} \equiv V_0 + \mu^G \bullet \mathbf{t} + \sigma^G \bullet B. \tag{2.5.11}$$

A direct calculation of the increment ΔG^ρ shows that

$$G^\rho = \left(\rho\left(\mu^G - rS_- - \sigma^G\eta\right)\right) \bullet \mathbf{t} + \left(\rho\sigma^G\right) \bullet \left(B + \eta \bullet \mathbf{t}\right).$$

By (2.5.8), G^ρ is a Q-martingale if and only if the first term, which is the drift of G^ρ under Q, is zero. Summarizing, we have shown that (Q, ρ) prices the contract (δ, V) if and only if

$$\mu^G = rS_- + \sigma^G\eta. \tag{2.5.12}$$

In words, the conditional expected gain over one period is the risk-free interest on the beginning-of-period ex-dividend value plus a risk adjustment that is measured by the volatility σ^G of the gain process priced by η.

We arrived at the pricing condition (2.5.12) through the martingale property of G^ρ under Q, an approach that has the advantage of generalizing to continuous-time formulations. The reader may wish to explore more direct links between the pricing of a contract and equation (2.5.12), for example, through the pricing condition (2.4.12). In general, the pricing of a contract corresponds to a backward recursion on the information tree, corresponding to the requirement that current value be equal to the present value of all future dividends up to some time plus the present value of the contract value at that time. That

condition (2.5.12) represents a backward recursion becomes apparent once we note that $G_{t-1} = \mathbb{E}_{t-1} G_t - \mu_t^G$ and $\sigma_t^G = \mathbb{E}_{t-1}[G_t \Delta B_t]$. Thus knowing the end-of-period value G_t gives us σ_t^G, which in turn gives us μ_t^G by (2.5.12), and finally gives us G_{t-1}. In general, any condition that specifies the drift as a function of the volatility represents a backward recursion. This simple insight is particularly useful in gaining intuition behind continuous-time formulations, where the notion of a single-period backward recursion is not directly available.

Assuming S_{t-1} is nonzero everywhere, pricing condition (2.5.12) is often expressed in terms of the return dynamics

$$\frac{\Delta G_t}{S_{t-1}} = \mu_t^R + \sigma_t^R \Delta B_t, \quad \mu^R \in \mathcal{P}_0, \quad \sigma^R \in \mathcal{P}_0^{1\times d}. \tag{2.5.13}$$

In this case, equation (2.5.12) can be restated as

$$\mu^R - r = \sigma^R \eta, \tag{2.5.14}$$

which can more directly be viewed as a corollary of the expected excess returns expression (2.3.8) and the SPD dynamics (2.5.6). Thus (conditional) expected excess returns relative to the risk-free rate over a single period are explained by exposure to the risk sources ΔB_t as measured by the return volatility σ^R. In the literature, this type of pricing is also known as *factor pricing,* with ΔB_t being the (risk) *factors,* the volatility representing (conditional) *factor loadings,* and η_t representing the *factor prices.*

In arbitrage-pricing applications it is common to assume that the market is implemented by exogenously specified contracts. Let us adopt the setting of Section 1.7, where X is implemented by $1 + J$ contracts, with contract zero being an MMA defining the short-rate process r and implied discount process ρ. Modifying our earlier notation, let $\delta = \left(\delta^1, \ldots, \delta^J\right)'$ and define $V, G \in \mathcal{L}^J$ analogously, with G having dynamics (2.5.11), where $\mu^G \in \mathcal{P}_0^J$ and $\sigma^G \in \mathcal{P}_0^{J\times d}$. Given $Q \in \mathcal{Q}$ with predictable representation $\eta \in \mathcal{H}$, we know that Q is an EMM if and only if it prices contracts 1 through J, a condition that is in turn equivalent to η satisfying equation (2.5.12). Therefore the market is arbitrage-free if and only if (2.5.12) admits a solution $\eta \in \mathcal{H}$, and such a solution is unique if and only if the market is complete. Of course, the computational details of the determination of η depends on the details of the specification, which typically includes a Markovian structure, as discussed in the following chapter.

Remark 2.5.3. The existence of some $\eta \in \mathcal{P}_0$ satisfying equation (2.5.12) does not exclude arbitrage opportunities; the positivity part of the definition of $\mathcal{H}$ is essential. The existence of of some $\eta \in \mathcal{P}_0$ satisfying equation (2.5.12) is equivalent to the weaker condition

$$\text{for all } \theta \in \mathcal{P}_0^{1\times J}, \quad \theta\sigma^G = 0 \implies \theta\left(\mu^G - rS_-\right) = 0. \tag{2.5.15}$$

To see the necessity of this condition, use an orthogonal decomposition at each spot to write $\mu^G - rS_- = \sigma^G\eta + \varepsilon'$, for some $\eta \in \mathcal{P}_0^d$ and $\varepsilon \in \mathcal{P}_0^{1\times J}$ such that $\varepsilon\sigma^G = 0$ and therefore $\varepsilon\left(\mu^G - rS_-\right) = 0$. Since $\varepsilon\left(\mu^G - rS_-\right) = \varepsilon\varepsilon'$, it follows that $\varepsilon = 0$, proving (2.5.15). The converse claim is immediate. Condition (2.5.15) states that every trading strategy with zero volatility must produce the same returns as the MMA. In other words, there is no MMA arbitrage. $\diamondsuit$

For a typical arbitrage-pricing application, consider a market maker who can trade in the $1 + J$ contracts implementing the market X, and sells contract (δ^*, V^*) to a customer at time zero. We assume that X is arbitrage-free and contract (δ^*, V^*) is synthetic in X, meaning that there exists a trading strategy (θ^0, θ) such that $(\delta^*, V^*) = \left(\delta^\theta, V^\theta\right)$. (By Proposition1.6.8, the contract is synthetic if and only if it is traded.) The customer could in principle bypass the market maker by following trading strategy (θ^0, θ), but the idea is that the customer is less qualified to carry out the necessary transactions, for example, because of lack of sophistication, market access or time. In a perfectly competitive, frictionless market-making market, the market maker charges V_0^* for the contract at time zero[10] and delivers the cash flow δ^*. The market maker entirely hedges the liability risk resulting from this sale by purchasing back the synthetic contract $\left(\delta^\theta, V^\theta\right)$. The market maker's pricing model consists of the short-rate process r and the gain or return dynamics specified earlier, which imply a market price of risk process η and associated EMM Q. This data can be used to recursively compute V^* as the present value process of δ^*. Given V^*, the market maker can compute the hedging strategy in two steps. The position in the MMA is dictated by the replication condition $V^* = V^\theta$, which is the fact that the (cum-dividend) value in the MMA plus the value of the remaining positions in the J contracts must equal the value of the sold contract. Therefore,

$$\theta_t^0 = \frac{V_t^* - \theta_t V_t}{1 + r_t}, \quad t = 1, \ldots, T; \quad \theta_0^0 = 0. \tag{2.5.16}$$

The remaining trading strategy θ is computed to eliminate volatility period by period. To see why, define the gain-process predictable representation

$$G^* \equiv V^* + \delta_-^* \bullet \mathbf{t} \equiv V_0^* + \mu^* \bullet \mathbf{t} + \sigma^* \bullet B \tag{2.5.17}$$

and use the budget equation in the form of Proposition 1.6.3 to find

$$V^* - \theta V = \text{ predictable term } + \left(\sigma^* - \theta\sigma^G\right) \bullet B. \tag{2.5.18}$$

[10]In practice, the market maker would charge a small spread over V_0^* reflecting trading and other business costs and possibly barriers to entry by other market makers. The assumption here is that for any individual trade such a spread is negligible, although the same spread multiplied by a large number of transactions may not be negligible.

By (2.5.16), $V^* - \theta V$ is predictable and therefore

$$\sigma^* = \theta \sigma^G, \tag{2.5.19}$$

which is assumed to have a solution in θ. (In the simplest formulation, σ^G is invertible.)

Assuming the market is arbitrage-free, a corollary of the preceding construction is that every synthetic contract is uniquely generated by a trading strategy if and only if the rows of the $J \times d$ matrix $\sigma^G(\omega, t)$ are linearly independent for all $(\omega, t) \in \Omega \times \{1, \ldots, T\}$. In this case, we say that the contracts $(\delta^1, V^1), \ldots, \left(\delta^J, V^J\right)$ are **dynamically independent.** Another corollary, via Proposition 1.6.7, is that the market is complete if and only if the rank of $\sigma^G(\omega, t)$ is d for all $(\omega, t) \in \Omega \times \{1, \ldots, T\}$. In particular, the MMA (δ^0, V^0) together with the dynamically independent contracts $(\delta^1, V^1), \ldots, \left(\delta^J, V^J\right)$ implement a complete market if and only if $J = d$ and $\sigma^G(\omega, t)$ is invertible for all $(\omega, t) \in \Omega \times \{1, \ldots, T\}$. In this case, for every $j \in \{1, \ldots, J\}$, the contract (δ^j, V^j) is necessarily **everywhere risky,** meaning that for every time $t > 0$ and every nonempty event $F \in \mathcal{F}_t$, the random variable $V_t^j 1_F$ is not $\mathcal{F}_{t-1}$-measurable.

We close with an argument that leverages this section's apparatus to easily show that every complete market on the given filtration can be implemented by $1 + d$ contracts. In contrast, if we were to implement a complete market by forming portfolios at time zero only, as many contracts as there are non-time-zero spots would be required, a number that rises exponentially in T if $d > 0$, rendering the assumption of a perfectly competitive market implausible.

THEOREM 2.5.4. *Suppose the underlying filtration is uniform with spanning number* $1+d$. *An arbitrage-free market is complete if and only if it can be implemented by a money-market account and* d *dynamically independent everywhere risky contracts.*

PROOF. The "if" part follows from the preceding discussion. Conversely, suppose the market X is complete and arbitrage-free, and let (Q, ρ) be the corresponding unique EMM-discount pair implying the short-rate process r and market price of risk process η. Market completeness implies that a contract is traded (in X) if and only if it is priced by (Q, ρ) (why?). The MMA (δ^0, V^0) is defined by r in (1.7.3). Clearly, (δ^0, V^0) is priced by (Q, ρ) and is therefore traded. Define the everywhere risky contracts $(\delta, V) \in \mathcal{L}^{d\times 2}$ by letting $\delta_- = 0$ and $G^\rho = \rho V = B + \eta \bullet \mathbf{t}$. By (2.5.7), G^ρ is a Q-martingale and therefore, by Proposition 2.3.4, (Q, ρ) prices the contracts (δ, V), which are therefore traded. The market implemented by the contracts (δ^0, V^0) and (δ, V) is included in X and is complete and is therefore equal to X. □

2.6. Independent increments and the Markov property

On a uniform filtration with spanning number $1+d$, where $d>0$, the number of spots rises exponentially with the number of periods. A spot-by-spot backward recursion that determines some value V_t in terms of V_{t+1} may be conceptually trivial, but computationally infeasible as t increases due to an explosion in the number of spots. In empirical applications, another concern is that a parameter can depend in too many ways on the entire history path, leading to data overfitting. These issues are commonly handled by introducing a so-called Markovian structure, where the entire history is summarized by a relatively low-dimensional state vector.

We review the basic idea in last section's setting, where the underlying filtration is generated by the d-dimensional dynamically orthonormal basis B, where $d>0$, relative to the full-support probability P. We further assume, throughout this section, that B has **independent increments** (relative to P), meaning that for every time $t>0$ and function $f:\mathbb{R}^d\to\mathbb{R}$,

$$\mathbb{E}_{t-1}\left[f\left(\Delta B_t\right)\right]=\mathbb{E}\left[f\left(\Delta B_t\right)\right]. \tag{2.6.1}$$

Note that by Proposition 1.1.4 and Corollary 2.1.5, B has independent increments if and only if for every time $t>0$, the algebras $\sigma\left(\Delta B_t\right)$ and $\mathcal{F}_{t-1}=\sigma\left(\Delta B_1,\ldots,\Delta B_t\right)$ are stochastically independent. While condition (2.6.1) is all we need to apply the independent-increments property in this section, Proposition 2.1.2 explains the terminology: B has independent increments if and only if the random variables $\Delta B_1,\ldots,\Delta B_T$ are stochastically independent. Example 2.1.8 gives a prototypical independent increments process. As in that example, an independent increments process B such that $\mathbb{E}\Delta B_t=0$ for all $t>0$ is necessarily martingale, but an arbitrary martingale need not have independent increments.

Recall that a time-t spot corresponds to a specific realization of the entire history of B up to that spot. There are $(1+d)^t$ such histories. The idea is to hypothesize that at each time-t there is a k-dimensional random variable Z_t that summarizes all that is relevant of the history of B up to t, where k is a computationally manageable positive integer. Fixing an initial value $Z_0\in\mathbb{R}^k$, we assume that Z is the k-dimensional adapted process defined recursively by

$$\Delta Z_t=\mu_t^Z\left(Z_{t-1}\right)+\sigma_t^Z\left(Z_{t-1}\right)\Delta B_t,\quad t=1,\ldots,T, \tag{2.6.2}$$

for given functions $\mu_t^Z:\mathbb{R}^k\to\mathbb{R}^k$ and $\sigma_t^Z:\mathbb{R}^k\to\mathbb{R}^{k\times d}$. This construction guarantees that Z is a **Markov process** (relative to P), meaning that for every time $t>0$ and function $f:\mathbb{R}^k\to\mathbb{R}$,

$$\mathbb{E}_{t-1}f\left(Z_t\right)=\mathbb{E}\left[f\left(Z_t\right)\mid Z_{t-1}\right]. \tag{2.6.3}$$

The **Markov property** (2.6.3) formalizes the idea that every statistic of Z_t can be calculated knowing only Z_{t-1} rather than the entire path of B up to time $t-1$. Note that equation (2.6.3) holds if and only if $\mathbb{E}_{t-1} f(Z_t)$ is $\sigma(Z_{t-1})$-measurable.

The claim that Z is a Markov process relative to P is a corollary of the following more general result.

LEMMA 2.6.1. *Suppose B has independent increments relative to P, and $Q \in \mathcal{Q}$ has a conditional density process $\xi_t = \mathbb{E}_t[dQ/dP]$ that satisfies*

$$\frac{\Delta \xi_t}{\xi_{t-1}} = -\eta_t(Z_{t-1})' \Delta B_t, \quad \xi_0 = 1,$$

for some $\eta_t : \mathbb{R}^k \to \mathbb{R}^{d\times 1}$, $t = 1, \dots T$. Then Z is a Markov process relative to Q.

PROOF. Consider any time $t > 0$, vector $z \in \mathbb{R}^k$ and function $h : \mathbb{R}^k \to \mathbb{R}$. Using the conditional change of measure formula of Lemma 2.4.10 and recursion (2.6.2) for Z, note that on the event $\{Z_{t-1} = z\}$, the conditional expectation $\mathbb{E}^Q_{t-1}[h(Z_t)]$ is equal to

$$\mathbb{E}_{t-1}\left[\left(1 - \eta(z)_t' \Delta B_t\right) h\left(z + \mu^Z_t(z) + \sigma^Z_t(z) \Delta B_t\right)\right].$$

Since ΔB_t is independent of $\mathcal{F}_{t-1}$, the above quantity is constant on $\{Z_{t-1} = z\}$ and therefore so is $\mathbb{E}^Q_{t-1}[h(Z_t)]$. This proves that $\mathbb{E}^Q_{t-1}[h(Z_t)]$ is $\sigma(Z_{t-1})$-measurable. □

For each time t, the set of possible time-t Markov states is $\mathcal{N}_t \equiv \{Z_t(\omega) : \omega \in \Omega\}$. For $x \in \mathcal{L}$, we abuse notation and write $x_t = x_t(Z_t)$ to mean that there exists a function $x_t : \mathcal{N}_t \to \mathbb{R}$ such that $x_t(\omega) = x_t(Z_t(\omega))$ for all $\omega \in \Omega$, or, equivalently, that x_t is $\sigma(Z_t)$-measurable. Similarly, if x is a predictable process, we write $x_t = x_t(Z_{t-1})$ to express the condition that there exists a function $x_t : \mathcal{N}_{t-1} \to \mathbb{R}$ such that $x_t(\omega) = x_t(Z_{t-1}(\omega))$ for all $\omega \in \Omega$ (a condition that is equivalent to the $\sigma(Z_{t-1})$-measurability of x_t). Even though x_t denotes two separate quantities, the meaning is clear from the context.

With these conventions, suppose that the market is arbitrage-free and is implemented by the MMA and J contracts, as specified in the last section, with the additional restriction that for every time $t > 0$,

$$\left(r_t, \delta_t, V_t, \mu^G_t, \sigma^G_t\right) = \left(r_t(Z_{t-1}), \delta(Z_t), V_t(Z_t), \mu^G_t(Z_{t-1}), \sigma^G_t(Z_{t-1})\right),$$

and therefore $S_t \equiv V_t - \delta_t = S_t(Z_t)$. A market price of risk process $\eta \in \mathcal{H}$, defining an EMM Q, is specified as a predictable solution to equation (2.5.12) such that $1 - \eta' \Delta B$ is strictly positive (with $\eta_0 = 0$). The reader can verify that in this context, we can select $\eta \in \mathcal{H}$ to satisfy $\eta_t = \eta_t(Z_{t-1})$, $t = 1, \dots, T$. Assuming such a selection, it follows, from Lemma 2.6.1, that Z is a Markov process under Q, as well as P.

As in the last section, we consider the pricing and replication of a traded contract (δ^*, V^*), with the added assumption that $\delta^*_t = \delta^*_t(Z_t)$

for all t, and $V_T^* = V_T^*(Z_T)$ (which is redundant by the convention $\delta_T^* = V_T^*$). Pricing the contract recursively using the EMM Q and the Markov property of Z relative to Q gives

$$V_{t-1}^* = \delta_{t-1}^*(Z_{t-1}) + \frac{\mathbb{E}^Q[V_t^*(Z_t) \mid Z_{t-1}]}{1 + r_t(Z_{t-1})} = V_{t-1}^*(Z_{t-1}).$$

Therefore, $V_t^* = V_t^*(Z_t)$ for all t. Using the notation (2.5.17) for the corresponding gain process, we also have $\mu_t^* = \mu_t(Z_{t-1})$ and $\sigma_t^* = \sigma_t^*(Z_{t-1})$ (why?). A trading strategy that replicates the contract (δ^*, V^*) can be selected to have an analogous Markovian structure. We saw in the last section that a trading strategy (θ^0, θ) such that $(\delta^*, V^*) = (\delta^\theta, V^\theta)$ can be constructed by letting θ_t^0 be defined by (2.5.16) where θ_t solves $\sigma^*(Z_{t-1}) = \theta_t \sigma_t^G(Z_{t-1})$. Select such a θ of the form $\theta_t = \theta_t(Z_{t-1})$. Let $S^* \equiv V^* - \delta^*$ and $S \equiv V - \delta$ and substitute the identities $V_t^* = S_{t-1}^* + \Delta G_t^*$ and $V_t = S_{t-1} + \Delta G_t$ in expression (2.5.16) for θ_t^0. Using the dynamics of G and G^*, it follows that $\theta_t^0 = \theta_t^0(Z_{t-1})$, where

$$\theta_t^0(z) = \frac{S_{t-1}^*(z) + \mu_t^*(z) - \theta_t(z)\left(S_{t-1}(z) + \mu_t^G(z)\right)}{1 + r_t(z)}.$$

This type of Markovian scaffolding can be easily extended to our earlier discussion of option pricing given the existence of a dominant choice. The following example illustrates the basic idea.

Example 2.6.2 (American call). We revisit the American call of Examples 1.5.5 and 2.4.11, but with the underlying stock potentially paying dividends. Assuming the above Markovian structure and notation convention, the option's value process and an associated optimal exercise time can determined by solving the backward recursion

$$V_t^*(z) = \max\left\{S_t(z) - K, \frac{\mathbb{E}^Q\left[V_{t+1}^*(Z_{t+1}) \mid Z_t = z\right]}{1 + r_{t+1}(z)}\right\}, \quad V_{\bar{\tau}+1}^*(z) \equiv 0,$$

where z ranges over all possible values of the Markov state. The idea is that at time t and Markov state z, assuming the option has not been already exercised, it is optimal to exercise the option and collect $S_t(z) - K$ if the maximum is achieved by the first term and it is optimal to keep the option alive if the maximum is achieved by the second term. Define the stopping time

$$\tau^* \equiv \min\{t \mid V_t^*(Z_t) = S_t(Z_t) - K\}$$

and for any stopping time $\tau \in \bar{\mathcal{T}}$, let V^τ be the present value process (under the EMM Q) of the dividend process $\delta_t^\tau \equiv (S_t - K)\,1_{\{\tau = t\}}$. It is left as an exercise to verify (using a backward-in-time inductive argument) that $V_t^\tau \leq V_t^*$, with equality if $\tau = \tau^*$. This confirms that τ^* maximizes present value and is therefore optimal, and that V^* represents the option value process in the sense of Section 1.5. $\diamondsuit$

2.7. A glimpse of the continuous-time theory

We have so far taken the unit of time to coincide with a period, which implies that the time horizon T is the same as the number of periods N. For this section, we fix the time horizon T in a given time unit, which we call a year, and discuss approximations and associated simplifications that arise as the number of periods N becomes very large. Continuous-time models of interest are idealized representations of such large discrete models in the sense that the quantitative predictions of the discrete and continuous models are very close provided N is large enough. Theoretical support for this claim is provided by limit theorems as N goes to infinity. The rigorous details of both the continuous-time limiting model and arguments of convergence are highly technical and well beyond the scope of this presentation. Moreover, infinite models bring into scope set-theoretic esoteric aspects of the theory that are far removed from applications. An introduction founded on this chapter's finite information tree analysis should help demystify some of the continuous-time tools, as well as provide a basis for prioritizing aspects of the continuous-time theory that are most relevant for economic theory.

While one can embed an arbitrary sequence of discrete risks into a continuous-time model, tractability benefits derive from the assumption that uncertainty evolves as a sequence of risks that are in a sense small or infinitesimal in the continuous-time limit. The fundamental building blocks for constructing such high frequency sequences of small risks are two types of stochastic process: Brownian motion and the Poisson process. The incremental change of either type of process over an infinitesimal time interval represents a small risk, but for opposing reasons. For Brownian motion, a change is certain to happen, but the magnitude of the change is infinitesimal. For a Poisson process, the change is fixed at a unit amount, but the probability of change is infinitesimal. Mixing of Brownian motions and Poisson processes can be used to construct an arbitrary Lévy process,[11] which can be thought of as a representation of an arbitrary sequence of independent identically distributed small risks in high frequency. Lévy processes are further extended to classes of so-called semimartingales,[12] where the conditional moments of each small risk can be time-varying and path dependent, presenting a rich palette for formulating statistically testable stochastic models. This being a pedagogical introduction, we focus on the one-dimensional Brownian case.

[11] Cinlar [2010] provides a broad introduction to Probability theory that includes Lévy processes. Applebaum [2004] provides an overview of Lévy processes with a focus on stochastic integration.

[12] See Jacod and Shiryaev [2003] and Jacod and Protter [2012].

The set of **times** for this section is $\{0, h, 2h, 3h, \ldots, Nh\}$, where $h \equiv T/N$ represents the time length of each period in years. The idea is to fix T and consider versions of the model as N goes to infinity and h goes to zero. We proceed informally with a given value of N, but with the understanding that N is big in the following sense. Suppose $h = 10^{-6}$ years and we are only interested in computing quantities to a six-decimal precision. We therefore think of h as small but not negligible and quantities of higher order, like $h^{1.5} = 10^{-9}$ and $h^2 = 10^{-12}$, as negligible. On the other hand $\sqrt{h} = 10^{-3}$ is 1000 times bigger than h and definitely not negligible. Note that as h goes to zero, $\sqrt{h}$ becomes infinitely larger than h.

The underlying filtration $\{\mathcal{F}_{nh} \mid n = 0, 1, \ldots, N\}$ is defined analogously to Example 1.1.1, where information available at time $t = nh$ can be thought of as being the outcome of n consecutive coin tosses. Equivalently, we assume that the underlying filtration is generated by a process B, where $B_0 \equiv 0$, and over every period the increment of B can take one of two possible values. As in Section 2.5, we fix an underlying full-support probability P, relative to which B is a (zero-mean) martingale. Moreover, we assume that $v \equiv \mathbb{E}_t\left[(B_{t+h} - B_t)^2\right]$ is the same for all $t = nh$, and B is normalized so that $\mathbb{E}\left[B_T^2\right] = T$. Since $B_T = \sum_{n=1}^{N} \Delta B_{nh}$, where $\Delta B_{nh} \equiv B_{nh} - B_{(n-1)h}$,

$$\mathbb{E}\left[B_T^2\right] = \sum_{n=1}^{N} \mathbb{E}\left[(\Delta B_{nh})^2\right] + 2 \sum_{m<n} \mathbb{E}\left[\Delta B_{mh} \Delta B_{nh}\right].$$

For $m < n$, $\mathbb{E}\left[\Delta B_{mh} \Delta B_{nh}\right] = \mathbb{E}\left[\Delta B_{mh} \mathbb{E}_{mh} \Delta B_{nh}\right] = 0$. Therefore $T = \mathbb{E}\left[B_T^2\right] = Nv$ or $v = T/N = h$. Since, conditionally on $\mathcal{F}_t$, $B_{t+h} - B_t$ can only take two values, $\mathbb{E}_t\left[B_{t+h} - B_t\right] = 0$ and $\mathbb{E}_t\left[(B_{t+h} - B_t)^2\right] = h$, the increments of B must satisfy

$$P\left[B_{t+h} - B_t = \sqrt{h} \mid \mathcal{F}_t\right] = P\left[B_{t+h} - B_t = -\sqrt{h} \mid \mathcal{F}_t\right] = \frac{1}{2}.$$

The results of Section 2.5 all apply here, too. What is new is the calibration in terms of h and our earlier conventions on what powers of h we can ignore. Based on those, for every smooth function $f : \mathbb{R} \to \mathbb{R}$, we have the (informally stated) second-order Taylor approximation

$$f\left(B_{t+h}\right) - f\left(B_t\right) \approx f'\left(B_t\right)\left(B_{t+h} - B_t\right) + \frac{1}{2} f''\left(B_t\right) h. \tag{2.7.1}$$

We now use this approximation to compute the limiting distribution of B as $N \to \infty$ (with T fixed). The argument presented is heuristic, but it is the essential part of a rigorous version. We assume familiarity with the normal (or Gaussian) distribution and its characteristic function (or Fourier transform). Consider approximation (2.7.1) with

$$f(x) \equiv \exp(i\theta x) \equiv \cos(\theta x) + i \sin(\theta x), \quad i \equiv \sqrt{-1}.$$

The fact that f is complex-valued presents no problem, since the approximation can be applied on the real and imaginary parts separately. Fixing any time s, we compute $m_s(t) \equiv \mathbb{E}_s f(B_t)$ for $t \in [s, T]$, which defines the characteristic function of the distribution of B_t conditionally on time-s information. The law of iterated expectations and the martingale property of B imply $\mathbb{E}_s[B_{t+h} - B_t] = 0$. Applying $\mathbb{E}_s$ on both sides of (2.7.1) and using the fact that $f''(x) = -\theta^2 f(x)$, we have

$$\frac{m_s(t+h) - m_s(t)}{h} \approx -\frac{\theta^2}{2} m_s(t).$$

For infinitesimal h, we conclude that the derivative of $\log(m_s(t))$ with respect to t equals $-\theta^2/2$. Since $m_s(s) = f(B_s)$, we can integrate from s to t to find that in the limit as $N \to \infty$,

$$\mathbb{E}_s \exp(i\theta(B_t - B_s)) = \exp\left(-\frac{\theta^2}{2}(t-s)\right).$$

This is the characteristic function of a normal distribution with mean zero and variance $t - s$, which must therefore be the distribution of $B_t - B_s$ conditionally on time-s information. Coupled with the fact that in the limit the paths of B are continuous, we are led to the usual[13] definition of (one-dimensional) **Standard Brownian Motion (SBM)**, or **Wiener process**, as *an independent-increments stochastic process with continuous paths that start at zero, and whose increment over any time interval of length Δ is normally distributed with mean zero and variance Δ.*

We henceforth write B^N for the version of B corresponding to the N-period model, and B for SBM. To get a sense of what the limiting paths of B should look like, note that

$$\sum_{s<t} \left|B^N_{s+h} - B^N_s\right| = t\sqrt{N} \quad \text{and} \quad \sum_{s<t} \left(B^N_{s+h} - B^N_s\right)^2 = t.$$

These are, respectively, the **total** and **quadratic variation** processes of B^N. Clearly, as $N \to \infty$, with T fixed, the total variation blows up while the quadratic variation remains the same. It can further be shown that, with probability one, the paths of SBM, while continuous, are nowhere differentiable. Despite the extremely erratic behavior of the Brownian paths, the existence of SBM motion has been established in a variety of ways, within the usual set-theoretic foundations of real analysis. The infinite total variation property means that every path of B, although continuous, has infinite length. The quadratic variation property means that over every time interval the variance of B_t can be perfectly estimated by taking an infinite sequence of time-series estimates of the variance, whose limit is the true variance t. Thus

[13] Mörters and Peres [2010] and Revuz and Yor [1999] are some excellent accounts of Brownian motion and its fascinating properties, some of which are informally discussed in this section.

observation of any given path of B perfectly reveals the variance, but not the mean, which can only be estimated with arbitrary precision in an infinite-horizon version of the model, where the law of large numbers implies that $\lim_{t\to\infty} B_t/t$ with probability one.

A key assumption in our heuristic derivation of the Brownian motion distribution has been that B is a zero-mean martingale satisfying $\mathbb{E}_t[(B_{t+h} - B_t)^2] = h$ or, equivalently, $B_t^2 - t = \mathbb{E}_t\left[B_{t+h}^2 - (t+h)\right]$. This insight is reflected in Lévy's characterization of Brownian motion: *A process B with continuous paths*[14] *starting at zero is SBM if and only if both B_t and $B_t^2 - t$ are martingales.* In fact, for the "if" part, it is sufficient to assume that B_t and $B_t^2 - t$ are *local* martingales. The distinction between a martingale and a local martingale is not present in the finite information setting, but is essential in the continuous-time limit. A **martingale** over the continuous time interval $[0, T]$ is an adapted stochastic process M such that $\mathbb{E}\,|M_t| < \infty$ for every time t, and $s < t$ implies $\mathbb{E}_s M_t = M_s$. A **local martingale** is an adapted process M for which there exists an increasing sequence of stopping times $\tau_1 \leq \tau_2 \leq \cdots$ converging to T with probability one and such that for every n, the process M is a martingale up to time τ_n (meaning that the process that equals M_t on $\{t \leq \tau_n\}$ and M_{τ_n} on $\{t > \tau_n\}$ is a martingale). On a three-date infinite state-space setting, we may think of an adapted process M whose increment $\Delta \equiv M_2 - M_1$ satisfies $\mathbb{E}_1\,|\Delta| < 0$ and $\mathbb{E}_1\Delta = 0$, but from the perspective of time zero, we have $\mathbb{E}\,|\Delta| = \infty$, which leads to an example[15] of a local martingale that is not a martingale. This type of example fully encapsulates the distinction between a martingale and a local martingale in discrete time,[16] but the issue is more subtle in the continuous-time limit.

The fact that the paths of B have infinite length implies that an integral $\sigma \bullet B$ cannot be defined path by path in a conventional way, even if σ is assumed to be bounded. A key insight of Ito's calculus is that the integral $\sigma \bullet B$ must be defined as a genuinely stochastic integral, taking into account the entire filtration and the requirement that σ is predictable. (In the Brownian continuous-time limit, predictability of σ can be thought of as the requirement that σ is adapted, since the distinction of the information at time t and $t - dt$ is negligible. For example, the SBM B is predictable, while none of the B^N are.) Where in discrete time we would recursively express the relationship $M = \sigma \bullet B$

[14] The continuity of paths is essential here. Suppose instead we applied an analogous argument for a process $B_t = C_t - t$, where the paths of C are right-continuous and constant except for jumps of size $+1$. Think of C_t as counting the number of arrivals over $[0, t]$. Then the martingale property of B characterizes C as a Poisson process with unit arrival rate. This is Watanabe's characterization of Poisson processes.

[15] See Example 1.49 of Chapter 1 in Jacod and Shiryaev [2003].

[16] See Proposition 1.64 of Chapter 1 in Jacod and Shiryaev [2003].

as $\Delta M_t = \sigma_t \Delta B_t$, in continuous time, we write $dM_t = \sigma_t dB_t$, where the differential notation dM_t can be thought of as standing for the infinitesimal increment $M_t - M_{t-dt}$, and analogously for dB_t. We also write $M_t = \int_0^t \sigma_s dB_s$, $t \in [0,T]$. Where in the finite state-space model $\sigma \bullet B$ is always a martingale, in the current context we can only claim that $\sigma \bullet B$ is a *local* martingale. As the heuristic $(dM_t)^2 = \sigma_t^2 (dB_t)^2 = \sigma_t^2 dt$ suggests, in order for the process $M \equiv \sigma \bullet B$ to be well-defined as a process of finite quadratic variation, it is necessary that the process $Q_t \equiv \int_0^t \sigma_t^2 dt$ is well-defined and finite, in which case, Q is exactly the quadratic variation process of M.

Ito's calculus corresponds to the limiting case of our earlier approximations in terms of h, which can be stated as the exact relationships

$$(dB_t)^2 = dt \quad \text{and} \quad dB_t dt = (dt)^2 = 0, \tag{2.7.2}$$

while the Taylor approximation (2.7.1), for twice continuously differentiable $f : \mathbb{R} \to \mathbb{R}$, becomes

$$df(B_t) = f'(B_t)\, dB_t + \frac{1}{2} f''(B_t)\, dt. \tag{2.7.3}$$

For example, for $f(x) = x^2$, we find $dB_t^2 = 2B_t dB_t + dt$, or

$$B_t^2 - t = 2 \int_0^t B_t dB_t. \tag{2.7.4}$$

Since B is predictable, the above integral defines a local martingale. For a SBM B, we already know that $B_t^2 - t$ is a martingale. The argument leading to the above identity, however, applies to any local martingale B with continuous paths and quadratic variation $(dB_t)^2 = dt$. By Lévy's characterization, any such local martingale starting at zero must necessarily be a SBM.

The preceding argument leads to an interesting insight on the relationship between volatility and time. Consider the local martingale $M_t \equiv \int_0^t \sigma_t dB_t$, for predictable σ, defining the volatility of M. The process M is close to being a Brownian motion—it is a local martingale with continuous paths, but it does not necessarily have unit volatility: $(dM_t)^2 = \sigma_t^2 dt$. Imagine now a movie of M which at time t is played at speed σ_t^2. Think of t as real time for the duration $[0,T]$ of the movie. Think of u as time shown on a clock within the movie. Both times are set to zero at the beginning of the movie. At real time t, the clock in the movie shows $u = \int_0^t \sigma_s^2 ds$, which is the quadratic variation of M up to time t. With u and t so related, let $W_u = M_t$. For example, if W represents a contract's gain process from the perspective of a character in the movie, M represents the same gain process from the perspective of the viewer. The movie character observes that W is a local martingale with continuous paths that satisfies $(dW_u)^2/du = (dM_t)^2/\sigma_t^2 dt = 1$, and therefore can correctly claim that W is SBM.

To get a glimpse of phenomena that arise in the continuous-time limit but are not present in a finite-tree model, suppose $\int_0^T \sigma_s^2 ds = \infty$ (for example, $\sigma_t = (T-t)^{-1/2}$). The movie is the complete biography of an immortal god, who correctly argues that with probability one W eventually hits the value one. One way to see this is through the reflection principle for SBM. For each (continuous) path ω of W that crosses 1 before time t, there is an equally likely path $\tilde{\omega}$ that coincides with ω up the first time t_1 that ω hits 1, and thereafter it is the vertical reflection of ω relative to the horizontal line through 1, that is, for $s > t_1$, $\tilde{\omega}_s \equiv 1 - (\omega_s - 1)$. By construction, $\tilde{\omega}_t < 1$ if and only if $\omega_t > 1$. Both ω and $\tilde{\omega}$ have the property that they cross 1. Conversely, every path that crosses 1 before time t belongs to such a pair $\{\omega, \tilde{\omega}\}$. Therefore, the probability that W crosses 1 by time t is twice the probability of the event $\{W_t > 1\} = \left\{W_t t^{-1/2} > t^{-1/2}\right\}$, a probability that converges to one as $t \to \infty$ since $W_t t^{-1/2}$ is normally distributed with zero mean and unit variance. Equivalently, the stopping time $\tilde{\tau}$ defined as the first time that W takes the value one is finite with probability one. But then the viewer of the movie must also conclude that there is a $[0,T]$-valued stopping time τ such that $M_\tau = 1$ with probability one. Doob's optional stopping theorem states that if M is a zero-mean martingale, then $\mathbb{E}M_\tau = 0$ for every stopping time τ that is bounded by some deterministic time. Within the movie, we have an example of a zero-mean martingale W and a stopping time $\tilde{\tau}$ such that $\mathbb{E}W_{\tilde{\tau}} = 1$, which is consistent with Doob's theorem, since $\tilde{\tau}$ is not bounded—it can take an arbitrarily long time for W to hit one. The movie viewer's stopping time τ, however, is bounded and therefore it must be the case that M is a local martingale that is not a martingale.

In a market interpretation, we can think of σ as a trading strategy in a contract with gain process B, while within the movie the god holds one share in a contract whose gain process is W. The cumulative gains of σ at real time t match the cumulative gains of the god's strategy at god time $u = \int_0^t \sigma_s^2 ds$. In a finite model, the expected gains from trading a contract whose gain process is a zero-mean martingale are zero. The god who holds the contract until W hits one is guaranteed a unit expected gain. Likewise, the trading strategy σ can guarantee a unit gain by time T. Both strategies are a form of arbitrage. Within the movie, the arbitrage disappears if we place a martingale lower bound on how negative the god's wealth can become before liquidation. Alternatively, the arbitrage disappears if we assume the god has a finite expected lifetime after all. In real time, an appropriate lower bound on wealth eliminates the arbitrage, and that's the modeling choice we will adopt in the following section. Limiting the god's expected horizon corresponds to the square integrability condition $\mathbb{E}\int_0^T \sigma_s^2 ds < \infty$, which implies that $M \equiv \sigma \bullet B$ is a zero-mean finite-variance martingale.

Conversely, every such martingale M can be represented as $M \equiv \sigma \bullet B$ for some square integrable predictable σ.

A natural type of process used to represent prices in a Brownian model is an **Ito process**, defined as a process of the form

$$x_t = x_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dB_s \quad \text{or} \quad dx_t = \mu_t dt + \sigma_t dB_t,$$

for adapted processes μ and σ, respectively referred to as the **drift** and **volatility** of x, where it is assumed that the integrals $\int_0^T |\mu_t|\, dt$ and $\int_0^T \sigma_t^2 dt$ are well-defined and finite with probability one. The Ito decomposition of x into a drift and a volatility term is unique (modulo some zero-probability event technicalities we will ignore). To see why, suppose $dx_t = \mu_t dt + \sigma_t dB_t = \tilde{\mu}_t dt + \tilde{\sigma}_t dB_t$. Then $(\mu_t - \tilde{\mu}_t)\, dt + (\sigma_t - \tilde{\sigma}_t)\, dB_t = 0$. Squaring this expression and using the Ito calculus (2.7.2) results in $(\sigma_t - \tilde{\sigma}_t)^2\, dt = 0$ and therefore $\sigma = \tilde{\sigma}$ (essentially), which in turn implies $\mu = \tilde{\mu}$ (essentially). Alternatively, we can heuristically, but correctly, think of the drift and volatility as being determined by $\mathbb{E}_{t-dt}[dx_t] = \mu_t dt$ and $\mathrm{cov}_{t-dt}[dx_t, dB_t] = \sigma_t dt$. For any predictable process θ, the integral $\theta \bullet x$ can be expressed as

$$\int_0^t \theta_s dx_s = \int_0^t \theta_s \mu_s ds + \int_0^t \theta_s \sigma_s dB_s \quad \text{or} \quad \theta_t dx_t = \theta_t \mu_t dt + \theta_t \sigma_t dB_t,$$

provided $\int_0^T |\theta_t \mu_t|\, dt$ and $\int_0^T (\theta_t \sigma_t)^2\, dt$ are well-defined and finite with probability one.

The type of second-order Taylor approximation that led to identity (2.7.1) can be applied to to an arbitrary Ito process x and associated time-dependent function $f : [0,T] \times \mathbb{R} \to \mathbb{R}$ that is continuously differentiable with respect to time and twice continuously differentiable with respect to the second argument:

$$df(t,x_t) = \frac{\partial f(t,x_t)}{\partial t} dt + \frac{\partial f(t,x_t)}{\partial x} dx_t + \frac{1}{2} \frac{\partial^2 f(t,x_t)}{\partial x^2} (dx_t)^2, \tag{2.7.5}$$

where $(dx_t)^2 = \sigma_t^2 dt$, according to the Ito multiplication rules (2.7.2). This is known as **Ito's formula** (or lemma) and extends to multidimensional Ito processes (and beyond that, to general semimartingales). We will not go further in this direction here, except state an important special case, known as the **integration by parts** formula: If x and y are Ito processes, then $d(x_t y_t) = y_t dx_t + x_t dy_t + dx_t dy_t$. (For a heuristic proof, expand the product $(x_{t-dt} + dx_t)(y_{t-dt} + dy_t)$ and use the continuity of x which implies that $x_{t-dt} = x_t$, and analogously for y.) By the Ito multiplication rules (2.7.2), if x and y have volatility σ^x and σ^y, respectively, then $dx_t dy_t = \sigma_t^x \sigma_t^y dt$. For example, for $x = y = B$, we recover[17] identity (2.7.4). For strictly positive x and y, the following

[17] Integration by parts can be used inductively to show that every polynomial in (t,x) satisfies Ito's formula, which can then form the basis for proving the Ito

version of integration by parts is often more convenient to use:

$$\frac{d\left(x_t y_t\right)}{x_t y_t} = \frac{dx_t}{x_t} + \frac{dy_t}{y_t} + \frac{dx_t}{x_t}\frac{dy_t}{y_t}.$$

We close this section with some comments on the interchange of limits and expectations which will arise in our discussion of state pricing in the following section. Consider first the example of the state space $\Omega = (0,1]$ with the underlying probability P being the uniform distribution: $0 < a < b \leq 1$ implies $P\left((a,b]\right) = b - a$. The sequence of random variables $X_n(\omega) \equiv n1_{\{\omega \leq 1/n\}}$ converges to 0 for every $\omega \in \Omega$, yet $\mathbb{E}X_n = 1$ for all n. Unlike the finite Ω case, we cannot freely interchange limits and expectation. An essential positive result in this regard, which is not far from the way expectations are defined, is the **monotone convergence theorem:** For all $[0,\infty)$-valued random variables $X, X_1, X_2, \ldots$, if $X_n \uparrow X$ then $\mathbb{E}X_n \uparrow \mathbb{E}X$ (where $X_n \uparrow X$ means that, with probability one, $X_{n+1} \geq X_n$ for all n, and $\lim_{n\to\infty} X_n = X$). The conclusion applies even if $\mathbb{E}X = \infty$. We will use the monotone convergence theorem through Fatou's lemma, which is stated and proved below.

LEMMA 2.7.1 (Fatou). *If $X_n \geq 0$ for all n, then*

$$\liminf_n \mathbb{E}X_n \geq \mathbb{E}\liminf_n X_n.$$

PROOF. Let $Y_n = \inf_{k\geq n} X_k$ and note that $0 \leq Y_n \uparrow Y \equiv \liminf_n X_n$. For all $m \geq n$, $\mathbb{E}X_m \geq \mathbb{E}Y_n$ and therefore $\inf_{m\geq n} \mathbb{E}X_m \geq \mathbb{E}Y_n$. Taking the limit as $n \to \infty$ and using the monotone convergence theorem, we conclude that $\liminf_n \mathbb{E}X_n \geq \mathbb{E}Y$. □

2.8. Brownian market example

As a simple example, in this section we discuss an arbitrage-free market with the underlying filtration generated by SBM B over the continuous time interval $[0,T]$. In other words, the information available at time t is the realization of the path of B up to time t. The process B is one-dimensional, which means that the filtration can be thought of as a high-frequency limiting version of the binomial filtration of Example 1.1.1. The extension to multi-dimensional B is mainly a matter of notation and will not be given here. The arguments that follow are for the most part similar to finite-information counterparts we have encountered already, with simplifications resulting due to small-risk approximations which are codified as exact infinitesimal relationships by the Ito calculus. As in the last section, we omit some mathematical details. For example, the mere definition of the information generated by the SBM B requires notions of σ-algebras and measurability, as

formula as stated above. The idea is to stop the process x before it leaves a compact interval and then uniformly approximate f and the relevant derivatives on that interval by polynomials.

well as an underlying probability measure that defines the expectation operator $\mathbb{E}$. The hope is that once the main economic arguments are established, and the formalities are tied to finite-tree constructs, the more advanced literature that includes all these mathematical details will make a lot more sense.

A general continuous-time cash flow on the given filtration can be defined analogously to the way cumulative distribution functions are used to specify probability distributions on the real line. This being a simple introduction, we instead consider a special type of market in which all cash flows consist of just two lump-sum payments at times 0 and T. More precisely, a **cash flow** is a pair $(c_0, c_T) \in \mathbb{R} \times L_2$, where L_2 is the set of every random variable x such that $\mathbb{E}[x^2] < \infty$ (with the usual convention of identifying any two random variables x and $\tilde{x}$ if $\mathbb{E}\left[(x-\tilde{x})^2\right] = 0$). Assuming the cash flow is traded, the time-0 payment c_0 is used to purchase an initial portfolio of value $-c_0$, which is then updated through a trading strategy up to time T, when the terminal portfolio value c_T is liquidated. Every trading strategy is assumed to be self-financing, meaning that no cash is generated or injected after the initial purchase and prior to the terminal liquidation.

The market is implemented by two contracts: a money-market account and a stock. The **money-market account (MMA)** has a value process identically equal to one unit of account, let's call it a "dollar," and a continuous interest rate $r \in \mathbb{R}$. A dollar in the MMA at time t generates interest rdt at time $t + dt$. Equivalently, a dollar invested in the MMA at time zero with all interest continuously reinvested, results in a time-t account balance of e^{rt} dollars. The **stock** has a prices process S which follows **geometric Brownian motion with drift**: $S_t = S_0 \exp(at + \sigma B_t)$, for given constants $a \in \mathbb{R}$ and $S_0, \sigma \in (0, \infty)$. The assumption that σ is positive is just a convention, since the implications of the model are invariant to replacing B with the SBM $-B$. The stock's (continuous) dividend yield is a constant $y \in \mathbb{R}$, meaning that a share purchased at time t pays ydt stock shares (or $yS_t dt$ dollars) in dividends at time $t + dt$. Equivalently, if a share of the stock is purchased at time zero and all dividends are reinvested in the stock, then the time-t stock position is e^{yt} shares, which is worth $e^{yt}S_t$ dollars. (Note that by this definition, we could have called r the dividend yield of the MMA.) Since interest and dividends are paid out continuously, the cum and ex-dividend price processes for each contract are the same. Ito's formula (2.7.5) implies that S is an Ito process satisfying

$$\frac{dS_t}{S_t} + ydt = \mu dt + \sigma dB_t, \quad \mu \equiv a + y + \frac{1}{2}\sigma^2. \tag{2.8.1}$$

A trading strategy takes the form of a pair of predictable processes (θ^0, θ), where θ^0_t is the time-t dollar balance in the MMA and θ_t is the the number of stock shares held at time t. The gain processes for the

two contracts are defined analogously to the finite case by

$$G_t^0 \equiv 1 + rt \quad \text{and} \quad G_t \equiv S_t + \int_0^t yS_u du.$$

The trading strategy(θ^0, θ) is **self-financing** if it satisfies the **budget equation** expressing the fact that the portfolio value is equal to the initial portfolio value plus accumulated gains from trading:

$$\theta_t^0 + \theta_t S_t = \theta_0^0 + \theta_0 S_0 + \int_0^t \theta_u^0 dG_u^0 + \int_0^t \theta_u dG_u, \tag{2.8.2}$$

where $dG_t^0 = rdt$ and $dG_t = dS_t + yS_t dt$.

The trading strategy (θ^0, θ) is **admissible** if it is self-financing (implying the integrals of the budget equation are well-defined) and there exist constants $\bar{\theta}^0$ and $\bar{\theta}$ such that, with probability one, the following **solvency constraint** is satisfied:

$$\bar{\theta}^0 e^{rt} + \bar{\theta} e^{yt} S_t + \theta_t^0 + \theta_t S_t \geq 0, \quad \text{for all } t \in [0, T]\,. \tag{2.8.3}$$

The constraint can be thought of as a simple form of a collateral requirement. A brokerage carries out your trades, but must ensure that at all times your total account balance can cover your trading losses. At time zero, you are asked to deposit $\bar{\theta}^0$ dollars in the MMA and $\bar{\theta}$ shares of the stock, with interest and dividends reinvested in the respective contracts. The solvency constraint requires that the net value of all your positions with the brokerage remains positive. While redundant in the finite state-space model, it is needed here in some form in order to preclude arbitrage trades based on increasingly larger bets of the type discussed in the last section. (As suggested by last section's discussion, a variant of this example replaces the solvency constraint with a square integrability restriction on admissible trading strategies. Either approach supports the example of Black-Scholes option pricing and hedging presented below.)

The budget equation should be invariant relative to a change of the unit of account implied by the strictly positive Ito process π. At time-t a dollar is the same as π_t of the new unit of account. The corresponding gain processes in the new unit are

$$G_t^{0\pi} \equiv \pi_t + \int_0^t r\pi_u du \quad \text{and} \quad G_t^{\pi} \equiv \pi_t S_t + \int_0^t y\pi_u S_u du.$$

The integration by parts formula implies that the budget equation (2.8.2) can be equivalently stated as

$$\theta_t^0 \pi_t + \theta_t \pi_t S_t = \theta_0^0 \pi_0 + \theta_0 \pi_0 S_0 + \int_0^t \theta_u^0 dG_u^{0\pi} + \int_0^t \theta_u dG_u^{\pi}. \tag{2.8.4}$$

The solvency constraint under the new units is inequality (2.8.3) multiplied through by π_t, and is therefore also invariant under this type of change of a unit of account.

We say that the cash flow (x_0, x_T) is **traded** if there exists an admissible trading strategy (θ^0, θ) such that $x_0 = -(\theta_0^0 + \theta_0 S_0)$ and $x_T = \theta_T^0 + \theta_T S_T$. The set of traded cash flows defines the **implemented market** X. A **state price density** (SPD) process in this context is a strictly positive Ito process π such that $\pi_T \in L_2$ and for every traded cash flow (x_0, x_T),

$$\pi_0 x_0 + \mathbb{E}\left[\pi_T x_T\right] \leq 0. \tag{2.8.5}$$

Note that the assumption $x_T, \pi_T \in L_2$ ensures that $\mathbb{E}\left|\pi_T x_T\right| < \infty$ thanks to the Cauchy-Schwarz inequality.

Based on the analysis for the finite case, we can guess the state price dynamics in terms of the short-rate process r and the market-price-of-risk process η, both of which are constant in this example:

$$\frac{d\pi_t}{\pi_t} = -r dt - \eta dB_t, \quad \text{where } \eta \equiv \frac{\mu - r}{\sigma}. \tag{2.8.6}$$

In integrated form, the proposed SPD is

$$\pi_t = \pi_0 e^{-rt}\xi_t, \quad \xi_t = \exp\left(-\frac{\eta^2}{2}t - \eta B_t\right). \tag{2.8.7}$$

To confirm that this form of π indeed follows the claimed dynamics, apply integration by parts and Ito's formula, which implies that ξ solves

$$\frac{d\xi_t}{\xi_t} = -\eta dB_t, \quad \xi_0 = 1. \tag{2.8.8}$$

This Ito decomposition implies that ξ is a local martingale. Relaxing temporarily the assumption that η is constant over time and states, it may well be the case that ξ is not a martingale. The local martingale property of ξ together with Fatou's lemma can be used to show that ξ is a supermartingale: $s > t$ implies $\mathbb{E}_t \xi_s \leq \xi_t$, and ξ is a martingale if and only if $\mathbb{E}\xi_T = 1$, in which case it defines a probability Q analogously to the finite case by $\xi_T = dQ/dP$. Here η is constant and the martingale property of ξ can be easily confirmed from (2.8.7) using the fact that B has independent and normally distributed increments.

As in the finite case, the coefficients of the Ito expansion of π are determined by the requirement that $G^{0\pi}$ and G^{π} are local martingales. To see that, suppose $d\pi_t/\pi_t = a_t dt + b_t dB_t$ and use integration by parts:

$$\begin{aligned}
\frac{dG_t^{0\pi}}{\pi_t} &= (a_t + r)\, dt + b_t dB_t, \\
\frac{dG_t^{\pi}}{\pi_t S_t} &= (\mu + a_t + \sigma b_t)\, dt + (\sigma + b_t)\, dB_t.
\end{aligned}$$

Setting the drift term in the first equation to zero is equivalent to setting $a_t = -r$, and given that, setting the drift term in the second equation to zero is equivalent to setting $b_t = -\eta$. In the following proposition we prove that π is a state-price density by showing that

the local martingale property of $G^{0\pi}$ and G^{π} implies the state-price density property of π.

PROPOSITION 2.8.1. *For any $\pi_0 \in (0,\infty)$, the process π defined by (2.8.6) is a state-price density.*

PROOF. Suppose the trade (x_0, x_T) is generated by the admissible trading strategy (θ_t^0, θ_t), satisfying the solvency constraint (2.8.3). Let

$$\bar{W}_t \equiv \bar{\theta}^0 e^{rt} + \bar{\theta} e^{yt} S_t \quad \text{and} \quad W_t \equiv \bar{W}_t + \theta_t^0 + \theta_t S_t.$$

We remarked earlier that $\xi_t = \pi_t e^{rt}$ is a martingale. Similarly, applying integration by parts to $\zeta_t = \pi_t e^{yt} S_t$, we find that $d\zeta_t/\zeta_t = (\sigma - \eta)\, dB_t$, which implies that ζ is a martingale (for the same reason that ξ is a martingale). The local martingale property of $G^{0\pi}$ and G^{π}, the budget equation (2.8.4), and the fact that $\pi_t \bar{W}_t$ is a martingale together imply that πW is a local martingale. There exists, therefore, an increasing sequence of stopping times τ_n that converges to T with probability one such that πW is a martingale up to time τ_n and therefore

$$\pi_0 W_0 = \mathbb{E}\left[\pi_{\tau_n} W_{\tau_n}\right], \quad n = 1, 2, \ldots \tag{2.8.9}$$

By the solvency constraint, $\pi_t W_t \geq 0$. By Fatou's Lemma 2.7.1,

$$\pi_0 W_0 = \liminf_{n\to\infty} \mathbb{E}\left[\pi_{\tau_n} W_{\tau_n}\right] \geq \mathbb{E}\left[\liminf_{n\to\infty} \pi_{\tau_n} W_{\tau_n}\right] = \mathbb{E}\left[\pi_T W_T\right].$$

Subtracting the martingale identity $\pi_0 \bar{W}_0 = \mathbb{E}\left[\pi_T \bar{W}_T\right]$ on both sides and using the fact that $W_0 = \bar{W}_0 - x_0$ and $W_T = \bar{W}_T + x_T$, we obtain (2.8.5). □

The equivalent martingale measure (EMM) argument of the finite case extends to the current context. Using the fact that ξ is a strictly positive unit-mean martingale, we define the probability Q through $\xi_T = dQ/dP$, where the latter is the density of Q with respect to P. The Girsanov-Lenglart theorem alluded to in Section 2.5 is known as simply Girsanov's theorem in the Brownian setting: $B_t^Q \equiv B_t + \eta t$ is a Q-martingale and therefore, by Lévy's characterization of SBM, B^Q is a SBM under the probability Q.

REMARK 2.8.2. Girsanov's theorem for Brownian motion can be related to the functional form of the standard normal density function ϕ as follows. Let $\Phi(x) \equiv \int_{-\infty}^{x} \phi(y)\, dy$ denote the standard normal cumulative distribution function. Girsanov's theorem implies that $\Phi(x) = Q[B_1^Q \leq x]$. We use expression (2.8.7) for ξ_1 and the change of measure formula $\mathbb{E}^Q z = \mathbb{E}[z\xi_1]$ for every $\mathcal{F}_1$-measurable bounded random variable z to conclude that

$$\Phi(x) = \mathbb{E}\left[1_{\left\{B_1^Q \leq x\right\}} \xi_1\right] = \int_{-\infty}^{x} \phi(y-\eta)\, e^{-\frac{\eta^2}{2} - \eta(y-\eta)} dy.$$

Equivalently, $\phi(y) = \phi(y-\eta)\exp(\eta^2/2 - \eta y)$. For $\eta = y$, this gives the standard normal density as

$$\phi(y) = \phi(0)\exp\left(-\frac{y^2}{2}\right),$$

where $\phi(0)$ is set to $(2\pi)^{-1/2}$ so that $\int_{-\infty}^{\infty}\phi(y) = 1$. Conversely, given the functional form of ϕ, the above steps can be reversed to confirm that $\Phi(x) = Q[B_1^Q \leq x]$. The argument can be easily extended to confirm that for all times $s > t$, $B_s^Q - B_t^Q$ has a normal distribution of zero mean and variance $s-t$, independently of $\mathcal{F}_t$, all under Q. Since B^Q has continuous paths starting at zero, this confirms that B^Q is a SBM under Q. $\diamondsuit$

The stock price dynamics can be equivalently written as

$$dS_t = (r-y)S_t dt + \sigma S_t dB_t^Q. \tag{2.8.10}$$

Repeating the above analysis with the EMM Q as the underlying probability, B^Q in place of B, and r in place of μ, results in the state-pricing condition $x_0 + \mathbb{E}^Q\left[e^{-rT}x_T\right] \leq 0$ for every traded cash flow x. If $\pm x$ are both traded, we have the pricing identity

$$-x_0 = \mathbb{E}\left[\frac{\pi_T}{\pi_0}x_T\right] = \mathbb{E}^Q\left[e^{-rT}x_T\right]. \tag{2.8.11}$$

The last equation also follows from the change of measure formula $\mathbb{E}^Q x_T = \mathbb{E}\left[(dQ/dP)\,x_T\right]$, since $\pi_T = e^{-rT}dQ/dP$. (These arguments apply with more general predictable r and η, with $e^{-\int_0^T r_t dt}$ in place of e^{-rT} and $B_t^Q \equiv B_t - \int_0^t \eta_s ds$, a SBM under Q.)

Example 2.8.3 (Forward pricing). Suppose a forward contract for delivery of one share of the stock at time T is traded at time zero, with corresponding **forward price** $F_{0,T} \in \mathbb{R}$. The cash flow (x_0, x_T) generated by entering a long position in the forward contract is given by $x_0 = 0$ and $x_T = S_T - F_{0,T}$. The same traded cash flow can be generated by borrowing $S_0 e^{-yT}$ from the MMA to purchase e^{-yT} stock shares at time zero, continuously rolling over the loan and reinvesting all dividends, resulting in a time-T long position of one stock share and a short dollar position $-S_0 e^{(r-y)T}$ in the MMA. Barring arbitrage, it follows that

$$F_{0,T} = S_0 e^{(r-y)T}. \tag{2.8.12}$$

This expression can also be dually derived using the pricing restriction (2.8.11) with $x_0 = 0$ and $x_T = S_T - F_{0,T}$ to conclude that $F_{0,T} = \mathbb{E}^Q S_T$ (since r is assumed to be constant). To show directly that the two expressions for $F_{0,T}$ are consistent, suppose $F_{0,T}$ is given

by (2.8.12) and note that

$$S_T = F_{0,T} \exp\left(-\frac{\sigma^2}{2}T + \sigma B_T^Q\right). \tag{2.8.13}$$

The exponential factor has unit mean under Q for the same reason that ξ_T as defined in (2.8.7) has unit mean under P. $\diamondsuit$

The replication and arbitrage-pricing argument of the last example follows a line of reasoning that was introduced in Section 2.5 and extends more generally to the current context. For simplicity, we consider a contract, we call the "star" contract, whose value process is V^* and whose only dividend is a terminal lump-sum payment $V_T^* \in L_2$. (In the preceding forward contract example, $V_T^* \equiv S_T - F_{0,T}$.) Using the EMM Q, let

$$V_t^* \equiv \mathbb{E}_t^Q \left[e^{-r(T-t)} V_T^*\right], \tag{2.8.14}$$

which defines the Q-martingale $V_t^* e^{-rt}$ thanks to the law of iterated expectations. Omitting some technical details on integrability conditions, a predictable martingale representation theorem implies the existence of a predictable process σ^* such that

$$V_t^* e^{-rt} = \mathbb{E}_t^Q \left[e^{-rT} V_T^*\right] = V_0^* + \int_0^t e^{-ru} \sigma_u^* dB_u^Q.$$

Integration by parts leads to the Ito decomposition

$$dV_t^* = rV_t^* dt + \sigma_t^* dB_t^Q = \left(rV_t^* + \sigma_t^* \eta\right) dt + \sigma_t^* dB_t. \tag{2.8.15}$$

This equation can be read as a condition that specifies the drift of V_t^* as a function of its volatility σ_t^*. In our discussion of equation (2.5.12) in Section 2.5, we saw that in the finite model such a restriction is equivalent to a backward recursion on the information tree. Equation (2.8.15) can be thought of as the continuous-time version of such a backward recursion, applied dt by dt backward in time, starting with the given terminal value V_T^*. Mathematically, (2.8.15) is an example of a backward stochastic differential equation (BSDE) to be solved jointly in V^* and σ^* given the terminal value V_T^*. The fact that the drift term in (2.8.15) is linear in V^* and σ^* makes this a linear BSDE, which is why we have (subject to some technical integrability requirements) a closed-form solution for V^* in (2.8.14)

Analogously to a construction in Section 2.5, we can think of the star contract as a synthetic contract, generated by the trading strategy (θ^0, θ), where

$$\theta_t^0 + \theta_t S_t = V_t^* \quad \text{and} \quad \theta_t S_t \sigma = \sigma_t^*. \tag{2.8.16}$$

The corresponding budget equation is satisfied, since, using (2.8.10),

$$\theta_t^0 dG_t^0 + \theta_t dG_t = rV_t^* dt + \sigma_t^* dB_t^Q = dV_t^*.$$

Unlike the finite case of Section 2.5, we must further verify the solvency constraint. Moreover, in order to arrive to an arbitrage pricing equation, as opposed to an inequality, we must confirm that both $\pm(\theta^0, \theta)$ are admissible. For instance, the condition is obviously satisfied for the forward contract of Example 2.8.3, as well as the case of a European call reviewed below.

EXAMPLE 2.8.4 (European call and the Black-Scholes formula). Suppose the star contract is a European call option on the stock with strike $K \in (0, \infty)$ and maturity T, and therefore $V_T^* \equiv (S_T - K)^+$. The star contract is traded by the above argument and the fact that the solvency constraint is satisfied for both the long and short position. Since V_T^* is strictly positive, the value of the call is always positive in an arbitrage-free market, and therefore no collateral is required in buying the option. For the strategy that shorts the call option at time zero, the fact that $-V_T^* \geq K - S_T$ implies that the solvency constraint is satisfied by $\bar{\theta}^0 = Ke^{-rT}$ and $\bar{\theta} = e^{-yT}$. The time-zero option price, also known as the option **premium**, is therefore $V_0^* = e^{-rT}\mathbb{E}^Q[(S_T - K)^+]$. By identity (2.8.13), the latter can be written as the following expression for the forward call premium per unit strike:

$$p \equiv \frac{V_0^* e^{rT}}{K} = \mathbb{E}^Q\left[\left(\exp\left(m - \frac{v^2}{2} + v\frac{B_T^Q}{\sqrt{T}}\right) - 1\right)^+\right], \tag{2.8.17}$$

where $v \equiv \sigma\sqrt{T}$ and $m \equiv \log(F_{0,T}/K)$, with the forward price $F_{0,T}$ given in (2.8.12). Let $\Phi(x) \equiv (2\pi)^{-1/2}\int_{-\infty}^{x} \exp(-y^2/2)\,dy$ denote the standard Gaussian cumulative distribution of function. Since the latter is the distribution of $B^Q/\sqrt{T}$ under Q, the expectation in (2.8.17) can be written as

$$p = e^m \Phi\left(\frac{m}{v} + \frac{v}{2}\right) - \Phi\left(\frac{m}{v} - \frac{v}{2}\right). \tag{2.8.18}$$

This is the Black and Scholes [1973] formula for pricing the European call. ♢

The Markovian structure introduced in Section 2.6 is also present in the current market model, with the stock price S being the Markov state. The analog of the forward recursion (2.6.2) are the Ito dynamics (2.8.1) of the stock price S, given the initial value S_0. This is an example of a (forward) stochastic differential equation (SDE). In a more general formulation, an SDE can represent the Ito dynamics of an underlying Markov process Z as $dZ_t = a(t, Z_t)\,dt + b(t, Z_t)\,dB_t$, given an initial value Z_0, and with the functions $a, b : [0, T] \times \mathbb{R} \to \mathbb{R}$ appropriately restricted to ensure a unique solution. Here $Z = S$, $a(t, Z) = \mu Z$ and $b(t, Z) = \sigma Z$, which allows a closed-form solution.

We henceforth assume that $V_T^* = f_T(S_T)$ for given $f_T : \mathbb{R}_+ \to \mathbb{R}$. Suppose the (suitably regular) function $f : [0, T] \times \mathbb{R}_+ \to \mathbb{R}$ solves the

partial differential equation (PDE)

$$\frac{\partial f}{\partial t} + (r-y)\, S\frac{\partial f}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial f^2}{\partial S^2} = rf, \quad f(T,\cdot) = f_T. \tag{2.8.19}$$

By Ito's lemma, it follows that

$$V_t^* = f(t, S_t) \quad \text{and} \quad \sigma_t^* = \frac{\partial f(t,S_t)}{\partial S} S_t \sigma$$

defines a solution to the linear BSDE (2.8.15). This is an illustration of a close connection between BSDEs and corresponding PDEs, which allows one to leverage the rich theoretical and computational insights of the PDE literature. (For example, PDE (2.8.19) can be transformed to the classical heat equations in Physics.) Conversely, probabilistic arguments in the BSDE literature can help shed light on the corresponding PDE. Given f, we can express the trading strategy defined in (2.8.16) as a function of the Markov state: $\theta_t^0 = \theta^0(t, S_t)$ and $\theta_t = \theta(t, S_t)$, where

$$\theta(t,S) = \frac{\partial f(t,S)}{\partial S}, \quad \theta^0(t,S) = f(t,S) - \theta(t,S)\, S. \tag{2.8.20}$$

Example 2.8.5 (European call replication). Suppose the star contract is the European call of the last example, and therefore $f_T(S) \equiv (S-K)^+$. Expressing the Black-Scholes equation in the form $V_0^* = f(0, S_0)$ results in an expression for $f(0,\cdot)$. The Markov property implies that $f(t,\cdot)$ is identical to $f(0,\cdot)$ after replacing T with $T-t$. The result is $f(t,S) = \theta^0(t,S) + \theta(t,S)\, S$, where

$$\theta^0(t,S) = -e^{-r(T-t)} K \Phi\left(\frac{\log(S/K) + (r-y)(T-t)}{\sigma\sqrt{T-t}} - \frac{\sigma\sqrt{T-t}}{2} \right),$$

$$\theta(t,S) = e^{-y(T-t)} \Phi\left(\frac{\log(S/K) + (r-y)(T-t)}{\sigma\sqrt{T-t}} + \frac{\sigma\sqrt{T-t}}{2} \right).$$

One can then easily check that these two function satisfy (2.8.20) and therefore define the trading strategy that implements the European call as a synthetic contract in the MMA and the stock. ♢

In the more applied literature, the notation (Θ, Δ, Γ) is used for the triple $(\partial f/\partial t, \partial f/\partial S, \partial f^2/\partial S^2)$. A market maker who sells the star contract can use f to price the contract, and buy Δ stock shares to hedge the position, an activity known as delta hedging. Γ contains information on how aggressively the market maker must trade to maintain the hedging position as a consequence of changes in the underlying stock price. By Ito's lemma, (Θ, Δ, Γ) encapsulates a snapshot of the value $V_{t+dt}^* = V_t^* + \Theta dt + \Delta dS_t + (\Gamma/2)(dS_t)^2$ a short time interval in the future as a function of the stock price, and PDE (2.8.19) expresses the fact that $V_t^* = e^{-rdt}\mathbb{E}_t^Q V_{t+dt}^*$ (why?). Notice that Δ represents a directional exposure that delta-hedging seeks to cancel out, while Γ represents an exposure to volatility that can be a source of error in

the delta hedging strategy. In theory, trading sufficiently frequently makes the error negligible, but in practice the market maker faces small transaction costs and may wish to control the overall Γ to reduce the need for frequent rebalancing, a practice known as gamma hedging. Moreover, what we called the star contract, may be a portfolio of contracts with the same Markov structure, for example, various options with different characteristics on the same stock. Since the derivatives defining Θ, Δ and Γ are linear, these parameters can be easily added up across portfolio positions, allowing the market maker to quickly assess the effect of new trades on the current portfolio.

2.9. Exercises

Exercise 1 Consider the setting of Section 2.2 with $T = 1$. Since there is only one period, we drop all time subscripts. (The model can be thought of as a single period on the information tree conditional on a given beginning-of-period spot.) Recall that the market is arbitrage-free, $\mathcal{R}$ is the set of all traded returns, and $R^0 \in \mathcal{R}$ is the (zero-variance) MMA return. Assume that the market is *not* priced risk neutrally: There exists some $R \in \mathcal{R}$ such that $\mathbb{E}R \neq R^0$. Finally, assume that the positive-variance return $R^* \in \mathcal{R}$ maximizes the Sharpe ratio (relative to R^0) over $\mathcal{R}$.

(a) Give a direct argument showing that R^* must be a minimum variance frontier return.

(b) Fix any positive-variance return $R \in \mathcal{R}$ such that $R - R^*$ is correlated with R^* (that is, $\operatorname{cov}[R - R^*, R^*] \neq 0$). On the plane, consider the locus

$$H \equiv \{(\operatorname{stdev}[(1-w)R^* + wR], \mathbb{E}[(1-w)R^* + wR]) \mid w \in \mathbb{R}\}$$

as well as the line L that connects $h^* \equiv (\operatorname{stdev}[R^*], \mathbb{E}R^*)$ to $(0, R^0)$. For every point $(s, m) \in H$ with $s > 0$, the slope of the line that connects $(0, R^0)$ to (s, m) is the Sharpe ratio $(m - R^0)/s$ of the corresponding return. Since the latter is maximized by R^* over all of $\mathcal{R}$ and h^* lies on H, it follows that H lies below L on the plane and it touches L at h^*. Since H is smooth, it must then be the case that H is tangent to L at h^*. Compute the slope of the tangent of H at h^* and show that the tangency condition is equivalent to the beta pricing equation

$$\mathbb{E}R - R^0 = \frac{\operatorname{cov}[R, R^*]}{\operatorname{var}[R^*]} \left(\mathbb{E}R^* - R^0\right).$$

(c) How would the arguments of parts (a) and (b) change if R^* were instead assumed to *minimize* the Sharpe ratio over all traded returns? Finally, explain why the minimum Sharpe ratio case applied to R^* is essentially the same as the maximum Sharpe ratio case applied to a suitably defined symmetric traded return.

Exercise 2 This exercise outlines a version of the theory of beta pricing without a traded money-market account. As in Section 2.2, the argument applies separately over a single period conditionally on each nonterminal spot. To simplify notation (and without loss of generality) we assume there is a single period ($T = 1$), so every cash flow x is a pair of a scalar x_0 and a random variable x_1. We fix throughout a reference market, which is arbitrage-free and therefore every marketed cash flow has a uniquely defined present value. Call the random variable x a **marketed payoff** if $(0, x)$ is a marketed cash flow, in which case $\Pi(x)$ denotes the present value of the cash flow $(0, x)$. This defines a linear functional Π on the set of marketed payoffs. (In terms of the present-value notation of Chapter 1, $\Pi(x)$ can be thought of as shorthand for $\Pi((0, x))$.) The market is implemented by $J \geq 2$ contracts with well-defined and linearly independent returns $R^1, \ldots, R^J$. Fixing an underlying full-support probability, let $\Sigma = [\Sigma_{ij}]$ denote the return variance-covariance matrix: $\Sigma_{ij} \equiv \operatorname{cov}(R^i, R^j)$, $i, j \in \{1, \ldots, J\}$. A **portfolio allocation** is any $\psi = (\psi^1, \ldots, \psi^J) \in \mathbb{R}^J$ such that $\sum_j \psi^j = 1$, generating the return $R^\psi \equiv \sum_j \psi^j R^j$. The set of traded returns is the linear manifold $\mathcal{R} \equiv \left\{R^\psi \mid \psi \in \mathbb{R}^J\right\}$. Assume throughout that there is *no traded money-market account:* $\operatorname{var}[R] > 0$ for all $R \in \mathcal{R}$. Let $L^{\mathcal{R}}$ denote the linear span of $\mathcal{R}$.

(a) Explain why $L^{\mathcal{R}}$ equals the set of marketed payoffs, and $x/\Pi(x) \in \mathcal{R}$ for every $x \in L^{\mathcal{R}}$ such that $\Pi(x) \neq 0$.

(b) Verify that covariance defines an inner product for the vector space $L^{\mathcal{R}}$. Treat $L^{\mathcal{R}}$ as an inner product space with the covariance inner product for the remainder of this exercise.

(c) Let x^Π denote the Riesz representation of the present-value functional Π in $L^{\mathcal{R}}$:

$$x^\Pi \in L^{\mathcal{R}} \quad \text{and} \quad \Pi(x) = \operatorname{cov}\left[x^\Pi, x\right] \text{ for all } x \in L^{\mathcal{R}}.$$

Explain why $\Pi\left(x^\Pi\right) > 0$ and therefore R^Π is well defined by

$$R^\Pi \equiv \frac{x^\Pi}{\Pi\left(x^\Pi\right)} \in \mathcal{R}.$$

Use Proposition B.1.6 to derive closed-form expressions, in terms of Σ, for the portfolio allocation that generates R^Π and the variance of R^Π.

(d) Show that $\operatorname{var}\left[R^\Pi\right] = \min\{\operatorname{var}[R] \mid R \in \mathcal{R}\}$ by using an orthogonal projection argument to argue that the minimum exists, is unique and is characterized by the orthogonality condition $R^\Pi \perp R - R^\Pi$ for all $R \in \mathcal{R}$, or equivalently,

$$\text{for all } R \in \mathcal{R},\ \operatorname{cov}\left[R^\Pi, R\right] = \operatorname{var}\left[R^\Pi\right].$$

Verify that R^Π satisfies this condition and is therefore the traded return of least variance. Also explain why the same conclusion can be reached by applying Corollary B.4.7.

(e) The return $R^* \in \mathcal{R}$ is a **frontier** return if

$$\operatorname{var}[R^*] = \min\{\operatorname{var}[R] \mid \mathbb{E}R = \mathbb{E}R^*, R \in \mathcal{R}\}.$$

Give a geometric interpretation of the property as an orthogonal projection and (briefly) explain why $R^* \in \mathcal{R}$ is a frontier return if and only if $R^* \perp R - R^*$ for every traded return R such that $\mathbb{E}R = \mathbb{E}R^*$. Call the market **degenerate** if all traded returns have the same mean, that is, there exists an m such that $\mathbb{E}R = m$ for all $R \in \mathcal{R}$. What is the set of frontier returns if the market is degenerate? Show that if the market is *not* degenerate, then the set of frontier returns is a line in $L^{\mathcal{R}}$ that contains and is orthogonal to R^{Π}.

Hint: The idea is to show that given a frontier return $R^* \neq R^{\Pi}$ and any given mean m, there is a point on the line through R^* and R^{Π} that is the projection of zero to the linear manifold $\{R \in \mathcal{R} \mid \mathbb{E}R = m\}$. The requisite orthogonality condition should follow from the corresponding orthogonality condition for R^* and the fact that R^{Π} is orthogonal to $\mathcal{R}$. This is the hard way—we will encounter an easier approach below.

(f) Let $x^{\mathbb{E}}$ be the Riesz representation of $\mathbb{E}$ restricted to $L^{\mathcal{R}}$, defined by the requirements:

$$x^{\mathbb{E}} \in L^{\mathcal{R}} \quad \text{and} \quad \mathbb{E}x = \operatorname{cov}\left[x^{\mathbb{E}}, x\right] \text{ for all } x \in L^{\mathcal{R}}.$$

Use Proposition B.1.6 to derive closed-form expressions, in terms of Σ and $\left(\mathbb{E}R^1, \ldots, \mathbb{E}R^J\right)$, for the row vector $\beta^{\mathbb{E}}$ such that $x^{\mathbb{E}} = \sum_j \beta_j^{\mathbb{E}} R^j$ as well as $\Pi\left(x^{\mathbb{E}}\right)$.

(g) Prove that the market is degenerate if and only if there exists a scalar m such that $x^{\mathbb{E}} = mx^{\Pi}$. Assume that the market is *not* degenerate for the remainder of this exercise.

(h) Use Corollary B.4.7 to show that $R^* \in \mathcal{R}$ is a frontier return if and only if it takes the form $\alpha x^{\Pi} + \beta x^{\mathbb{E}}$ for scalars α and β. Use this fact to give another proof that the set of frontier returns is a line in $L^{\mathcal{R}}$.

(i) Fix an arbitrary frontier return R^* other than R^{Π} and show that there exists a unique frontier return R^0 that is uncorrelated with R^*. Give a geometric interpretation of this result. How is R^0 positioned relative to R^{Π} and R^* on the frontier line? How does R^0 behave as R^* approaches R^{Π}?

(j) Suppose that R^* is a frontier return and R^0 is the unique frontier return that is uncorrelated with R^*. Show that

$$\mathbb{E}\left[R - R^0\right] = \frac{\operatorname{cov}\left[R^*, R\right]}{\operatorname{var}\left[R^*\right]} \mathbb{E}\left[R^* - R^0\right], \quad R \in \mathcal{R}.$$

Conversely, show that if this beta pricing equation holds for some $R^*, R^0 \in \mathcal{R}$, then R^* and R^0 are uncorrelated, R^* is necessarily a frontier return, and R^0 can always be selected to be a frontier return.

Exercise 3 Assume the setting of Section 2.2 with a single period ($T = 1$). The (arbitrage-free) market is implemented by $1 + J$ linearly independent contracts: an MMA with return $R^0 \equiv 1 + r$ and J risky contracts with returns $R^1, \ldots, R^J$. Define $1 + \mu^j \equiv \mathbb{E}R^j$ and $\Sigma_{ij} \equiv \operatorname{cov}[R^i, R^j]$, and write $\mu \equiv \left(\mu^1, \ldots, \mu^J\right)$, a row vector, and $\Sigma \equiv [\Sigma_{ij}]$, a $J \times J$ matrix. A **portfolio allocation** is any row vector $\psi = \left(\psi^1, \ldots, \psi^J\right) \in \mathbb{R}^J$, generating the return $R^\psi \equiv R^0 + \sum_j \psi^j \left(R^j - R^0\right)$. The set of traded returns is the linear manifold $\mathcal{R} \equiv \left\{R^\psi \mid \psi \in \mathbb{R}^J\right\}$. As in Exercise 2, the set of marketed payoffs is $L^\mathcal{R} \equiv \operatorname{span}(\mathcal{R})$ and for every $x \in L^\mathcal{R}$, we write $\Pi(x)$, rather than $\Pi((0, x))$, for the present value of x. Moreover, by the argument of Exercise 2(a), $x/\Pi(x) \in \mathcal{R}$ for every $x \in L^\mathcal{R}$ such that $\Pi(x) \neq 0$.

(a) Explain why covariance is *not* an inner product in $L^\mathcal{R}$. Instead, for the remainder of this exercise assume $L^\mathcal{R}$ is an inner product space with the inner product $\langle x \mid y \rangle \equiv \mathbb{E}[xy]$. As a notational convention, exponentiation takes precedence over expectation, and therefore the implied squared norm of x is $\mathbb{E}x^2 \equiv \mathbb{E}[x^2]$.

(b) Let x^Π denote the Riesz representation of Π in $L^\mathcal{R}$:

$$x^\Pi \in L^\mathcal{R} \quad \text{and} \quad \Pi(x) = \mathbb{E}\left[x^\Pi x\right] \text{ for all } x \in L^\mathcal{R}.$$

Define the corresponding return

$$R^\Pi \equiv \frac{x^\Pi}{\Pi\left(x^\Pi\right)} \in \mathcal{R}.$$

Use Proposition B.1.6 to derive closed-form expressions, in terms of r, μ and Σ, for the portfolio allocation that generates R^Π.

Hint: Show and use the fact that $x^\Pi - \mathbb{E}x^\Pi$ is the Riesz representation of the present-value functional in the linear subspace $\left\{x - \mathbb{E}x \mid x \in L^\mathcal{R}\right\}$.

(c) Use an orthogonal projection argument to show that

$$\mathbb{E}\left(R^\Pi\right)^2 = \min\left\{\mathbb{E}R^2 \mid R \in \mathcal{R}\right\}.$$

What is the associated orthogonality condition satisfied by R^Π?

(d) Show that the set of frontier returns is

$$\left\{\alpha R^0 + (1 - \alpha) R^\Pi \mid \alpha \in \mathbb{R}\right\}.$$

(e) Call the market **degenerate** if all traded returns have the same mean. Show that the set of frontier returns is a point if and only if the market is degenerate, and otherwise it is a line.

(f) Arguing directly from the definition of frontier returns, show that the set of frontier returns other than R^0 is exactly the set of traded returns of maximum absolute Sharpe ratio.

(g) The locus of all pairs $(\operatorname{stdev}[R^*], \mathbb{E}R^*)$ as R^* ranges over all frontier returns is known as the **minimum-variance frontier**. Describe and plot the minimum-variance frontier, and specify geometrically the exact location of the point $\left(\operatorname{stdev}\left[R^\Pi\right], \mathbb{E}R^\Pi\right)$.

Exercise 4 Consider the setting of Section 2.2. To simplify notation, assume there is a single period ($T = 1$), which entails no loss of generality since the argument of this exercise can be made over a single period conditionally on any nonterminal spot. In Proposition 2.2.3 we encountered the **beta** of a traded return R with respect to the traded return R^*:

$$\beta \equiv \frac{\text{cov}[R^*, R]}{\text{var}\,[R^*]}.$$

Suppose that in an empirical implementation of the beta-pricing equation of Proposition 2.2.3, a proxy $R^* + \varepsilon$ is used instead of a "true" frontier return R^*, where the error ε is judged to be small. The corresponding beta is

$$\beta^\varepsilon \equiv \frac{\text{cov}[R^* + \varepsilon, R]}{\text{var}\,[R^* + \varepsilon]}.$$

Give a simple example illustrating the claim that an arbitrarily small value of $\mathbb{E}\varepsilon^2$ is consistent with an arbitrarily large value of $|\beta^\varepsilon - \beta|$.

Exercise 5 (a) Show that an adapted process M is a martingale if and only if $M_0 = \mathbb{E}M_\tau$ for every stopping time $\tau : \Omega \to \{0, \ldots, T\}$. (Note that τ is not allowed the value ∞.)

(b) Suppose the market is implemented by the contracts $(\delta, V) \in \mathcal{L}^J \times \mathcal{L}^J$. Show that a strictly positive adapted process π is an SPD if and only if for every finite stopping time $\tau : \Omega \to \{0, \ldots, T\}$,

$$V_0 = \frac{1}{\pi_0}\mathbb{E}\left[\sum_{t=0}^{\tau-1} \pi_t \delta_t + \pi_\tau V_\tau\right].$$

Exercise 6 (Binomial pricing) This is a continuation of Exercise 7 of Chapter 1, whose setting and notation is assumed here. We also use the notation $R^0 \equiv 1 + r$, $Y \equiv 1 + y$ and

$$q \equiv \frac{(R^0/Y) - D}{U - D}. \tag{2.9.1}$$

The condition of part (b) of Exercise 7, which is implied by the no-arbitrage assumption, is equivalent to $q \in (0, 1)$. Assume that this condition holds for the remainder of this exercise.

(a) Compute all EMM-discount pairs for the given market. Is the market complete and why? Explain why the condition $1 + r \in (0, \infty)$ and $q \in (0, 1)$ is sufficient for the market to be arbitrage-free.

(b) For the remainder of this exercise, assume (Q, ρ) is an EMM-discount pair. Is Z a Markov process under Q and why?

(c) Consider a contract (δ^*, V^*) that is specified in terms of the payoff function $f_T : \mathcal{N}_T \to \mathbb{R}$ by

$$\delta_T^* = V_T^* = f_T(Z_T) \quad \text{and} \quad \delta_-^* = 0.$$

Use an EMM to show a pricing relationship of the form $V_t^* = f_t(Z_t)$, where the functions $f_t : \mathcal{N}_t \to (0, \infty)$ are computed recursively, backward in time, starting with the known terminal function f_T. Confirm that your result is consistent with that of Exercise 7.

(d) What is the order of magnitude of the number of operations needed to compute V_0^* in part (e)? How does that compare to the number of spots on the information tree? What key assumptions make this type of computational efficiency possible?

(e) Postulate an underlying probability P that is defined in terms of the constant $p \in (0, 1)$ by

$$P(\{\omega\}) = p^{N(\omega)} (1-p)^{T-N(\omega)}, \quad \omega \in \Omega,$$

where $N \equiv \sum_{t=1}^{T} b_t$. The process B is defined recursively by

$$B_0 = 0, \quad \Delta B_t = b_t \sqrt{\frac{1-p}{p}} - (1-b_t) \sqrt{\frac{p}{1-p}}, \quad t = 1, \ldots, T.$$

Verify that B is a dynamic orthonormal basis under P, and specify the set of all dynamically orthonormal bases in terms of B.

(f) Compute coefficients α and β so that

$$\frac{\Delta Z}{Z_-} = \alpha + \beta \Delta B.$$

Be as specific as you can, given the primitives of the model.

(g) Define the processes μ^R and σ^R by the return representation

$$\frac{\Delta G}{S_-} = \mu^R + \sigma^R \Delta B, \quad \mu^R \in \mathcal{P}_0, \quad \sigma^R \in \mathcal{P}_0^{1 \times d}.$$

Derive formulas for μ^R and σ^R in terms of α, β and Y.

(h) Assume that $Q = P^\eta$ is an EMM, where $\eta \in \mathcal{H}$. Compute $\mathbb{E}_{t-1}^Q [\Delta G_t / S_{t-1}]$ and use the fact that $\mathbb{E}_{t-1}^Q \Delta B_t = \eta$ to conclude that

$$\eta = \frac{\mu^R - r}{\sigma^R}.$$

Give an expression for η in terms of p and q, and use it together with the definition of ΔB to confirm that

$$1 - \eta \Delta B = \frac{q}{p} b + \frac{1-q}{1-p} (1-b),$$

and that the conditional density process $\xi_t = \mathbb{E}_t [dQ/dP]$ is given by

$$\xi_t = \left(\frac{q}{p} \right)^{N_t} \left(\frac{1-q}{1-p} \right)^{t-N_t}, \quad N_t = \sum_{u=0}^{t} b_u.$$

How can you use this result to recover the EMM expression you derived in part (a)?

Exercise 7 Assume the stochastic setting of Section 2.7.

(a) Suppose $x_t = x_0 + \alpha t + \beta B_t$, for constant $\alpha, \beta \in \mathbb{R}$. Using Ito's lemma to derive the Ito decomposition of $\exp(x_t)$, and then use this decomposition to compute $\mathbb{E}\exp(x_t)$. You can use without proof the facts that in this context, the expectation $\mathbb{E}$ and integral $\int_0^t$ can be interchanged, and the local martingale part in the Ito expansion of $\exp(x_t)$ is in fact a martingale. Jensen's inequality requires that the ratio $\mathbb{E}\exp(x_t) / \exp(\mathbb{E}x_t)$ is greater than one. Explain how your calculation quantifies this ratio.

(b) Suppose x_t is a strictly positive Ito process and $\alpha \in \mathbb{R}$. Use Ito's calculus to give expressions for $d\log x_t$ and dx_t^α / x_t^α as a function of dx_t/x_t. Show how these expressions simplify when x is geometric Brownian motion with drift: $dx_t/x_t = \mu dt + \sigma dB_t$ for constant μ and σ.

Exercise 8 Use integration by parts to show the equivalence of the budget equation (2.8.2) and the budget equation (2.8.4) after the change of unit of account implied by the strictly positive Ito process π.

Exercise 9 This exercise provides some insight on the relationship between the Brownian model of Section 2.8 and a high-frequency version of the binomial model of Exercise 5.

(a) As preparation, you will prove a special version of the strong law of large numbers. Suppose $x_{n,N}$, $n = 1, 2, \ldots N$; $N = 1, 2, \ldots$, are i.i.d. (that is, identically distributed and stochastic independent) random variables. These are defined on a common state space with a given probability measure P and corresponding expectation operator $\mathbb{E}$. Assume that for all n, N, $\mathbb{E}x_{n,N} = 0$ and $\mathbb{E}\left[x_{n,N}^2\right] = 1$. Moreover, assume that there exists a constant C such that for all n, N, $|x_{n,N}| \leq C$ everywhere. (The argument you are about to give applies if the last assumption is weakened to $\mathbb{E}\left[x_{n,N}^4\right] < \infty$, but we do not need this generality here.) Define the averages

$$\bar{x}_N \equiv \frac{1}{N}\sum_{n=1}^{N} x_{n,N}, \quad N = 1, 2, \ldots$$

Prove that the random variable $\sum_{N=1}^{\infty} \bar{x}_N^4$ has finite expectation and explain why it must then be true that $\lim_{N\to\infty} \bar{x}_N = 0$ with probability one. You can use the fact that $\mathbb{E}\sum_{N=1}^{\infty} \bar{x}_N^4 = \sum_{N=1}^{\infty} \mathbb{E}\left[\bar{x}_N^4\right]$, which is a consequence of the monotone convergence theorem.

(b) For each $N = 1, 2, \ldots$, consider version of the model of Exercise 5 (introduced in Exercise 7 of Chapter 1), where instead of normalizing the time unit to correspond to one period, we assume that $T = 1$ and there are N periods. The time length of each period is therefore $1/N$. We wish to select the remaining parameters so that the Brownian model of Section 2.8 is approximated as N gets large (without actually proving a convergence result). The transition probability p is the same for all N. Suppose $z_{n,N}$, $n = 1, 2, \ldots, N$; $N = 1, 2, \ldots$

are i.i.d. random variables (under some given probability P on some common state space), and

$$p = P\left[z_{n,N} = \sqrt{\frac{1-p}{p}}\right] = 1 - P\left[z_{n,N} = -\sqrt{\frac{p}{1-p}}\right].$$

The martingale basis for the N^{th} model is B^N, where

$$B_0^N = 0, \quad B_{\frac{n}{N}}^N - B_{\frac{n-1}{N}}^N = \sqrt{\frac{1}{N}} z_{n,N}, \quad n = 1, \ldots, N. \tag{2.9.2}$$

The stock return parameters μ^R and σ^R are specified in terms of constants μ and σ by $\mu^R = \mu/N$ and $\sigma^R = \sigma/\sqrt{N}$. The stock cumulative return process R^N is given by

$$R_{\frac{n}{N}}^N - R_{\frac{n-1}{N}}^N \equiv \frac{S_{\frac{n}{N}} e^{\frac{y}{N}}}{S_{\frac{n-1}{N}}} - 1 = \mu \frac{1}{N} + \sigma \sqrt{\frac{1}{N}} z_{n,N}, \quad R_0^N \equiv 1. \tag{2.9.3}$$

The interest rate process of the MMA of the N^{th} model is analogously scaled, so that one dollar invested in the account at time $(n-1)/N$ becomes $1 + r/N$ at time n/N. The constant parameters (r, μ, σ) do not vary with N. Using part (a), show that with probability one,

$$\lim_{N \to \infty} \sum_{n=1}^{N} \left(R_{\frac{n}{N}}^N - R_{\frac{n-1}{N}}^N \right)^2 = \sigma^2.$$

This shows that no matter what the probability $p \in (0,1)$ is, and given any required level of precision, we can take N large enough so that the quadratic variation of R^N equals σ^2 up to the required precision. In the continuous-time limit, where R is an Ito process, this fact corresponds to the quadratic variation calculation $(dR_t)^2 = \sigma^2 dt$.

(c) Explain why in the finite binomial model the choice of the parameter $p \in (0, \infty)$ is irrelevant the pricing of the European call option. Also explain why in the Brownian model, the value of a (or μ) is irrelevant for the pricing of the European call option. (The absence of μ in the Black-Scholes formula reflects this fact. Here you are asked to provide a deeper reason for this absence.)

(d) Given the insight of part (c), set $p = 1/2$. Recall that the stock return S_t/S_{t-h} over every nonterminal period is either $U \equiv U^N$ or $D \equiv D^N$, where the superscript N has been added to emphasize the dependence of these parameters on N. (Technically speaking, the convention $S_T = 0$ requires us to slightly modify this statement, but the last period becomes infinitesimal in the limit and does not matter.) What are suitable expressions for U^N and D^N in terms of the parameters (r, y, σ) so that, as $N \to \infty$, the binomial model approximates the Brownian model of Section 2.8 with $a = 0$. Provide a brief explanation of your claim, without proving anything formal.

CHAPTER 3

Optimality and Equilibrium Pricing

We have so far discussed markets that contain no cash flows that are desirable in the sense of arbitrage. In this chapter, we expand the notion of a desirable cash flow by introducing preferences. The absence of traded desirable cash flows relative to a reference consumption plan defines the plan's optimality and the related problem of optimal portfolio selection. Multi-agent optimality together with a market-clearing condition defines competitive equilibrium.

3.1. Preferences and optimality

We adopt the setting of Chapter 1, with information unfolding over times $0, \ldots, T$, defining periods $1, \ldots, T$ and $1 + K$ spots. There is no information at time zero and the state is revealed at time T. A **consumption plan** is an adapted process. The set of all consumption plans is therefore $\mathcal{L}$, which can be identified with $\mathbb{R}^{1+K}$. A consumption plan represents an agent's total contingent consumption at each spot. A cash flow available in a market can be incrementally added to a given consumption plan to modify it to a preferred consumption plan. In reality, consumption consists of bundles of multiple goods. By assuming that consumption is one-dimensional at every spot, we are implicitly taking relative spot prices of goods as given and we measure consumption in some unit of account.

A **consumption set** is a set C of consumption plans such that for all $c \in C$, if x is an arbitrage, then $c + x \in C$. Given $c \in C$, we wish to specify a set $\mathcal{D}(c)$ with the interpretation that every $x \in \mathcal{D}(c)$ represents a **desirable** cash flow in the sense that $c+x$ is a consumption plan that is strictly preferred to c from the perspective of time zero. We impose some minimal requirements on preferences, listed below. Further structure will be imposed as needed later on.

Definition 3.1.1. A **preference correspondence** is a function $\mathcal{D}$ whose domain, denoted $\mathrm{dom}(\mathcal{D})$, is a consumption set, and such that for all $c \in \mathrm{dom}(\mathcal{D})$, $\mathcal{D}(c)$ is a set of cash flows satisfying

- admissibility: $x \in \mathcal{D}(c)$ implies $c + x \in \mathrm{dom}(\mathcal{D})$.
- irreflexivity: $0 \notin \mathcal{D}(c)$.
- continuity: $\mathcal{D}(c)$ is open.
- monotonicity: For every arbitrage cash flow y, if $x = 0$ or $x \in \mathcal{D}(c)$ then $x + y \in \mathcal{D}(c)$.

The first two properties state that $c + x$ must be an admissible consumption plan in order to be strictly preferred to c, and c cannot be strictly preferred to itself. The third condition states that if $c+x$ is strictly preferred to c then there is a sufficiently small tolerance $\epsilon > 0$ such that $\|x - x'\| < \epsilon$ implies that $c + x'$ is also strictly preferred to c. Finally, the fourth condition means that the agent always strictly prefers to increase consumption at some spot, provided consumption is not reduced at any other spot.

We henceforth fix a reference market X, relative to which we formulate notions of individual and allocational optimality.

DEFINITION 3.1.2. A consumption plan c is **optimal** for the preference correspondence $\mathcal{D}$ given X if $c \in \mathrm{dom}\,(\mathcal{D})$ and $X \cap \mathcal{D}\,(c) = \emptyset$.

The first part of the optimality condition requires that c is an admissible consumption plan, while the second part states that there is no trade that results in a desirable incremental cash flow relative to c. Since, by monotonicity, $\mathcal{D}\,(c)$ contains all arbitrage cash flows, optimality of c implies the no-arbitrage condition $X \cap \mathbb{R}^{1+K}_{+} = \{0\}$.

Individual optimality extends usefully to a notion of allocational optimality as follows. Fix a positive integer I. Informally, we think of $i \in \{1, \ldots, I\}$ as labeling agents representing potential market participants. A **consumption allocation** is a tuple $c = \left(c^1, \ldots, c^I\right)$, with the interpretation that c^i is a consumption plan for agent i. A **preference profile** is a preference correspondence tuple $\left(\mathcal{D}^1, \ldots, \mathcal{D}^I\right)$, with $\mathcal{D}^i$ representing the admissible consumption plans and preferences of agent i. The sum of the $\mathcal{D}^i$ is the correspondence $\mathcal{D}$ on $\mathrm{dom}\,(\mathcal{D}) \equiv \prod_{i=1}^{I} \mathrm{dom}\,(\mathcal{D}^i)$ defined by

$$\mathcal{D}\,(c) \equiv \sum_{i=1}^{I} \mathcal{D}^i\left(c^i\right) \equiv \left\{ \sum_{i=1}^{I} x^i \mid x^i \in \mathcal{D}^i\left(c^i\right) \text{ for all } i \right\}. \tag{3.1.1}$$

DEFINITION 3.1.3. An allocation $c = \left(c^1, \ldots, c^I\right)$ is **optimal** for the preference profile $\left(\mathcal{D}^1, \ldots, \mathcal{D}^I\right)$ given the market X if $c \in \mathrm{dom}\,(\mathcal{D})$ and $X \cap \mathcal{D}\,(c) = \emptyset$.

The existence of some $x \in X \cap \mathcal{D}\,(c)$ can be thought of as a profitable market-making opportunity. Suppose $x = \sum_{i=1}^{I} x^i$ where $x^i \in \mathcal{D}^i\,(c^i)$. A market maker can offer x^i to each agent i in exchange for some positive amount, since agent i strictly prefers to add x^i to c^i. The market maker can then simply trade away the aggregate cash flow x, since $x \in X$. The net result is a positive profit for the market maker, a type of arbitrage that is not in X but can be obtained by using X and simultaneously contracting with the I agents in an incentive compatible way. Optimality of the allocation c given X means there are no market-making opportunities of this sort.

The monotonicity and continuity properties of preference correspondences imply that the optimality of an allocation c is equivalent to the apparently stronger condition of optimality of the allocation for every subset of the agents.

PROPOSITION 3.1.4. *The allocation* $\left(c^1, \ldots, c^I\right)$ *is optimal for the preference profile* $\left(\mathcal{D}^1, \ldots, \mathcal{D}^I\right)$ *given* X *if and only if for every* $A \subseteq \{1, \ldots, I\}$*, the allocation* $(c^i)_{i \in A}$ *is optimal for* $(\mathcal{D}^i)_{i \in A}$ *given* X*.*

PROOF. Suppose $x = \sum_{i \in A} x^i \in X$, where $x^i \in \mathcal{D}^i(c^i)$ for all $i \in A \subseteq \{1, \ldots, I\}$. For $i \notin A$ let $x^i = 0$. The idea is to increase every x^i at the expense of a single agent $\alpha \in A$ by a sufficiently small amount so that agent α still finds the resulting incremental cash flow desirable. More formally, let y be any arbitrage cash flow, say $y = 1$, and fix any agent $\alpha \in A$. Given that $\mathcal{D}^\alpha(c^\alpha)$ is an open set, choose scalar $\epsilon > 0$ small enough so that $\bar{x}^\alpha \equiv x^\alpha - (1 - I)\epsilon y \in \mathcal{D}^\alpha(c^\alpha)$, and for every agent $i \neq \alpha$ let $\bar{x}^i \equiv x^i + \epsilon y$. By preference monotonicity, $\bar{x}^i \in \mathcal{D}(c^i)$ for all i, and by construction $\sum_{i=1}^I \bar{x}^i = x \in X$. Therefore, $\left(c^1, \ldots, c^I\right)$ is not optimal given X. The converse is immediate. □

Since individual optimality is the same as the degenerate case of single-agent allocational optimality, it follows that allocational optimality implies individual optimality.

COROLLARY 3.1.5. *If* $\left(c^1, \ldots, c^I\right)$ *is optimal for* $\left(\mathcal{D}^1, \ldots, \mathcal{D}^I\right)$ *given* X*, then for all* $i \in \{1, \ldots, I\}$*,* c^i *is optimal for* $\mathcal{D}^i$ *given* X*.*

We proceed under the assumption that c represents either a consumption plan of a single agent with preference correspondence $\mathcal{D}$, or an allocation $c = \left(c^1, \ldots, c^I\right)$ for I agents with preference profile $\left(\mathcal{D}^1, \ldots, \mathcal{D}^I\right)$. In the latter case, $\mathcal{D}$ denotes the sum of the $\mathcal{D}^i$ as specified in (3.1.1) with $\operatorname{dom}(\mathcal{D}) \equiv \prod_{i=1}^I \operatorname{dom}(\mathcal{D}^i)$. In the following discussion, any statement of optimality of c is understood to be for $\mathcal{D}$ if c is a consumption plan and for $\left(\mathcal{D}^1, \ldots, \mathcal{D}^I\right)$ if c is an allocation. With these conventions, we introduce a dual notion of optimality that will be key in relating optimality given a market to optimality given a budget constraint in terms of present values.

DEFINITION 3.1.6. For any linear functional $\Pi : \mathcal{L} \to \mathbb{R}$, the consumption plan or allocation c is **Π-optimal** if $c \in \operatorname{dom}(\mathcal{D})$ and $x \in \mathcal{D}(c)$ implies $\Pi(x) > 0$.

Since $\mathcal{D}(c)$ contains every arbitrage-cash flow, if c is Π-optimal, then Π is necessarily positive ($\Pi(x) > 0$ if x is an arbitrage). A consequence of this is that Π-optimality is the same as optimality given the complete market for which Π is the unique present-value function.

LEMMA 3.1.7. *A consumption plan or allocation is* Π*-optimal if and only if it is optimal given the market* $X^\Pi \equiv \{x \in \mathcal{L} \mid \Pi(x) = 0\}$*.*

PROOF. Suppose c is not Π-optimal and therefore $\Pi(x) \leq 0$ for some $x \in \mathcal{D}(c)$. Then the constant cash flow $\delta = -\Pi(x)$ is either zero or an arbitrage and therefore $x + \delta \in \mathcal{D}(c) \cap X^{\Pi}$, which implies c is not optimal given X^{Π}. The converse is immediate. □

The lemma allows us to apply Proposition 3.1.7 and its corollary to Π optimality. Therefore, if an allocation $c = \left(c^1, \ldots, c^I\right)$ is Π-optimal, then c^i is Π-optimal for every agent i. The fact that the Π in this statement is the same for every agent will be key.

The duality between optimality given a market and Π-optimality for a present-value function Π has the geometric structure of the duality between an arbitrage-free market and a present-value function (Theorem 1.4.3). The set $\mathcal{D}(c)$ enlarges the set of arbitrage cash flows and the condition that $\Pi(x)$ is positive for all $x \in \mathcal{D}(c)$ strengthens the condition that Π is positive. A present-value function Π (strictly) separates the market from the set of arbitrage cash flows, while a present value Π such that c is Π-optimal separates the market from $\mathcal{D}(c)$. This picture is the basis for the following proposition, where we take the reference market X as given, and we say that the consumption plan or allocation c is **optimal** to mean that c is optimal given the market X. Note that the second part assumes convexity of $\mathcal{D}(c)$, which in the multi-agent case is implied by the convexity of every $\mathcal{D}^i(c)$.

PROPOSITION 3.1.8. *The following are true for all $c \in dom(\mathcal{D})$.*

(1) *If c is Π-optimal for some present-value function Π, then c is optimal.*
(2) *If c is optimal and $\mathcal{D}(c)$ is convex, then there exists a present-value function Π such that c is Π-optimal.*
(3) *Suppose the market complete and Π is the unique present-value function. Then c is optimal if and only if it is Π-optimal.*

PROOF. (1) If $x \in \mathcal{D}(c)$, then $\Pi(x) > 0$ and therefore $x \notin X$.

(2) Suppose c is optimal and $\mathcal{D}(c)$ is convex. Viewing the space of adapted processes as $\mathbb{R}^{1+K}$ with the Euclidean inner product, the separating hyperplane theorem (Corollary B.5.3) implies that there exists a nonzero vector $p \in \mathbb{R}^{1+K}$ such that $p \cdot x \leq 0$ for all $x \in X$ and $p \cdot y \geq 0$ for all $y \in \mathcal{D}(c)$. Suppose that $p \cdot x = 0$ for some $x \in \mathcal{D}(c)$. Since $\mathcal{D}(c)$ is open, there is $\epsilon > 0$ such that $x - \epsilon p \in \mathcal{D}(c)$, leading to the absurdity $0 \leq p \cdot (x - \epsilon p) < 0$. Therefore, $p \cdot x > 0$ for all $x \in \mathcal{D}(c)$, which implies that p is strictly positive (since $\mathcal{D}(c)$ contains every arbitrage). It follows that $\Pi(x) = p \cdot x / p_0$ defines a present-value function such that c is Π-optimal.

(3) This is Lemma 3.1.7 with $X = X^{\Pi}$. □

A preference correspondence $\mathcal{D}$ has been defined from the perspective of time zero. A preference correspondence $\mathcal{D}_{F,t}$ can be analogously

defined from the perspective of spot (F,t), with $\mathcal{D}_{F,t}(c) \subseteq \mathcal{L}_{F,t}$ for every $c \in C$. Given a corresponding spot-(F,t) market $X_{F,t} \subseteq \mathcal{L}_{F,t}$, a key observation is that if $X_{F,t} \subseteq X$ and $\mathcal{D}_{F,t}(c) \subseteq \mathcal{D}(c)$, then $X \cap \mathcal{D}(c) = \emptyset$ implies $X_{F,t} \cap \mathcal{D}_{F,t}(c) = \emptyset$. This condition states that if c is optimal from the perspective of spot zero, then it is also optimal from the perspective of spot (F,t). The condition guarantees that an agent that selects an optimal consumption plan c at time zero has no incentive to deviate from that choice as uncertainty unfolds. In discussing Proposition 1.2.5, we saw that $X_{F,t} \subseteq X$ is a dynamic consistency assumption on the market: If an incremental cash flow x is available in the market at spot (F,t), then it is also available at time zero, since the agent can make contingent plans to carry out whatever trades generate x if spot (F,t) is realized. Analogously, $\mathcal{D}_{F,t}(c) \subseteq \mathcal{D}(c)$ is a **dynamic consistency** assumption on preferences: If from the perspective of spot (F,t), consumption plan $c + x$ is strictly preferred to c for some incremental cash flow $x \in \mathcal{L}_{F,t}$, then the agent at time zero anticipates this preference contingent on spot (F,t), the only contingency on which x is non-zero, and therefore decides that $c + x$ is strictly preferred to c from the perspective of time zero as well.

A useful way of thinking about dynamic choice is as if the decision maker is multiple agents, one for every spot. Spot-(F,t) agent has preferences represented by $\mathcal{D}_{F,t}$ and selects trades in $X_{F,t}$. It is not hard to envision situations where dynamic consistency is violated. For example, suppose that the time-zero copy of the agent considers a contingent trade at spot (F,t) optimal, let's say a trade to rebalance into a falling stock market, yet, the spot-(F,t) copy of the same agent, faced with immediate market losses becomes afraid and stays out of the market. (As former heavyweight world champion Mike Tyson put it, "Everyone has a plan 'till they get punched in the mouth.") Perhaps the spot-zero agent has poor foresight and would have chosen differently given a better understanding of the future self. Alternatively, the spot-zero agent may be well aware of the role of future temptations and may therefore seek to commit[1] to a plan at time zero in a way that cannot be reversed at spot (F,t). The strategic interaction among copies of the same agent can become complex (and interesting). Here we bypass this complexity, by removing any conflict among copies of the same agent. Dynamic consistency ensures that what is best for the spot-zero copy of the agent is also best for the spot-(F,t) agent, and is therefore sufficient to only consider time-zero optimal decisions. Analogous reasoning applies to the notions of allocational optimality and equilibrium, which we therefore only discuss from the perspective of time zero.

[1]The concept is discussed by Strotz [1957], who quotes from Homer's Odyssey. Odysseus, being well aware his future self cannot resist the song of the Sirens, instructs his crew to tie him to the mast.

3.2. Equilibrium

In order to introduce a simple notion of (competitive) equilibrium, we formally define an **agent** to be a pair of a preference correspondence and a consumption plan, called the agent's **endowment**. We take as primitive the I agents

$$\left(\mathcal{D}^1, e^1\right), \ldots, \left(\mathcal{D}^I, e^I\right). \tag{3.2.1}$$

The initial allocation $\left(e^1, \ldots, e^I\right)$ can be modified to a new allocation $c \equiv \left(c^1, \ldots, c^I\right)$ through access to a market X.

DEFINITION 3.2.1. An allocation c is **X-feasible** if $c^i - e^i \in X$ for all i, and **clears the market** if $\sum_{i=1}^I c^i = e$, where $e \equiv \sum_{i=1}^I e^i$ is the **aggregate endowment**. An **equilibrium** is a pair (X, c), where X is a market, c is an X-feasible allocation that clears the market, and for all i, c^i is optimal for $\mathcal{D}^i$ given X.

Equilibrium is commonly formulated in the literature in terms of contracts implementing the market, whose dividend processes are given and whose prices are set in equilibrium. The following example outlines the relationship of this approach to the equilibrium notion just introduced.

EXAMPLE 3.2.2 (Contract-market equilibrium). Consider J contracts with given dividend processes $\delta \equiv \left(\delta^1, \ldots, \delta^J\right)'$. A **contract-market equilibrium** is a corresponding vector of value processes $V \equiv (V^1, \ldots, V^J)'$ and trades $\theta \equiv (\theta^1, \ldots, \theta^J)$ in the contracts (δ, V) such that $\sum_i \theta^i = 0$ and θ^i is an optimal trading strategy for agent i, in the sense that if θ^i generates cash flow x^i and $c^i \equiv e^i + x^i$, then $X(\delta, V) \cap \mathcal{D}^i(c^i) = \emptyset$. Recall that $X(\delta, V)$ denotes the market implemented by (δ, V). The last optimality condition, therefore, states that there is no trading strategy in (δ, V) that agent i desires to add to θ^i. If (V, θ) is a contract-market equilibrium, then $\sum_i \theta^i = 0$ implies $\sum_i x^i = 0$, and therefore $(X(\delta, V), c)$ is an equilibrium. In the converse direction, suppose that $(X(\delta, V), c)$ is an equilibrium and for each agent i, let θ^i be a trade that **finances** c^i, meaning that θ^i generates a cash flow $x^i \in X(\delta, V)$ such that $c^i = e^i + x^i$. While it is clear that c^i is optimal for agent i given the market $X(\delta, V)$ and that $\sum_i x^i = 0$, it is not necessary that $\sum_i \theta^i = 0$, since there are potentially more than one trading strategies generating a given traded cash flow. It is possible, however, to choose $\theta = \left(\theta^1, \ldots, \theta^J\right)$ so that $\sum_i \theta^i = 0$, and therefore so that (V, θ) is a contract-market equilibrium. The trick is to pick θ^i to finance c^i for $i > 1$, and then define $\theta^1 = -\sum_{i>1} \theta^i$. The fact that c clears the market implies that θ^1 finances c^1, and we have the desired contract-market equilibrium. $\diamondsuit$

Market equilibrium as defined here is closely related to the classical Walras [1874] notion of a competitive exchange equilibrium in the absence of time and uncertainty. For example, given an initial allocation of bread and wine, a competitive exchange equilibrium is a price for each good and a new allocation that is resource feasible (no bread or wine is created) and individually optimal (agents cannot do better by selling some wine to buy more bread or vice versa). A key insight of Arrow [1953, 1963] and Debreu [1959] was that the same notion can be applied in contexts with time and uncertainty by reinterpreting the notion of a good, in the example wine or bread, to be what we have called Arrow cash flows (also known as Arrow-Debreu securities). Bundles of goods correspond to cash flows and prices of goods correspond to state prices. A state price vector p can always be identified with the corresponding linear functional $\Pi(c) = p \cdot c$, which leads to the following equilibrium notion.

DEFINITION 3.2.3. An **Arrow-Debreu equilibrium** is a pair (Π, c) of a linear functional $\Pi : \mathcal{L} \to \mathbb{R}$ and an allocation $c \equiv \left(c^1, \ldots, c^I\right)$ such that $\sum_{i=1}^{I} c^i \leq e$ and for all i, $\Pi(c^i) \leq \Pi(e^i)$ and

$$\tilde{c} - c^i \in \mathcal{D}^i\left(c^i\right) \quad \Longrightarrow \quad \Pi(\tilde{c}) > \Pi\left(e^i\right). \tag{3.2.2}$$

The following equivalent form of this definition will be useful for our purposes.

LEMMA 3.2.4. *The pair (Π, c) is an Arrow-Debreu equilibrium if and only if $\Pi : \mathcal{L} \to \mathbb{R}$ is a positive linear functional, c clears the market, and for all i, c^i is Π-optimal for $\mathcal{D}^i$ and $\Pi(c^i) = \Pi(e^i)$.*

PROOF. Suppose (Π, c) is an Arrow-Debreu equilibrium and x is an arbitrage. For every scalar $\epsilon > 0$, $\epsilon x \in \mathcal{D}^i(c^i)$ and therefore

$$\Pi\left(c^i\right) + \epsilon\Pi(x) = \Pi\left(c^i + \epsilon x\right) > \Pi\left(e^i\right) \geq \Pi\left(c^i\right).$$

This implies that $\Pi(x) > 0$ and, since x can be any arbitrage, that Π is positive. Letting $\epsilon \downarrow 0$, we conclude that $\Pi(c^i) = \Pi(e^i)$ for all i. Individual Π-optimality and market clearing follow. The converse claim is immediate. □

A positive linear functional Π on $\mathcal{L}$ is the present-value function for the complete market $X^\Pi \equiv \{x \in \mathcal{L} \mid \Pi(x) = 0\}$. By Proposition 3.1.8 and the above lemma, it follows that (Π, c) is an Arrow-Debreu equilibrium if and only if $\left(X^\Pi, c\right)$ is an equilibrium. How does an equilibrium (X, c) relate to an Arrow-Debreu equilibrium if X is not complete? By Proposition 3.1.8, assuming convex preferences, individual optimality implies Π-optimality for some present-value function Π, which need not be common for all agents, unless the entire allocation c is optimal given the market. This leads us to the notion of an effectively compete market equilibrium.

DEFINITION 3.2.5. An equilibrium (X, c) is **effectively complete** if the allocation c is optimal for $\left(\mathcal{D}^1, \ldots, \mathcal{D}^I\right)$ given the market X.

As in the last section, we define $\mathcal{D} \equiv \sum_{i=1}^{I} \mathcal{D}^i$ and note that $\mathcal{D}(c)$ is convex if every $\mathcal{D}^i(c^i)$ is convex. The following claims are a consequence of Proposition 3.1.8 and our earlier discussion.

PROPOSITION 3.2.6. *The following are true for any market X.*

1. *If the allocation c is X-feasible and there exists a present-value function Π for X such that (Π, c) is an Arrow-Debreu equilibrium, then (X, c) is an effectively complete market equilibrium.*
2. *Suppose $\mathcal{D}(c)$ is convex. If (X, c) is an effectively complete-market equilibrium, then there exists a present-value function Π for X such that (Π, c) is an Arrow-Debreu equilibrium.*
3. *Suppose the market X is complete market with (necessarily unique) present-value function Π. Then (X, c) is an equilibrium if and only if (Π, c) is an Arrow-Debreu equilibrium.*

COROLLARY 3.2.7. *Suppose (X, c) is an equilibrium. If there exists exists a complete market $\bar{X} \supseteq X$ such that $\left(\bar{X}, c\right)$ is an equilibrium, then (X, c) is an effectively complete market equilibrium. The converse claim is also true if $\mathcal{D}(c)$ is convex.*

We conclude this section with two highly stylized examples of (effectively complete) equilibrium pricing that have played an important role in the early development of asset pricing theory. The first one is a version of the CAPM (Capital Asset Pricing Model), which is a myopic, single-period model in which covariance with the market portfolio explains expected excess returns relative to a traded (credit-risk-free) money market account. The second example, which works just as well in the multi-period case, is one in which individual optimality can be characterized as the optimality of the aggregate endowment for a suitably constructed "representative" agent. Representative-agent arguments of this type have been used extensively in the literature to justify the formulation of simple single-agent equilibrium models that relate aggregate consumption to market risk premia and interest rates. Both examples rely on special preference and endowment structure and a high degree of similarity among agents.

EXAMPLE 3.2.8. (CAPM) There is a single period ($T = 1$) and an underlying full-support probability, which can be thought of as representing common beliefs among the I agents. Consider an equilibrium (X, c) with a traded money-market account (MMA) defining the short rate r and return $R^0 \equiv 1 + r$. The set of **traded returns** is

$$\mathcal{R} \equiv \{-x_1/x_0 \mid x \in X,\ x_0 \neq 0\}.$$

Each endowment e^i is assumed to be marketed, with present value w^i (which is uniquely defined since X is arbitrage-free). The present

value of the aggregate endowment e is therefore $w \equiv \sum_{i=1}^{I} w^i$. The **market return** R^m is the return of the traded contract whose payoff is the time-one aggregate endowment e_1. Since e_1 has present value $w - e_0$, $R^m \equiv e_1/(w - e_0) \in \mathcal{R}$. To ensure that the market return is well-defined and has positive variance, we assume that

$$\text{var}[e_1] > 0 \quad \text{and} \quad e_0 \neq w. \tag{3.2.3}$$

The CAPM states that R^m is a beta-pricing return:

$$\mathbb{E}R - R^0 = \frac{\text{cov}(R, R^m)}{\text{var}(R^m)} \left(\mathbb{E}R^m - R^0\right), \quad R \in \mathcal{R}. \tag{3.2.4}$$

We will show the necessity of the CAPM with $\mathbb{E}R^m \neq R^0$ under the additional key assumption that all agents are **variance averse:**

$$\mathcal{D}^i(c) \supseteq \{x \in \mathcal{L} \mid (x_0, \mathbb{E}x_1) = (0,0) \text{ and } \text{var}[c_1 + x_1] < \text{var}[c_1]\}.$$

Moreover, assuming preference transitivity, we will show that the equilibrium is necessarily effectively complete.

Ignoring technicalities, the essential argument is simple given the beta-pricing discussion of Section 2.2. Because of the assumptions of common beliefs, marketed endowments and variance aversion, in equilibrium, agents sell their endowments and buy plans on the minimum-variance frontier. Such frontier plans are characterized by two-fund separation: All agents can finance their equilibrium consumption by holding the MMA and the same frontier risky investment. The key is that these positions can be added up. By market clearing, it follows that the market return takes the same form and must therefore also be a frontier return. The CAPM equation (3.2.4) then follows from Proposition 2.2.3.

We proceed with the technical details, starting with a review of minimum-variance analysis in the current context. On the space of random variables, we use the inner product $\langle x \mid y \rangle = \mathbb{E}[xy]$ with induced norm $\|\cdot\|$. Let x_1^Π denote the Riesz representation of the present-value functional on $X_1 \equiv \{x_1 \mid x \in X\}$, which is the unique $x_1^\Pi \in X_1$ such that $-x_0 = \langle x_1^\Pi \mid x_1 \rangle$ for all $x \in X$. Define $x^\Pi \in X$ by letting $-x_0^\Pi = \langle x_1^\Pi \mid x_1^\Pi \rangle$, a positive number, since $x_1^\Pi \neq 0$ by the MMA pricing equation $\langle x_1^\Pi \mid R^0 \rangle = 1$. The return $R^\Pi \equiv -x_1^\Pi / x_0^\Pi$ is therefore well-defined. Call $x \in X$ a **frontier cash flow** if $\text{var}[x_1]$ minimizes $\text{var}[y_1]$ as y ranges over X subject to the constraints $y_0 = x_0$ and $\mathbb{E}y_1 = \mathbb{E}x_1$, a condition that can be equivalently stated as

$$\|x(1)\| = \min\left\{\|z\| \mid \langle x_1^\Pi \mid z \rangle = -x_0,\ \langle 1 \mid z \rangle = \mathbb{E}x_1,\ z \in X_1\right\}.$$

By Corollary B.4.7, $x \in X$ is a frontier cash flow if and only if $x_1 \in \text{span}\left(x_1^\Pi, 1\right) = \text{span}\left(R^\Pi, R^0\right)$.

The main CAPM argument follows. By the marketed-endowment assumption, $c^i = w^i 1_{\Omega \times \{0\}} + x^i$, where $w^i \in \mathbb{R}$ and $x^i \in X$. By the optimality of c^i for agent i and variance-aversion, x^i is a frontier cash

flow and therefore $x_1^i = a^i R^\Pi + b^i R^0$ for some $a^i, b^i \in \mathbb{R}$. Let $a \equiv \sum_i a^i$ and $b \equiv \sum_i b^i$. Adding up over all agents and using market clearing and the regularity assumption (3.2.3), we find

$$e = w1_{\Omega\times\{0\}} + x, \quad \text{where } x_1 = aR^\Pi + bR^0 \text{ and } a \neq 0. \tag{3.2.5}$$

By the same regularity assumption, $x_0 \neq 0$ and the market return R^m has positive variance. This shows that $(-1, R^m)$ is a frontier cash flow and therefore R^m is a **frontier return**, in the sense that for all $R \in \mathcal{R}$, $\mathbb{E}R = \mathbb{E}R^m$ implies $\text{var}[R] \leq \text{var}[R^m]$. As in Lemma 2.2.2, this means that R^m is the projection of zero onto the linear manifold $\mathcal{R}_m \equiv \{R \in \mathcal{R} \mid \mathbb{E}R = \mathbb{E}R^m\}$, a condition that is equivalent to the orthogonality of R^m to $\mathcal{R}_m$, which is in turn equivalent to $\text{cov}[R, R^m] = \text{var}[R^m]$ for all $R \in \mathcal{R}_m$. The CAPM equation (3.2.4) with $\mathbb{E}R^m \neq R^0$ now follows exactly as in the $(1 \implies 2)$ part of Proposition 2.2.3.

Finally, we show that (X, c) is an effectively complete market equilibrium, assuming preference **transitivity:** if $y_1 \in \mathcal{D}^i(c^i)$ and $y_2 \in \mathcal{D}^i(c^i + y_1)$ then $y_1 + y_2 \in \mathcal{D}^i(c^i)$. Suppose $y^i \in \mathcal{D}^i(c^i)$ for every agent i and $\sum_i y^i \in X$. We will show that this violates individual optimality and hence cannot happen in equilibrium. Let $\bar{y}^i \in X$ be defined by the requirement that $\bar{y}_1^i$ is the projection of y_1^i onto X_1. Write $y^i = \bar{y}^i + (\delta^i, \varepsilon^i)$, where $\delta^i \in \mathbb{R}$ and ε^i is orthogonal to X_1. Since $1 \in X_1$, $\mathbb{E}\varepsilon^i = 0$. By variance aversion and preference transitivity,

$$\bar{y}^i + \left(\delta^i, 0\right) \in \mathcal{D}^i\left(c^i\right). \tag{3.2.6}$$

Let $y \equiv \sum_i y^i$, $\bar{y} \equiv \sum_i \bar{y}^i$ and $\varepsilon \equiv \sum_i \varepsilon^i$. Since $\varepsilon = y_1 - \bar{y}_1 \in X_1$ and ε is also orthogonal to X_1, $\varepsilon = 0$ and therefore $\sum_i \delta^i 1_{\Omega\times\{0\}} = y - \bar{y} \in X$. Since X is arbitrage-free, $\sum_i \delta^i = 0$. If for some i, $\delta^i < 0$, preference monotonicity and (3.2.6) implies $\bar{y}^i \in \mathcal{D}^i(c^i)$, which violates the optimality of c^i. Therefore, $\sum_i \delta^i = 0$ and $\delta^i \geq 0$ for all i, which implies $\delta^i = 0$ for all i. Equation (3.2.6) reduces to $\bar{y}^i \in \mathcal{D}^i(c^i)$, which again violates the optimality of c^i. ♢

Example 3.2.9. (Representative-agent pricing) The following is an example of representative-agent pricing based on the assumed scale invariance (also known as homotheticity) of preferences. Other variations are explored in the exercises.

We will characterize an equilibrium (X, c) for agents who are specified in terms of a fixed reference agent $(\mathcal{D}^0, e^0)$ with $\text{dom}(\mathcal{D}^0) \equiv \mathcal{L}_{++}$. A key assumption is that $\mathcal{D}^0$ is **scale invariant (SI)**:

$$\mathcal{D}^0(sc) = s\mathcal{D}^0(c) \text{ for all } s \in (0, \infty) \text{ and } c \in \text{dom}\left(\mathcal{D}^0\right),$$

where $s\mathcal{D}^0(c) \equiv \{sx \mid x \in \mathcal{D}^0(c)\}$. For $i = 1, \ldots, I$, agent $(\mathcal{D}^i, e^i)$ is defined in terms of the parameters $(b^i, w^i, x^i) \in \mathcal{L} \times (0, \infty) \times X$ by

$$\mathcal{D}^i(c) = \mathcal{D}^0\left(c - b^i\right) \text{ for all } c \in \text{dom}\left(\mathcal{D}^i\right) \equiv b^i + \mathcal{L}_{++},$$

and

$$e^i \equiv b^i + w^i e^0 + x^i.$$

The consumption plan b^i can be thought of as a subsistence plan, the scalar w^i as the agent's wealth above subsistence measured in multiples of the reference plan e^0, and x^i as an endowed traded cash flow. For example, if $e^0 = 1_{\Omega\times\{0\}}$ and X is arbitrage-free, then the present value of the endowment in excess of subsistence, $e^i - b^i$, is equal to w^i. We do not assume, however, that e^0 is marketed. Note that the aggregate endowment can be written as

$$e \equiv \sum\nolimits_i e^i = b + we^0 + x,$$

where $b \equiv \sum_i b^i$, $w \equiv \sum_i w^i$ and $x \equiv \sum_i x^i$. Let us now define the allocation c resulting from first allocating to all agents their subsistence plans b^i and then allocating the remaining aggregate endowment $e - b$ in proportion to each agent's wealth w^i above subsistence:

$$c^i \equiv b^i + \frac{w^i}{w}(e - b), \quad i = 1, \ldots, I.$$

The **representative agent** $(\mathcal{D}, e)$, whose endowment is the aggregate endowment, is defined by the preference correspondence

$$\mathcal{D}(c) \equiv \mathcal{D}^0(c - b) \text{ for all } c \in \operatorname{dom}(\mathcal{D}) \equiv b + \mathcal{L}_{++}.$$

The claim is that (X, c) is an equilibrium if and only if the aggregate endowment e is optimal for the representative agent given the market X, that is, $X \cap \mathcal{D}(e) = \emptyset$. To verify this claim, one can easily check, using the definition of endowments and the allocation, that the allocation c is X-feasible and clears the market. Moreover, using the preference and allocation definitions and the scale-invariance of $\mathcal{D}^0$, we have

$$\frac{1}{w^i}\mathcal{D}^i\left(c^i\right) = \mathcal{D}^0\left(\frac{c^i - b^i}{w^i}\right) = \mathcal{D}^0\left(\frac{e - b}{w}\right) = \frac{1}{w}\mathcal{D}(e), \tag{3.2.7}$$

and therefore $X \cap \mathcal{D}^i(c^i) = \emptyset$ if and only if $X \cap \mathcal{D}(e) = \emptyset$.

Finally, assuming (X, c) is an equilibrium and $\mathcal{D}(e)$ is convex, we show that (X, c) is an effectively complete market equilibrium. Suppose that $y^i \in \mathcal{D}^i(c^i)$ for all i and $y \equiv \sum_i y^i$. From (3.2.7), we have $\bar{y}^i \equiv (w/w_i)\, y^i \in \mathcal{D}(e)$, and therefore, by the convexity of $\mathcal{D}(e)$, $y = \sum_i (w^i/w)\, \bar{y}^i \in \mathcal{D}(e)$. Since e is an optimal consumption plan for the representative agent, it follows that $y \notin X$. $\diamondsuit$

3.3. Utility functions and optimality

In the remainder of this chapter we introduce more structured formulations based on utility representations of preferences. In this section we review general characterizations of optimality and equilibrium using utility functions, which are further specialized in the following section in a way that will allow us to relate time preferences and risk aversion to equilibrium pricing as well as optimal consumption

and portfolio choice. For simplicity, we assume throughout that every agent's consumption set is $C \equiv \mathcal{L}_{++}$, although the theory clearly applies more generally and some of the exercises assume specifications that require different consumption sets. We define utility functions to be continuous and monotone, which is not standard in the literature, but is convenient for our purposes. The existence of a utility representation is characterized in Section A.1.

DEFINITION 3.3.1. A **utility function** is a continuous function of the form $U : C \to \mathbb{R}$ that is **increasing**: for all $c \in C$, if x is an arbitrage, then $U(c+x) > U(c)$. The function $U : C \to \mathbb{R}$ is a **utility representation** of the preference correspondence $\mathcal{D}$ if $\operatorname{dom}(\mathcal{D}) = C$ and $x \in \mathcal{D}(c)$ is equivalent to $U(c+x) > U(c)$, in which case U is said to **represent** $\mathcal{D}$. Two utility functions are **ordinally equivalent** if they represent the same preference correspondence. A utility function U is **normalized** if $U(U(c)) = U(c)$ for all c. The **compensation function** of $\mathcal{D}$ is the function

$$U(c) \equiv \inf\{s \in \mathbb{R} \mid s - c \in \mathcal{D}(c)\}, \quad c \in \operatorname{dom}(\mathcal{D}).$$

Note that the utility functions U and $\tilde{U}$ on C are ordinally equivalent if and only if $\tilde{U} = f \circ U$ for a (strictly) increasing function f from the range of U onto the range of $\tilde{U}$. If $\mathcal{D}$ admits a utility representation $\tilde{U}$, the compensation function U is the unique normalized utility that is ordinally equivalent to $\tilde{U}$; it can be expressed as $U = \phi^{-1} \circ \tilde{U}$ where $\phi(s) \equiv \tilde{U}(s)$, $s \in (0, \infty)$. While the numerical value $\tilde{U}(c)$ in isolation is meaningless, $U(c)$ represents a per-period payment (in the assumed unit of account) of an annuity that is equally desirable as c.

A preference correspondence $\mathcal{D}$ is **convex** if $\mathcal{D}(c)$ is a convex set for every $c \in \operatorname{dom}(\mathcal{D})$, a condition that can be thought of as preference for consumption smoothing across spots. Concavity of a utility function clearly implies the convexity of the preference correspondence it represents. The converse claim is not generally true—a convex preference correspondence admitting a utility representation may admit no concave utility representation.[2] The following example shows that the converse *is* true for scale-invariant preferences.

EXAMPLE 3.3.2. (Scale-invariant preferences) Suppose the preference correspondence $\mathcal{D}$ is **scale invariant (SI)**:

$$\mathcal{D}(sc) = s\mathcal{D}(c) \quad \text{for all } s \in (0, \infty) \text{ and } c \in \mathcal{L}_{++}.$$

The compensation function U of $\mathcal{D}$ is **homogeneous of degree one**:

$$U(sc) = sU(c) \quad \text{for all } s \in (0, \infty) \text{ and } c \in \mathcal{L}_{++}.$$

[2]For example, $U(x, y) \equiv x + \sqrt{x^2 + y}$ is convex on on $\mathbb{R}^2_{++}$, but there is no strictly increasing function f such that $f \circ U$ is concave. This is shown in by Reny [2013] based on an idea outlined by Aumann [1975].

It follows that a preference correspondence admitting a utility representation is SI if and only if it admits a utility representation that is homogeneous of degree one.[3] If $\mathcal{D}$ is both SI and convex, then U is necessarily concave. Since U is homogeneous of degree one, concavity of U is equivalent to its **super-additivity**:

$$U(x+y) \geq U(x) + U(y) \quad \text{for all } x, y \in \mathcal{L}_{++}. \tag{3.3.1}$$

To show super-additivity given convexity of $\mathcal{D}$, fix any $x, y \in \mathcal{L}_{++}$ and define $\alpha, \beta \in \mathbb{R}_{++}$ so that

$$S \equiv U(x+y) = U(\alpha x) = U(\beta y),$$

that is, $\alpha \equiv S/U(x)$ and $\beta \equiv S/U(y)$. Let $p \equiv \alpha/(\alpha+\beta)$. For every positive integer n, we have

$$S = U(\alpha x) < U\left(\alpha x + \frac{1}{n}\right), \quad S = U(\alpha x) = U(\beta y) < U\left(\beta y + \frac{1}{n}\right),$$

and therefore $S < U((1-p)\alpha x + p\beta y + n^{-1})$, since $\mathcal{D}(\alpha x)$ is convex. Letting n go to infinity, we have

$$S \leq U((1-p)\alpha x + p\beta y) = \frac{S\alpha\beta}{\alpha+\beta}.$$

Therefore, $\alpha^{-1} + \beta^{-1} \leq 1$, which rearranges to (3.3.1). ♢

Let us now fix a reference market X and a preference correspondence $\mathcal{D}$ with utility representation U. We call a consumption plan c **optimal** if it is optimal for $\mathcal{D}$ given X. On the space of all adapted processes, we use the inner product

$$\langle x \mid y \rangle = \mathbb{E} \sum_{t=0}^{T} x_t y_t, \quad x, y \in \mathcal{L}. \tag{3.3.2}$$

In applications, U is commonly assumed to be differentiable, in which case a simple characterization of optimality can be given in terms of the gradient vector $\nabla U(c)$, defined by

$$\langle \nabla U(c) \mid x \rangle = \lim_{\epsilon \downarrow 0} \frac{U(c+\epsilon x) - U(c)}{\epsilon}, \quad x \in \mathcal{L}.$$

Example 3.3.3. (Gradient of additive utility) Suppose the utility takes the additive form $U(c) = \mathbb{E}^Q \sum_t u_t(c_t)$, where each $u_t : (0,\infty) \to \mathbb{R}$ is a differentiable function and Q is a full-support probability that can be thought of as expressing beliefs. Let ξ denote the conditional density process $\xi_t = \mathbb{E}_t dQ/dP$. Lemma 2.4.4 implies $U(c) = \mathbb{E} \sum_t u_t(c_t)\xi_t$ and therefore $\nabla U(c)_t = u_t'(c_t)\xi_t$. For example, Section A.4 shows that if U represents scale-invariant preferences, then u_t must take a power or logarithmic form, which implies that the utility function is necessarily differentiable. ♢

[3] In the literature, preferences that are representable by a homogeneous-of-degree-one utility function are commonly referred to as "homothetic."

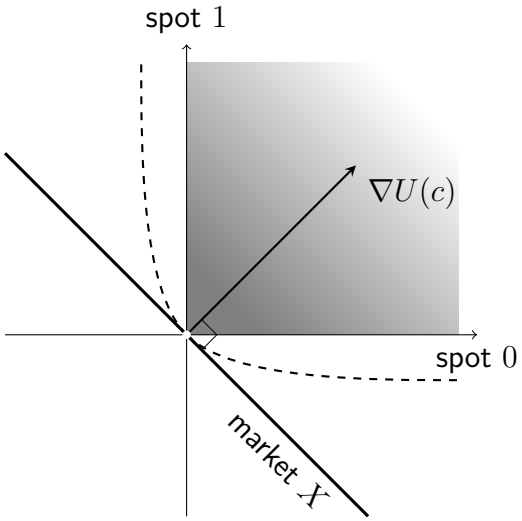


FIGURE 3.3.1. The dotted line is the boundary of the set $\mathcal{D}(c) = \{x \mid c + x \in \mathcal{L}_{++},\ U(c+x) > U(c)\}$, which includes the shaded region of all arbitrage cash flows. The gradient $\nabla U(c)$ is orthogonal to the dotted line and points to the direction of steepest utility gain. Optimality means that X does not intersect $\mathcal{D}(c)$ and is therefore tangent to the dotted line, which in turn implies that $\nabla U(c)$ is orthogonal to the market.

The role of the utility gradient for optimality is illustrated in Figure 3.3.1 (which extends Figure 1.4.1) and is formally shown in the following proposition. Note that the assumed strict positivity of $\nabla U(c)$ is not implied by the monotonicity of U. (Set $u_t(c) = (c-1)^3$ in Example 3.3.3 and compute $\nabla U(1)$.) The reader can show, however, that if U is concave, then the fact that U is monotone does imply the strict positivity of $\nabla U(c)$.

PROPOSITION 3.3.4. *Suppose $c \in C$ and the gradient vector $\nabla U(c)$ exists and is strictly positive. If c is optimal, then $\nabla U(c)$ is a state-price density. Conversely, if U is concave and $\nabla U(c)$ is a state-price density, then c is optimal.*

PROOF. Suppose c is optimal. Given any $x \in X$, we use the fact that C is open to select $\varepsilon > 0$ so that $c + \alpha x \in C$ for all $\alpha \in [0, \varepsilon]$. The function $f(\alpha) = U(c + \alpha x)$, $\alpha \in [0, \varepsilon]$, is maximized at zero and therefore has a nonpositive right derivative at zero: $\langle \nabla U(c) \mid x \rangle \leq 0$. By assumption, $\nabla U(c) \in \mathcal{L}_{++}$ and therefore $\nabla U(c)$ is a state-price density. Conversely, suppose that U is concave and $\pi \equiv \nabla U(c)$ is a state-price density. For all $x \in X$ such that $c + x \in C$, the gradient inequality and the fact that $\langle \pi \mid x \rangle \leq 0$ imply that $U(c+x) \leq U(c) + \langle \pi \mid x \rangle \leq U(c)$. □

In cases where the utility U is known to be concave but differentiability cannot be guaranteed, optimality at $c \in C$ can be characterized in terms of the superdifferential

$$\partial U(c) = \{\pi \in \mathcal{L} \mid c + x \in C \implies U(c+x) \leq U(c) + \langle \pi \mid x \rangle\}.$$

As shown in Section B.5, assuming U is concave, $\partial U(c)$ is nonempty, and $d = \nabla U(c)$ if and only if $\partial U(c) = \{d\}$.

PROPOSITION 3.3.5. *Consider any $c \in C$. If there exists a state-price density $\pi \in \partial U(c)$, then c is optimal. Conversely, if c is optimal and U is concave, then there exists a state-price density $\pi \in \partial U(c)$.*

PROOF. Suppose $\pi \in \partial U(c)$ is a state-price density. For all $x \in X$ such that $c + x \in C$, $\langle \pi \mid x \rangle \leq 0$ and therefore $U(c+x) \leq U(c) + \langle \pi \mid x \rangle \leq U(c)$. This proves the optimality of c. Conversely, suppose c is optimal and U is concave. Consider the convex sets in $\mathbb{R}^{1+K} \times \mathbb{R}$:

$$A \equiv \{(x, \alpha) \mid x \in X,\, U(c) < \alpha\},$$
$$B \equiv \{(y, \beta) \mid c + y \in C,\, \beta \leq U(c+y)\}.$$

If $(x, \alpha) \in A \cap B$, then $U(c) < \alpha \leq U(c+x)$ for some $x \in X$, contradicting the optimality of c. Note also that $(0, U(c))$ is in the closure of both sets. Therefore, by the separating hyperplane Theorem B.5.3, there exists some nonzero $(p, r) \in \mathbb{R}^{1+K} \times \mathbb{R}$ such that

$$(x, \alpha) \in A \implies p \cdot x + r\alpha \leq rU(c), \tag{3.3.3}$$
$$(y, \beta) \in B \implies p \cdot y + r\beta \geq rU(c). \tag{3.3.4}$$

If $r > 0$, condition (3.3.4) is violated by taking β to minus infinity. If $r = 0$, and therefore $p \neq 0$, condition (3.3.4) is violated by taking $y = -\varepsilon p$ for $\varepsilon > 0$ small enough so that $c + y \in C$. Therefore, $r < 0$ and, after rescaling, we can set $r = -1$. Given this normalization, condition (3.3.3) implies that $p \cdot x \leq 0$ for all $x \in X$ and condition (3.3.4) implies that $p \in \partial U(c)$. Since U is increasing, p is necessarily strictly positive. □

As we saw in Proposition 3.1.8, optimality given the market is closely related to the notion of Π-optimality for a present-value function Π. For example, suppose the market is arbitrage-free and complete and therefore the present-value function Π exists and is unique. In this case, if an agent is endowed with a plan $e \in \mathcal{L}_{++}$, then $w \equiv \Pi(e)$ represents the agent's time-zero wealth, and a plan c maximizes utility given the budget constraint $\Pi(c) \leq w$. Let π denote the corresponding state-price density with $\pi_0 = 1$, so that $\Pi(c) = \langle \pi \mid c \rangle$ for all c. The following proposition relates the scaling factor λ that makes $\lambda\pi$ a member of $\partial U(c)$ to the marginal value of the wealth level w.

PROPOSITION 3.3.6. *Suppose* Π *is a linear functional on* $\mathcal{L}$ *with Riesz representation* $\pi \in \mathcal{L}_{++}$ *and*

$$\mathcal{V}(w) \equiv \sup\{U(c) \mid \Pi(c) \leq w,\ c \in \mathcal{L}_{++}\}, \quad w \in \mathbb{R}_{++}. \tag{3.3.5}$$

Then for all $w \in \mathbb{R}_{++}$ *and* $c \in \mathcal{L}_{++}$ *such that* $\Pi(c) = w$*:*

1. *If* $\lambda\pi \in \partial U(c)$ *for some* $\lambda \in (0, \infty)$*, then* c *is* Π*-optimal and* $\lambda \in \partial\mathcal{V}(w)$*.*
2. *Suppose* U *is concave and* c *is* Π*-optimal. Then* $\mathcal{V}$ *is concave,* $\partial\mathcal{V}(w) \neq \emptyset$ *and* $\lambda\pi \in \partial U(c)$ *for all* $\lambda \in \partial\mathcal{V}(w)$*. Assuming further that* $\nabla U(c)$ *exists, then* $\mathcal{V}$ *is differentiable at* w *and*

$$\lambda = \mathcal{V}'(w) = \langle \nabla U(c) \mid x \rangle \text{ for all } x \in \mathcal{L} \text{ such that } \Pi(x) = 1. \tag{3.3.6}$$

PROOF. The result is a corollary of Theorem B.6.2 and Lemma B.6.3. Condition 2 of Theorem B.6.2 in this context reduces to $\lambda\pi \in \partial U(c)$ for some $\lambda \in (0, \infty)$, where the positivity of λ is a consequence of the monotonicity of U. To show condition (3.3.6), note that if U is concave and $\nabla U(c)$ exists for a Π-optimal c, then (by Theorem B.5.5) $\partial U(c) = \{\nabla U(c)\}$ and therefore $\lambda \in \partial\mathcal{V}(w)$ implies $\nabla U(c) = \lambda\pi$. This proves that $\partial\mathcal{V}(w)$ is a singleton and therefore $\lambda = \mathcal{V}'(w)$. Finally, if $\Pi(x) = 1$, then $\langle \nabla U(c) \mid x \rangle = \lambda \langle \pi \mid x \rangle = \lambda$. □

Informally speaking, condition (3.3.6) can be thought of in terms of an optimizing agent who has initial financial wealth w and pays $\Pi(c)$ to consume c. Suppose c is optimal, and therefore $\mathcal{V}(w) = U(c)$ and $\Pi(c) = w$ (since U is increasing, the wealth constraint $\Pi(c) \leq w$ must bind). Condition (3.3.6) means that if the initial wealth w is increased by a small amount δ, the resulting optimal utility value $\mathcal{V}(w + \delta)$ is approximately equal to $U(c + \delta x)$, where x is any unit-cost cash flow. The intuition is that once all but the last penny of one's wealth has been optimally allocated, then what one does with the last penny is of no quantitative significance. This sort of intuition applies more broadly to a class of so-called envelope theorems and is particularly important in interpreting observed choices involving monetary amounts that are trivial in comparison to one's total wealth. Condition (3.3.6) is also the justification for the term **marginal value of wealth** for the Lagrange multiplier λ of the budget constraint $\Pi(c) \leq w$.

EXAMPLE 3.3.7. In the context of Proposition 3.3.6, if U is homogeneous of degree one, then $\mathcal{V}(w) = \mathcal{V}(1)\, w$ and the marginal value of wealth $\lambda = \mathcal{V}'(w)$ does not depend on w. ♢

As in Section 3.1, individual optimality characterizations extend to allocational optimality given the market. To see, how, let us fix a reference preference correspondence profile $(\mathcal{D}^1, \ldots, \mathcal{D}^I)$ and assume that each $\mathcal{D}^i$ has a utility representation U^i. We call an allocation $c \in C^n$ **optimal** given X if it is optimal for $(\mathcal{D}^1, \ldots, \mathcal{D}^I)$ given X. Thus an allocation c is optimal given X if and only if for every other

allocation $c+x \in C^n$, if $U^i(c^i+x^i) > U^i(c^i)$ for all i, then $\sum_i x^i \notin X$. Moreover, by Proposition 3.1.4, c is optimal given X if and only if for all $c+x \in C^n$, if $U^i(c^i+x^i) \geq U^i(c^i)$ for all i and $U^i(c^i+x^i) > U^i(c^i)$ for some i, then $\sum_i x^i \notin X$. Allocational optimality given the trivial market $\{0\}$ is known as **Pareto optimality**. Note that allocational optimality given any market X implies Pareto optimality.

Proposition 3.3.4 extends to allocational optimality as follows.

PROPOSITION 3.3.8. *Suppose c is an allocation such that each gradient $\nabla U^i(c^i)$ exists and is strictly positive. If c is optimal given X, then there exists a state-price density π and some $\mu \in \mathbb{R}^I_{++}$ such that*

$$\pi = \mu_i \nabla U^i\left(c^i\right), \quad i = 1, \ldots, I. \tag{3.3.7}$$

Conversely, assuming every U^i is concave, if π is a state-price density such that (3.3.7) holds for some $\mu \in \mathbb{R}^I_{++}$, then c is optimal given X.

PROOF. Suppose $c \in C^n$ is optimal given X. Then $x = 0$ maximizes $U^i(c^i+x)$ subject to the constraint $U^j(c^j-x) \geq U^j(c^j)$ as x ranges over the set of cash flows such that $c^i+x, c-x \in C$. Applying Theorem B.6.4 results in condition (3.3.7) for some $\mu \in \mathbb{R}^I_{++}$ and $\pi \in \mathcal{L}$. Since allocational optimality implies individual optimality, Proposition 3.3.4 implies that π is a state-price vector. To show the converse, assuming utility concavity, consider any $c^i+x^i \in C$ such that $x \equiv \sum_i x^i \in X$. Multiply the gradient inequality $U^i(c^i+x^i) \leq U^i(c^i) + \langle \nabla U^i(c^i) \mid x^i \rangle$ by μ_i, add up over i, and use condition (3.3.7) and the fact that $\langle \pi \mid x \rangle = 0$ to conclude that $\sum_i \mu_i U^i(c^i+x^i) \leq \sum_i \mu_i U^i(c^i)$. Since $\mu \in \mathbb{R}^I_{++}$, it is not the case that $U^i(c^i+x^i) > U^i(c^i)$ for all i. □

The reader is encouraged to attempt a visualization of the preceding result based on Figure 3.3.1, but with two agents and a common market X. The less obvious case is when X is incomplete, which can be visualized by adding a third orthogonal axis, while preserving X as a line. In this case, $\nabla U^1(c^1)$ and $\nabla U^2(c^2)$ can both be orthogonal to X without satisfying the collinearity condition (3.3.7). Individual optimality implies that both gradients are state-price densities, but not their collinearity. Allocational optimality given X implies Pareto optimality, which in turn implies the gradient collinearity. To visualize the last claim, note that if a cash flow x^i forms an acute angle with $\nabla U^i(c^i)$, then for all sufficiently small $\epsilon > 0$, ϵx^i is strictly desirable for agent i. The non-collinearity of the gradients allows us to choose such x^i that sum up to zero, implying the violation of Pareto optimality.

A useful construct for characterizing effectively complete market equilibria is the so-called **central planner**, defined in terms of the parameter $\mu \in \mathbb{R}^I_{++}$ as the agent whose endowment is the aggregate endowment e, and whose preferences are represented by the utility

function $U^\mu : \mathcal{L}_{++} \to \mathbb{R}$, where

$$U^\mu(c) \equiv \sup\left\{\sum\nolimits_{i=1}^I \mu_i U^i\left(c^i\right) \mid \sum\nolimits_{i=1}^I c^i \leq c,\ c^i \in \mathcal{L}_{++}\right\}. \tag{3.3.8}$$

For scalar μ_i, we use the notation $\mu_i \partial U^i(c^i) \equiv \{\mu_i d \mid d \in \partial U^i(c^i)\}$.

PROPOSITION 3.3.9. *Suppose U^i is concave for all i. For every X-feasible allocation c, (X, c) is an effectively complete market equilibrium if and only if there exists $\mu \in \mathbb{R}^I_{++}$ such that $e = \sum_i c^i$ is optimal for the central planner given X and c solves the maximization problem defining $U^\mu(e)$. In this case, there exists a state-price density π for X such that*

$$\pi \in \partial U^\mu(e) = \bigcap\nolimits_{i=1}^I \mu_i \partial U^i\left(c^i\right),$$

and if the gradient vectors $\nabla U^i(c^i)$ exist, then $\nabla U^\mu(e)$ also exists and

$$\pi = \nabla U^\mu(e) = \mu_i \nabla U^i\left(c^i\right), \quad i = 1, \ldots, I.$$

PROOF. Since every U^i is assumed to be concave, U^μ is also concave and $\partial U^\mu(e) \neq \emptyset$ (by Theorem B.5.4). Apply Theorem B.6.2 to (3.3.8) with $c = e$, $F(\delta) = U^\mu(e + \delta)$ and $\lambda = \pi$. The theorem's second condition in this application can be stated as

$$U^\mu(e) = \sum\nolimits_{i=1}^I \max\left\{\mu_i U^i\left(c^i\right) - \left\langle \pi \mid c^i - e^i \right\rangle \mid c \in \mathcal{L}^I_{++}\right\},\ \pi \in \mathcal{L}_{++},$$

where strict positivity of π follows from the assumption that every U^i is increasing. The i^{th} term of this sum is maximized by c^i if and only if $\pi \in \mu_i \partial U^i(c^i)$. Therefore, for every $\mu \in \mathbb{R}^I_{++}$ and allocation c such that $\sum_i c^i = e$, $U^\mu(e) = \sum_{i=1}^I \mu_i U^i(c^i)$ and $\pi \in \partial U^\mu(e)$ both hold if and only if $\pi \in \bigcap_{i=1}^I \mu_i \partial U^i(c^i)$.

Suppose (X, c) is an effectively complete market equilibrium. By Proposition 3.2.6, there exists a present-value function Π such that (Π, c) is an Arrow-Debreu equilibrium. Given a state price density π representing Π, the Π-optimality of c^i for agent i implies that there exists $\lambda_i > 0$ such that $\lambda_i \pi \in \partial U^i(c^i)$. The argument of the first paragraph with $\mu_i \equiv 1/\lambda_i$ shows that $U^\mu(e) = \sum_i \mu_i U^i(c^i)$ and $\pi \in \partial U^\mu(e)$ and therefore, by Proposition 3.3.5, e is optimal for the central planner given X. Reversing these steps yields the converse. The claim involving gradients is a consequence of Theorem B.5.5(3). □

EXAMPLE 3.3.10. Suppose that every agent's utility takes the form $U^i(c) = \mathbb{E}\sum_t u^i_t(c_t)$, where each $u^i_t : (0, \infty) \to \mathbb{R}$ is a concave differentiable function. For every vector of agent weights $\mu \in \mathbb{R}^I_{++}$, the utility U^μ defined in (3.3.8) is of the same form with $i = \mu$, where

$$u^\mu_t(x) \equiv \left\{\sum\nolimits_{i=1}^I \mu_i u^i_t(x_i) \mid \sum_i x_i \leq x,\ x_i \in (0, \infty)\right\}.$$

If (X, c) is an effectively complete market equilibrium, then there exists $\mu \in \mathbb{R}^I_{++}$ such that $\nabla U^\mu(e)$ is a state-price vector and $\pi_t = u^{\mu\prime}_t(e_t)$

defines a corresponding state-price density, which is a deterministic function of the aggregate endowment at every spot. ♢

While the above example suggests that an additive utility structure has some analytical advantages, the following example highlights a limitation of additive utility in capturing aversion to risk associated with the possible persistence of outcomes.

EXAMPLE 3.3.11. Suppose the random variables $\epsilon_1, \ldots, \epsilon_T$ are independent and identically distributed. For example, think of each ϵ_t as taking the value $+1$ or -1, depending on the outcome of a coin toss. (Mathematically, let $\epsilon_t(\omega) = \omega_t$ in Example 1.1.1 and assign equal probability to every state ω as in Example 2.1.8.) Consider two consumption plans a and b with some common initial value $a_0 = b_0 > 1$ and such that for every time $t > 0$, $a_t = a_0 + \epsilon_1$ and $b_t = b_0 + \epsilon_t$. In the coin-toss interpretation, consumption under a is either permanently increased or permanently decreased depending on the outcome of a single coin toss at time one, while consumption under b is increased or decreased in every period, depending on a new coin toss for every period. While both a and b result in the same expected consumption in every period, there is a clear sense in which a is riskier than b. Yet, every utility of the form $U = \mathbb{E}\sum_t u_t$, for any functions $u_t : (0, \infty) \to \mathbb{R}$, must necessarily assign the same value to both a and b, since $U = \sum_t \mathbb{E}u_t$ and $\mathbb{E}u_t(a_t) = \mathbb{E}u_t(b_t)$ for all t. ♢

3.4. Dynamic consistency and recursive utility

In this section we show that the dynamic consistency of preferences admitting a utility representation leads to a recursive relationship of utilities on the information tree. The resulting utility class includes additive specifications such as expected discounted utility. Example 3.3.11 gave a first indication of the inadequacy of additive utilities. We will further see in this section that additive utilities impose an overly strong relationship between time preferences in the absence of risk and attitudes toward risk. This limitation motivates a broader class of recursive utilities, which allow a partial separation of time preferences and attitudes toward risk. Recursive utility is used in the remainder of this chapter to discuss equilibrium pricing as well as optimal consumption and portfolio choice.

As at the end of Section 3.1, we consider an agent whose preferences at spot (F, t) are represented by a preference correspondence $\mathcal{D}_{F,t}$ whose domain C is common across all spots. Throughout the rest of this text, we restrict attention to the case in which $\mathcal{D}_{F,t}$ has a utility representation and is **forward looking**, meaning that for all $a, b \in C$, if $a = b$ on $F \times \{t, \ldots, T\}$ then $\mathcal{D}_{F,t}(a) = \mathcal{D}_{F,t}(b)$. In other words, spot-$(F, t)$ preferences do not depend on past or unrealized consumption. While this assumption rules out potentially important aspects

of preferences (such as habit formation), it is methodologically appropriate to start with the simpler case and layer complexity on top as needed.[4] A corresponding utility function $U_{F,t} : C \to \mathbb{R}$ must also be **forward looking**: for all consumption plans a, b,

$$a = b \text{ on } F \times \{t, \ldots, T\} \quad \Longrightarrow \quad U_{F,t}(a) = U_{F,t}(b),$$

which allows us to consistently define $U_{F,t}\left(c1_{F\times\{t,\ldots,T\}}\right) \equiv U_{F,t}(c)$ for all $c \in C$, thus extending the domain of $U_{F,t}$ to $C_{F,t} \equiv \left\{c1_{F\times\{t,\ldots,T\}} \mid c \in C\right\}$. We call the function $U_{F,t}$ **increasing on** $C_{F,t}$ if for every $c \in C$ and arbitrage $x \in \mathcal{L}_{F,t}$, $U_{F,t}(c+x) > U_{F,t}(c)$.

Definition 3.4.1. A **spot-(F,t) utility function** is any continuous and forward-looking function $U_{F,t} : C \to \mathbb{R}$ that is increasing on $C_{F,t}$; it is said to **represent** $\mathcal{D}_{F,t}$ if $x \in \mathcal{D}_{F,t}(c)$ is equivalent to $U_{F,t}(c+x) > U_{F,t}(c)$. A **dynamic utility** is a function of the form $U : C \to \mathcal{L}$ such that for every spot (F,t), a spot-(F,t) utility $U_{F,t} : C \to \mathbb{R}$ is well-defined by $U_{F,t}(c) \equiv U(c)(F,t)$, $c \in C$.

We refer to $U(c)$ as the **utility process** of c and write $U_t(c)$ for its time-t value. Since $U_0(c) = U_{\Omega,0}(c)$ for all c, we write U_0 for the time-zero utility function $U_{\Omega,0} : C \to \mathbb{R}$.

Two dynamic utilities U and $\tilde{U}$ are **ordinally equivalent** if for every spot (F,t), the utilities $U_{F,t}$ and $\tilde{U}_{F,t}$ represent the same preference correspondence. The dynamic utility U is **normalized** if $U(\alpha) = \alpha$ for all $\alpha \in (0,\infty)$, and therefore $U_{F,t}(c) = U_{F,t}(U_{F,t}(c))$ for every $c \in C_{F,t}$, meaning that at spot (F,t) an agent whose preferences are represented by $U_{F,t}$ is indifferent between c and a constant annuity with payment $U_{F,t}(c)$ in every period. The discussion of Section 3.3 can be applied to each $U_{F,t}$ to show that every dynamic utility has a unique ordinally equivalent normalized version. A property of a dynamic utility U is **ordinal** if its validity is equivalent to its validity for every dynamic utility that is ordinally equivalent to U, and is therefore effectively a property of the preference correspondences represented by U at every spot. We next introduce the ordinal conditions of dynamic consistency and the irrelevance of current consumption for risk aversion, which together characterize recursive utility.

A dynamic utility U **dynamically consistent** if for every spot spot (F,t) and all $a, b \in C_{F,t}$,

$$U_{F,t}(a) > U_{F,t}(b) \quad \Longrightarrow \quad U_0(a) > U_0(b).$$

The latter condition corresponds to the dynamic consistency assumption $\mathcal{D}_{F,t}(c) \subseteq \mathcal{D}(c) \equiv \mathcal{D}_{\Omega,0}(c)$ first encountered and discussed at the

[4]For example, Schroder and Skiadas [2002] provide a way to mechanically extend the forward looking formulation to incorporate a linear form of habit formation.

end of Section 3.1. The assumed continuity and monotonicity of utilities allows the following characterization:

LEMMA 3.4.2. *U is dynamically consistent if and only if*

$$U_{F,t}(a) \geq U_{F,t}(b) \iff U_0(a) \geq U_0(b), \quad \text{for all } a, b \in C_{F,t}.$$

PROOF. Suppose U is dynamically consistent and $U_{F,t}(a) \geq U_{F,t}(b)$. Then for all $\varepsilon \in (0,\infty)$, $U_{F,t}(a+\varepsilon) > U_{F,t}(b)$ and therefore $U_0(a+\varepsilon) > U_0(b)$. It follows that $U_0(a) \geq U_0(b)$ by continuity. The remaining claims are immediate. □

REMARK 3.4.3. The lemma implies that if U is dynamically consistent, then every ordinal property of $U_{F,t}$ can be equivalently stated as an ordinal property of U_0. In particular, U_0 uniquely defines a normalized version of $U_{F,t}$. ♢

Consider now any nonterminal spot (F,t) and let

$$(F_0, t+1), \ldots, (F_d, t+1) \tag{3.4.1}$$

denote its immediate successor spots. The number d can vary from spot to spot, although in typical applications it is a constant throughout the filtration. Assuming U is normalized, a consequence of dynamic consistency is that the spot-(F,t) utility remains the same if for each i we replace the restriction of c on the subtree rooted on $(F_i, t+1)$ with an annuity whose payments are all equal to the (normalized) utility value $U_{F_i,t+1}(c)$. More formally, we have

LEMMA 3.4.4. *Suppose the dynamic utility U is a normalized and dynamically consistent. Given any $c \in C_{F,t}$, let $\bar{c} \in C_{F,t}$ be defined by letting $\bar{c}(F,t) = c(F,t)$ and $\bar{c} = U_{F_i,t+1}(c)$ on $F_i \times \{t+1, \ldots, T\}$. Then $U_{F,t}(c) = U_{F,t}(\bar{c})$.*

PROOF. Let $x_i \equiv (U_{F_i,t+1}(c) - c)\, 1_{F_i \times \{t+1,\ldots,T\}}$. Adding x_i to c replaces c on the subtree rooted at $(F_i, t+1)$ with an annuity whose payment is $U_{F_i,t+1}(c)$. We use induction in k to show that

$$U_{F,t}(c) = U_{F,t}\left(c + \sum\nolimits_{i=0}^{k} x_i\right), \quad k = 0, 1, \ldots, d. \tag{3.4.2}$$

For $k = 0$, the claim is trivially true. For the inductive step, assume $U_{F,t}(c) = U_{F,t}(b)$, where $b = c + \sum_{i=0}^{k-1} x_i$, with $b = c$ for $k = 0$. The utility normalization implies that $U_{F_k,t+1}(c) = U_{F_k,t+1}(c + x_k)$. Since c equals b on the subtree rooted at $(F_k, t+1)$, $U_{F_k,t+1}(b) = U_{F_k,t+1}(b + x_k)$. By dynamic consistency, $U_0(b) = U_0(b + x_k)$, and therefore $U_{F,t}(b) = U_{F,t}(b + x_k)$. The last equation combined with the inductive hypothesis gives $U_{F,t}(c) = U_{F,t}(b + x_k)$, proving (3.4.2). □

The Lemma's conclusion is equivalent to the existence of a function $\Phi_{F,t} : (0,\infty)^{2+d} \to \mathbb{R}$ such that

$$U_{F,t}(c) = \Phi_{F,t}(c(F,t), U_{F_0,t+1}(c), \ldots, U_{F_d,t+1}(c)), \quad c \in C. \tag{3.4.3}$$

The functions $\Phi_{F,t}$, as (F,t) ranges over all nonterminal spots, specify a backward recursion on the information tree, which starts with the terminal values $U_T(c)$ (equal to c_T if U is normalized) and computes $U_{F,t}(c)$ in terms of spot-(F,t) consumption and the already computed end-of-period utility values. Let us call **generalized recursive utility** any dynamic utility that admits such a recursive specification, which is an ordinal property (why?). One can easily check that every generalized recursive utility is dynamically consistent. We have therefore shown:

LEMMA 3.4.5. *A dynamic utility is generalized recursive utility if and only if it is dynamically consistent.*

We call a (non-normalized) dynamic utility $\tilde{U}$ **additive** if it takes the form $\tilde{U}_t(c) = \mathbb{E}_t \sum_{s=t}^{T} u_s(c_s)$, for functions $u_t : (0,\infty) \to \mathbb{R}$ that are increasing and continuous and hence invertible, and an expectation operator $\mathbb{E}$ under some underlying full-support probability. Every additive utility is dynamically consistency and is therefore generalized recursive utility. As pointed out at the end of last section, utility additivity has some analytical advantages but also some significant limitations. Besides the issue illustrated in Example 3.3.11, preferences under uncertainty represented by additive utilities are entirely determined by preferences in the absence of uncertainty. To clarify and show this claim, first note that for every consumption plan that is **deterministic** (that is, each c_t is constant across states), an additive time-zero utility reduces to $\tilde{U}_0(c) = \sum_t u_t(c_t)$. As a corollary of the uniqueness Theorem A.2.4, we have the following conclusion:

PROPOSITION 3.4.6. *Any two additive utilities that are ordinally equivalent when restricted to deterministic plans are necessarily ordinally equivalent over all consumption plans.*

In order to achieve a partial separation of time preferences and attitudes toward risk, we consider a normalized generalized recursive utility (3.4.3) such that for all $x, x', y_0, \ldots, y_d, y \in (0,\infty)$,

$$\Phi_{F,t}(x, y_0, \ldots, y_d) = \Phi_{F,t}(x, y, \ldots, y) \iff \Phi(x', y_0, \ldots, y_d) = \Phi_{F,t}(x', y, \ldots, y). \tag{3.4.4}$$

This is an ordinal property of $U_{F,t}$ (or U_0 given dynamic consistency) since, by Lemma 3.4.4,

$$\Phi_{F,t}(x, y_0, \ldots, y_d) = U_{F,t}\left(x 1_{F\times\{t\}} + \sum\nolimits_i y_i 1_{F_i\times\{t+1,\ldots,T\}}\right). \tag{3.4.5}$$

Given x, the value $\Phi_{F,t}(x, y_0, \ldots, y_d)$ can be thought of as a utility value of a single-period uncertain payoff $(y_0, \ldots, y_d) \in (0,\infty)^{1+d}$, with corresponding certainty equivalent value $y \equiv \nu_{F,t}(y_0, \ldots, y_d)$ defined implicitly by any of the equivalent equations of condition (3.4.4), independently of the value of the value x. From the perspective of spot (F,t), an agent whose preferences are represented by $U_{F,t}$ is indifferent

between the uncertain annuity $\sum_i y_i 1_{F_i \times \{t+1,\ldots,T\}}$ and the certain annuity that pays $\nu_{F,t}(y_0, \ldots, y_d)$ in every period, independently of spot (F,t) consumption. The lower this payment is, the more risk averse the agent. We say that a dynamic utility U satisfies the **irrelevance of current consumption for risk aversion** to mean that condition (3.4.4) holds at every nonterminal spot (F,t), where $\Phi_{F,t}$ is defined by (3.4.5). If we also define

$$f_{F,t}(x,y) \equiv \Phi_{F,t}(x,y,\ldots,y), \quad x,y \in (0,\infty), \tag{3.4.6}$$

then $\Phi_{F,t}(x,y_0,\ldots,y_d) = f_{F,t}(x,\nu_{F,t}(y_0,\ldots,y_d))$ and (3.4.3) becomes

$$U_{F,t}(c) = f_{F,t}(c(F,t), \nu_{F,t}(U_{F_0,t+1}(c),\ldots,U_{F_d,t+1}(c))). \tag{3.4.7}$$

By **recursive utility** we mean a dynamic utility that satisfies a recursion of this form for some functions $f_{F,t} : (0,\infty)^2 \to (0,\infty)$ and $\nu_{F,t} : (0,\infty)^{1+d} \to (0,\infty)$, as (F,t) ranges over all nonterminal spots.

Lemma 3.4.5 together with the above argument lead to the following main conclusion.

PROPOSITION 3.4.7. *A dynamic utility is recursive utility if and only if it is dynamically consistent and satisfies the irrelevance of current consumption for risk aversion.*

Whether a given dynamic utility is recursive is an ordinal property of the utility and is therefore valid if and only if the normalized version of the utility is recursive. Note that the recursive utility defined by (3.4.7) is normalized if and only if the functions $f_{F,t}$ and $\nu_{F,t}$ in recursion (3.4.7) are also **normalized**:

$$\text{for all } s \in (0,\infty),\ f_{F,t}(s,s) = s \text{ and } \nu_{F,t}(s,\ldots,s) = s.$$

Unless otherwise indicated, in the remainder of this chapter we will work with normalized dynamic utilities only. We use the terms **conditional aggregator** and **conditional certainty equivalent (CE)** to mean continuous, increasing and normalized functions of the form $f_{F,t} : (0,\infty)^2 \to (0,\infty)$ and $\nu_{F,t} : (0,\infty)^{1+d} \to (0,\infty)$, respectively.

Our earlier construction of recursion (3.4.7) makes precise the sense in which the conditional aggregator $f_{F,t}$ and conditional CE $\nu_{F,t}$ represent single-period time preferences and risk aversion from the perspective of spot (F,t). Suppose U' is another normalized recursive utility with spot-(F,t) conditional aggregator $f'_{F,t}$ and conditional CE $\nu'_{F,t}$. By identity (3.4.6), $f_{F,t} = f'_{F,t}$ if and only if $U_{F,t}$ and $U'_{F,t}$ are equal when restricted to consumption plans of the form $x1_{F\times\{t\}} + y1_{F\times\{t+1,\ldots,T\}}$ (which is an ordinal property by normalization). In this sense, $f_{F,t}$ represents time preferences over a single period. A complete separation between time preferences and risk attitudes is not generally possible. We can make the more modest claim, however, that given single-period time preferences, the conditional CE represents single-period risk attitudes. We call the conditional CE $\nu'_{F,t}$ **more risk averse** than $\nu_{F,t}$ if

$\nu'_{F,t} \leq \nu_{F,t}$, meaning that $\nu'_{F,t}(z) \leq \nu_{F,t}(z)$ for all $z \in (0,\infty)^{1+d}$. Assuming that $f_{F,t} = f'_{F,t}$, making $\nu'_{F,t}$ more risk averse than $\nu_{F,t}$ means that from the perspective of spot (F,t), the constant annuity that is of equal utility as the contingent annuity $\sum_i y_i 1_{F_i \times \{t+1,\ldots,T\}}$ makes a lower payment $\nu'_{F,t}(y_0,\ldots,y_d)$ under $U'_{F,t}$ compared to the payment $\nu_{F,t}(y_0,\ldots,y_d)$ under $U_{F,t}$. In this sense, risk aversion toward single-period payoffs is higher under $U'_{F,t}$ than $U_{F,t}$.

EXAMPLE 3.4.8. The functional form of $\nu_{F,t}$ can be founded on any static theory of choice under uncertainty, the most common being expected utility theory: $\nu_{F,t}(z) = u^{-1}\left(\sum_i u(z_i) P[F_i \mid F]\right)$ for some probability P and increasing continuous function $u : (0,\infty) \to \mathbb{R}$. Let us temporarily assume that $\nu_{F,t}$ has the above expected utility specification and similarly $\nu'_{F,t}(z) = u'^{-1}\left(\sum_i u'(z_i) P[F_i \mid F]\right)$, with the same probability P. By Theorem A.6.1, $\nu'_{F,t} \leq \nu_{F,t}$ if and only if $u' = \phi \circ u$ for a concave function $\phi : u(0,\infty) \to \mathbb{R}$. Section A.6 discusses various conditions that express the idea that $\nu_{F,t}$ is risk averse in an absolute sense, all of which are equivalent to the concavity of u. ♢

Recall that L_t denotes the set of all $\mathcal{F}_t$-measurable random variables. Let L_t^{++} denote the set of strictly positive elements of L_t. Just as it is convenient to represent the conditional expectations $\mathbb{E}[\cdot \mid F]$ for every spot (F,t) by an operator $\mathbb{E}_t$, it will be convenient to represent conditional CEs as operators from L_{t+1}^{++} to L_t^{++} by letting

$$v_t(z)(\omega) \equiv v_{F,t}(z) \equiv \nu_{F,t}(z(F_0),\ldots z(F_d)),\ \omega \in F,\ z \in L_{t+1}^{++}, \tag{3.4.8}$$

where $z(F_i)$ denotes the value of the random variable z on the event F_i. For instance, for the expected-utility conditional CE specification of Example 3.4.8, $v_t = u^{-1}\mathbb{E}_t u$, meaning $v_t(z) = u^{-1}(\mathbb{E}_t u(z))$ for all $z \in L_{t+1}^{++}$. For conditional aggregators, we write

$$f_{F,t}(x,y) \equiv f(\omega,t,x,y) \equiv f_t(\omega,x,y) \quad \omega \in F,\ x,y \in (0,\infty). \tag{3.4.9}$$

As with random variables, the state variable ω is typically elided. These conventions allow us to write recursion (3.4.7) in the more succinct form

$$U_t(c) = f_t(c_t, v_t(U_{t+1}(c))), \quad U_T(c) = c_T. \tag{3.4.10}$$

EXAMPLE 3.4.9. Consider the additive utility $\tilde{U}_t = \mathbb{E}_t \sum_{s=t}^T u_s$. The corresponding normalized version is $U_t(c) \equiv \phi_t^{-1} \circ \tilde{U}_t(c)$, where $\phi_t(z) \equiv \sum_{s=t}^T u_s(z)$, $z \in (0,\infty)$. The utility U recursive, as can be seen starting with the recursion $\tilde{U}_t(c) = u_t(c_t) + \mathbb{E}_t \tilde{U}_{t+1}(c)$ and ending with a recursion of the form (3.4.10), where both f and v are defined in terms of the functions u_t, thus tying time preferences to risk aversion as discussed earlier. Rather than go over the general case, the point is made more clearly in the special case where, for some $\beta \in (0,1)$ and

$u : (0, \infty) \to \mathbb{R}$,

$$\tilde{U}_t(c) = \mathbb{E}_t \left[\sum_{s=t}^{T-1} \beta^{s-t} u(c_s) + \frac{\beta^{T-t}}{1-\beta} u(c_T) \right]. \tag{3.4.11}$$

To motivate the utility weight for terminal consumption, define the hypothetical infinite-horizon plan $\bar{c}$ by letting $\bar{c}_t = c_t$ for $t < T$ and $\bar{c}_t = c_T$ for all $t \geq T$. Then $\tilde{U}_t(c) = \mathbb{E}_t \sum_{s=t}^{\infty} \beta^{s-t} u(\bar{c}_s)$. The normalized version U of $\tilde{U}$ satisfies recursion (3.4.10) with

$$f_t(\omega, x, y) \equiv u^{-1}((1-\beta) u(x) + \beta u(y)), \quad v_t \equiv u^{-1} \mathbb{E}_t u. \tag{3.4.12}$$

Since preferences over deterministic consumption plans determine f, they also determine u and hence v, which captures attitudes toward risk. This illustrates the more general conclusion of Proposition 3.4.6. By relaxing the additivity assumption, we can use a different function u in specifying f and v, thus freeing the conditional CE specification from any assumptions on preferences in the absence of uncertainty. ♢

Motivated by the streamlined notation of recursion (3.4.10), we henceforth adopt the following terminology.

An **aggregator** f is a mapping that assigns to every state ω and non-terminal time t a normalized, increasing and continuous function

$$f(\omega, t, \cdot) : (0, \infty)^2 \to (0, \infty),$$

such that the process $(\omega, t) \mapsto f(\omega, t, x, y)$ is predictable, for all (x, y). We say that f is **state** (resp. **time**) **independent** if $f(\omega, t, \cdot)$ does not vary with the state ω (resp. time t). For instance, the aggregator of Example 3.4.9 is state and time independent. For every spot (F, t), the aggregator f **implies** a conditional aggregator $f_{F,t}$ defined in (3.4.9).

A **certainty equivalent (CE)** v is a mapping that assigns to each nonterminal time t a continuous function

$$v_t : L_{t+1}^{++} \to L_t^{++}$$

that is normalized ($v_t(\alpha) = \alpha$ for all $\alpha \in (0, \infty)$) and for all $z \in L_{t+1}^{++}$, the value of $v_t(z)$ at spot (F, t), which we denote by $v_{F,t}(z)$, is an increasing function of the restriction[5] of z on F. Given this condition, we consistently extend the domain of $v_{F,t}$ by letting

$$v_{F,t}(z) \equiv v_{F,t}(z 1_F), \quad z \in L_{t+1}^{++}.$$

For every spot (F, t), the CE v **implies** a conditional CE $\nu_{F,t}$ defined in (3.4.8).

Using this terminology, we restate the definition of recursive utility as it applies to a normalized dynamic utility.

[5]More formally, this means that for all $x, y \in L_{t+1}^{++}$, if $x 1_F = y 1_F$ then $v_t(x) 1_F = v_t(y) 1_F$, and if $x 1_F \geq y 1_F \neq x 1_F$ then $v_t(x) 1_F > v_t(y) 1_F$.

DEFINITION 3.4.10. A normalized dynamic utility U is **recursive utility** if there exist an aggregator f and a CE υ such that for every $c \in C$, the process $U(c)$ solves the backward recursion (3.4.10).

If f is state independent and c is a deterministic consumption plan, then the utility process $U(c)$ is also deterministic and therefore $\upsilon_t(U_{t+1}(c)) = U_{t+1}(c)$. A state-independent aggregator, therefore, determines and is determined by preferences over deterministic consumption plans. For two normalized recursive utilities U and U' sharing the state independent aggregator f and corresponding CEs υ and υ', the reader can show that υ' is **more risk averse** than υ in the sense that $\upsilon'_t \leq \upsilon_t$ for all t, if and only if U'_0 is **more risk averse** than U_0 in the sense that $U'_0 \leq U_0$.

We conclude this section with a recursive utility gradient calculation that is essential in formulating optimality conditions. As in Section 3.3, the gradient is defined relative to the inner product (3.3.2), where $\mathbb{E}$ is the expectation operator relative to some underlying full-support probability that is fixed throughout. We will derive a gradient expression in terms of the CE derivative, which we define using the conditional norm notation $\|h\|_t^2 \equiv \mathbb{E}_t[h^2]$, $h \in L_{t+1}$, as follows: The **derivative** of the CE υ is a mapping κ that assigns to each $t \in \{1, \ldots, T\}$ and $z \in L_{t+1}^{++}$ a random variable $\kappa_{t+1}(z) \in L_{t+1}$ such that for all $z + h \in L_{t+1}^{++}$,

$$\upsilon_t(z+h) = \upsilon_t(z) + \mathbb{E}_t[\kappa_{t+1}(z)\, h] + \varepsilon_t(h)\, \|h\|_t,$$

where $\varepsilon_t(h)$ is small for small h in the sense that for all $\epsilon > 0$, there exists $\delta > 0$ such that $\|h\|_t < \delta$ implies $|\varepsilon_t(h)| < \epsilon$.

A CE is **differentiable** if it has a derivative. For a more concrete expression of the derivative, consider the spot-(F,t) conditional CE $\nu_{F,t}$ implied by υ, and let z_i denote the value $z(F_i)$ that z takes at the immediate successor spot $(F_i, t+1)$ of (F,t). Writing $\kappa_{F_i,t+1}(z)$ for the value of $\kappa_{t+1}(z)$ at $(F_i, t+1)$, we have

$$\kappa_{F_i,t+1}(z)\, P[F_i \mid F] = \frac{\partial \nu_{F,t}(z_0, \ldots, z_d)}{\partial z_i}.$$

A derivative of υ corresponds to the derivative of $\nu_{F,t}$ in the multivariate calculus sense. By a standard calculus result, if the above partial derivatives of $\nu_{F,t}$ exist and are continuous for all nonterminal spots (F,t), then κ is the derivative of υ.

EXAMPLE 3.4.11. Suppose $\upsilon_t = u^{-1}\mathbb{E}_t u$ for some continuously differentiable increasing function $u : (0,\infty) \to \mathbb{R}$. Then the derivative κ of υ exists and is given by

$$\kappa_t(z) \;=\; \frac{u'(z)}{u'(\upsilon_{t-1}(z))}, \quad t = 1, \ldots, T,$$

where u' denotes the derivative of u. $\diamondsuit$

The gradient of a recursive utility is computed in the following proposition, where $\partial f_t/\partial c$ and $\partial f_t/\partial v$ denote the partial derivatives of the time-t aggregator f_t with respect to its consumption and CE arguments, respectively.

PROPOSITION 3.4.12. *Suppose U is a recursive utility with aggregator f and CE υ. Suppose also that $f_t(\omega,\cdot)$ is differentiable for every state ω and time $t<T$, and υ has derivative κ. Given a reference consumption plan c, let the processes λ and $\mathcal{E}$ be defined by*

$$\lambda_t \equiv \frac{\partial f_t}{\partial c}\left(c_t, \upsilon_t\left(U_{t+1}(c)\right)\right), \quad t<T, \quad \lambda_T = 1, \tag{3.4.13}$$

$$\mathcal{E}_0 \equiv 1, \quad \frac{\mathcal{E}_t}{\mathcal{E}_{t-1}} \equiv \frac{\partial f_{t-1}}{\partial v}\left(c_{t-1}, \upsilon_{t-1}\left(U_t(c)\right)\right)\kappa_t\left(U_t(c)\right). \tag{3.4.14}$$

Then for every adapted process x and time t,

$$\lim_{\epsilon\downarrow 0}\frac{U_t(c+\epsilon x)-U_t(c)}{\epsilon} = \mathbb{E}_t\left[\sum\nolimits_{s=t}^{T}\frac{\mathcal{E}_s}{\mathcal{E}_t}\lambda_s x_s\right], \tag{3.4.15}$$

and therefore

$$\nabla U_0(c) = \mathcal{E}\lambda.$$

PROOF. Given $x\in\mathbb{R}^{1+K}$, define $\phi_t(\epsilon)\equiv U_t(c+\epsilon x)$ for all sufficiently small $\epsilon\in\mathbb{R}$. The left-hand-side in (3.4.15) defines the derivative $\phi_t'(0)$. The utility recursion implies

$$\phi_t(\epsilon) = f_t\left(c_t+\epsilon x_t, \upsilon_t\left(\phi_{t+1}(\epsilon)\right)\right).$$

Differentiating at $\epsilon=0$ and using the chain rule, we have

$$\phi_t'(0) = \lambda_t x_t + \frac{\partial f_t}{\partial v}\left(c, \upsilon_t\left(U_{t+1}(c)\right)\right)\mathbb{E}_t\left[\kappa_{t+1}\left(\phi_{t+1}(0)\right)\phi_{t+1}'(0)\right].$$

Letting $V_t\equiv\phi_t'(0)$ and $\delta_t\equiv\lambda_t x_t$, the recursion can be restated as

$$V_t = \delta_t + \frac{1}{\mathcal{E}_t}\mathbb{E}_t\left[\mathcal{E}_{t+1}V_{t+1}\right], \quad V_T = x_T.$$

As discussed in Section 2.3, this recursion expresses the fact that $\mathcal{E}$ (viewed as a state-price density) prices the contract (δ, V), a condition that can be restated as equation (3.4.15). □

REMARK 3.4.13. Suppose that at time zero an agent with an endowment e maximizes the utility U_0 given a complete market with present-value function Π. If c is the agent's optimal consumption plan, then $w\equiv\Pi(e)=\Pi(c)$ is the agent's time-zero wealth. Proposition 3.3.6 expresses the value λ_0 of the above proposition as the marginal value of wealth $\mathcal{V}'(w)$. The same argument can be applied from the perspective of any other spot. For this reason, we will refer to λ as a **marginal value of wealth** process.

3.5. Scale invariant recursive utility

The optimality and equilibrium theory to follow assumes a recursive utility that represents scale invariant preferences in the sense of Examples 3.2.9 and 3.3.2. The result is a simple theory that allows analytical insights while making not entirely unreasonable assumptions on how preferences and portfolio choices scale with wealth. With this motivation, we analyze a scale invariant (SI) normalized recursive utility U with aggregator f and CE v (Definition 3.4.10). By dynamic consistency and Lemma 3.4.4, U is SI (meaning $U_{F,t}$ represents SI preferences for every spot (F,t)) if and only if U_0 is SI. Moreover, for every spot (F,t), since $U_{F,t}$ is assumed normalized, it is SI if and only if it is homogenous of degree one: For every consumption plan c,

$$s \in (0,\infty) \quad \Longrightarrow \quad U_{F,t}(sc) = sU_{F,t}(c).$$

A conditional aggregator or CE can also be viewed as a normalized utility function and is therefore defined to be **SI** if it is homogeneous of degree one. We call an aggregator f (resp. CE v) **SI** if for every non-terminal spot (F,t), the implied conditional aggregator $f_{F,t}$ (resp. conditional CE $\nu_{F,t}$) is SI. Given last section's construction of recursive utility, it is straightforward to verify that *a recursive utility is SI if and only if the corresponding aggregator and conditional CE are both SI.*

Example 3.5.1 (Epstein-Zin-Weil utility). By Theorem A.4.3, the expected utility CE of Example 3.4.8 is SI if and only if

$$v_t = u_\gamma^{-1}\mathbb{E}_t u_\gamma, \quad \text{where} \quad u_\gamma(x) \equiv \frac{x^{1-\gamma}-1}{1-\gamma} \quad (\text{with } u_1 \equiv \log), \tag{3.5.1}$$

for a **coefficient of relative risk aversion (CRRA)** γ. Analogously, assuming an additive utility over deterministic plans (characterized by Theorem A.2.3), the aggregator is SI and state and time independent if and only if it takes the form

$$f_t(\omega, c, v) = u_\delta^{-1}\left((1-\beta)\,u_\delta(c) + \beta u_\delta(v)\right), \tag{3.5.2}$$

for parameters $\beta \in (0,1)$ and $\delta \in \mathbb{R}$, with u_δ defined in (3.5.1). The normalized recursive utility specified in terms of the constant parameters β, δ and γ by the CE (3.5.1) and aggregator (3.5.2) is known as **Epstein-Zin-Weil (EZW) utility.**[6] (All our later results with EZW utility apply with minor changes to parameters that are predictable processes, which can be thought of as the result of applying Theorem A.4.3 separately to each conditional aggregator $f_{F,t}$ and CE $\nu_{F,t}$.) The constant parameters (β,δ) of EZW utility determine and are determined by the utility of deterministic consumption plans. Given (β,δ), increasing γ increases risk aversion. EZW utility reduces to expected discounted utility if and only if $\gamma = \delta$, in which case, the ordinally

[6]Named after the contributions of Epstein and Zin [1989] and Weil [1989, 1990].

equivalent utility $\tilde{U} = (1-\beta)^{-1} u_\delta \circ U$ takes the additive form (3.4.11) of Example 3.4.9 with $u = u_\delta$. Of course, on deterministic consumption plans, $\tilde{U}$ takes the same additive form, with $\mathbb{E}_t$ omitted, no matter what the value of γ is.

In the additive specification (3.4.11), the parameter β determines both the weight on the terminal term and the discounting of every other term. Normalizing the more general additive utility

$$\tilde{U}_t(c) = \mathbb{E}_t\left[\sum_{s=t}^{T-1} b^{s-t} u_\delta(c_s) + \frac{b^{T-t}}{1-\beta} u_\delta(c_T)\right], \tag{3.5.3}$$

for some constant $b > 0$, not necessarily equal to β, yields the EZW form with a time-dependent parameter β. Alternatively, assuming $\delta \neq 1$, utility (3.5.3) becomes EZW utility (with constant parameters) after a change of the unit of account. Let us call the original unit a "bushel" and the new unit a "dollar," and let $\mathbf{u}$ represent the unit conversion process: at time t one bushel equals $\mathbf{u}_t$ dollars. For every consumption plan c in bushels, let $c^{\mathbf{u}} \equiv c\mathbf{u}$ denote the same consumption plan in dollars, and let $\tilde{U}^{\mathbf{u}}(c^{\mathbf{u}}) \equiv \tilde{U}(c)$. Clearly, $\tilde{U}^{\mathbf{u}}$ represents the same preferences over consumption plans in dollars as $\tilde{U}$ does over consumption plans in bushels. For the specific unit conversion choice $\mathbf{u}_t \equiv (1+\alpha)^t$, where α solves $\beta(1+\alpha)^{1-\delta} = b$, $\tilde{U}^{\mathbf{u}}_t(c^{\mathbf{u}})$ is a positive affine transformation of

$$\mathbb{E}_t\left[\sum_{s=t}^{T-1} \beta^{s-t} \frac{(c^{\mathbf{u}}_s)^{1-\delta}}{1-\delta} + \frac{\beta^{T-t}}{1-\beta} \frac{(c^{\mathbf{u}}_T)^{1-\delta}}{1-\delta}\right], \qquad (\delta \neq 1),$$

and therefore the normalized version $U^{\mathbf{u}}$ of $\tilde{U}^{\mathbf{u}}$ is EZW utility with constant parameters β and $\gamma = \delta$. For example, if one is interested in maximizing $\tilde{U}_0(c)$ subject to $\langle \pi \mid c \rangle \leq \pi_0 w$ for some SPD π, all expressed in bushels, one can change the units to dollars and equivalently maximize the EZW utility $U^{\mathbf{u}}(c^{\mathbf{u}})$ subject to $\langle \pi^{\mathbf{u}} \mid c^{\mathbf{u}} \rangle \leq \pi^{\mathbf{u}}_0 w$, where $\pi^{\mathbf{u}} \equiv \mathbf{u}^{-1}\pi$. Note that $\pi^{\mathbf{u}}$ and π differ only in their respective implied short-rate processes $r^{\mathbf{u}}$ and r, which are related by $(1+r^{\mathbf{u}}) = (1+\alpha)(1+r)$. ♢

At the center of the equilibrium and optimality theory to follow is the utility gradient expression $\mathcal{E}\lambda$ of Proposition 3.4.12, with added regularity implied by scale invariance that we now review. It is a general property of the gradient of a homogeneous-of-degree-one functional that it is homogeneous of degree zero and satisfies the so-called Euler equation: The functional's rate of change at a vector x in the direction of x equals the functional's value at x. The proof is a matter of observing simple properties of the relevant difference quotient. Below we state and prove this claim in the language of SI CEs.

LEMMA 3.5.2. *Suppose υ is an SI CE with derivative κ. Then for all processes $s, U \in \mathcal{L}_{++}$,*

$$\kappa_t(s_{t-1}U_t) = \kappa_t(U_t) \quad \text{and} \quad \upsilon_{t-1}(U_t) = \mathbb{E}_{t-1}[\kappa_t(U_t)\, U_t].$$

PROOF. Since the derivative κ of υ is in particular a directional derivative, we have the identity

$$\begin{aligned}\mathbb{E}_{t-1}[\kappa_t(s_{t-1}U_t)\, x_t] &= \lim_{\epsilon \downarrow 0} \frac{\upsilon_{t-1}(s_{t-1}U_t + \epsilon s_{t-1}x_t) - \upsilon_{t-1}(s_{t-1}U_t)}{\epsilon s_{t-1}} \\ &= \mathbb{E}_{t-1}[\kappa_t(U_t)\, x_t],\end{aligned}$$

where the second equality follows from the fact that υ is SI and therefore the term s_{t-1} can be factored out in the numerator. Since x_t can be any $\mathcal{F}_t$-measurable random variable, $\kappa_t(s_{t-1}U_t) = \kappa_t(U_t)$. For $x_t = U_t$ and $s_{t-1} = 1$, scale invariance allows us to factor out $(1+\epsilon)$ in the numerator of the above difference quotient, for all $\epsilon > 0$, and conclude that it must equal $\upsilon_{t-1}(U_t)$. □

The Euler equation for homogeneous functions applied to SI recursive utility gives the following result.

LEMMA 3.5.3. *Suppose U is SI recursive utility and c is a consumption plan. Assume that U satisfies the smoothness assumptions of Proposition 3.4.12, and therefore $\pi \equiv \mathcal{E}\lambda$ is the utility gradient of U_0 at c, where λ and $\mathcal{E}$ are defined by (3.4.13) and (3.4.14), respectively. Then*

$$U_t(c) = \lambda_t W_t, \quad t = 0, \ldots, T, \tag{3.5.4}$$

where

$$W_t \equiv \mathbb{E}_t\left[\sum\nolimits_{s=t}^{T} \frac{\pi_s}{\pi_t} c_s\right]. \tag{3.5.5}$$

PROOF. Let $x = c$ in (3.4.15) and note that the left-hand side equals $U_t(c)$. □

REMARK 3.5.4. The preceding lemma is entirely a statement about the utility function—no market is specified. Suppose, however, that the consumption plan c is optimal given a complete market for an agent maximizing a utility function of the form specified in the lemma. By Proposition 3.1.8, π is also an SPD, and therefore W_t represents the time t market value of the remaining consumption plan $c_t, \ldots, c_T$; that is, the agent's wealth just prior to time-t, which is used to finance all remaining consumption. Equation (3.5.4) in such a context states that the agent's utility at a time-t spot is proportional to the agent's wealth at the given spot, with the constant of proportionality being the marginal value of wealth. The relationship reflects the idea that, because of the homogeneity of the agent's budget constraint and utility function, scaling up the agent's wealth at an optimum scales up proportionately the agent's consumption and utility, while preserving optimality. ♢

For readability, we henceforth write c to denote either a consumption plan or a dummy variable representing a consumption value, and we write v to denote either a CE or a dummy variable representing a conditional CE value. An SI aggregator can be written as

$$f_t(\omega, c, v) = v g_t\left(\omega, \frac{c}{v}\right), \quad c, v \in (0, \infty), \tag{3.5.6}$$

where $g_t(\omega, x) \equiv f_t(\omega, x, 1)$, which leads us to the following terminology and SI aggregator characterization.

DEFINITION 3.5.5. A **proportional aggregator** is a mapping that assigns to each time $t < T$ a function $g_t : \Omega \times (0, \infty) \to (0, \infty)$, where $g_t(\cdot, x)$ is $\mathcal{F}_t$-measurable for all $x \in (0, \infty)$, and for all $\omega \in \Omega$, the function $g_t(\omega, \cdot)$ is continuous and increasing, $g_t(\omega, 1) = 1$, and the mapping $x \mapsto g_t(\omega, x)/x$ is decreasing. A proportional aggregator g **concave** or **differentiable** if $g_t(\omega, \cdot)$ has the respective property for every (ω, t).

LEMMA 3.5.6. *An aggregator f is SI if and only if it takes the form (3.5.6) for some proportional aggregator g, in which case f is concave if and only if g is concave.*

PROOF. The first part of the proposition is a straightforward consequence of the definitions. If f is concave, then clearly so is g. Conversely, suppose f is given by (3.5.6) for a concave proportional aggregator g. We fix the reference pair (ω, t) and abuse notation by writing f and g for the functions $f_t(\omega, \cdot)$ and $g_t(\omega, \cdot)$. Consider any pair of distinct points (c^1, v^1) and (c^2, v^2) in $(0, \infty)^2$ and let (c, v) denote their sum. Concavity of g implies

$$g\left(\frac{c}{v}\right) \geq \frac{v^1}{v^1 + v^2} g\left(\frac{c^1}{v^1}\right) + \frac{v^2}{v^1 + v^2} g\left(\frac{c^2}{v^2}\right).$$

Multiplying through by v shows that $f(c, v) > f(c^1, v^1) + f(c^2, v^2)$, which implies concavity of f, since f is assumed to be homogeneous of degree one. □

The terminology and notation for proportional aggregators is analogous to that for aggregators. Thus we say that g is **state** (resp. **time**) **independent** if $g_t(\omega, \cdot)$ does not vary ω (resp. t). We also omit the state variable in expressions like the utility recursion

$$U_t(c) = v_t(U_{t+1}(c)) g_t\left(\frac{c_t}{v_t(U_{t+1}(c))}\right), \quad t < T, \quad U_T = c_T.$$

We next introduce some useful transformations of a proportional aggregator g that is from now on assumed to be differentiable, with $g_t'(\omega, \cdot)$ denoting the derivative of $g_t(\omega, \cdot)$. The **elasticity** h of g is defined for every time $t < T$ by

$$h_t(x) \equiv \frac{d \log g_t(x)}{d \log x} = \frac{x g_t'(x)}{g_t(x)}, \quad x \in (0, \infty). \tag{3.5.7}$$

Since x, $g_t(x)$ and $g_t'(x)$ are all positive, so is $h_t(x)$. The fact that $g_t(x)/x$ is decreasing is equivalent to

$$h_t(x) \in (0,1), \quad x \in (0,\infty).$$

Note also that since $g_t(1) = 1$, we have $g_t'(1) = h_t(1) \in (0,1)$. The functions g_t' and h_t arise naturally in the following calculations.

LEMMA 3.5.7. *Under the same assumptions as Lemma 3.5.3,*

$$\lambda_t = g_t'(x_t) \quad \text{and} \quad \frac{c_t}{W_t} = h_t(x_t), \quad \text{for all } t < T, \tag{3.5.8}$$

where W_t is defined in (3.5.5) and

$$x_t \equiv \frac{c_t}{\upsilon_t(U_{t+1}(c))}. \tag{3.5.9}$$

PROOF. The expression for λ_t follows from the definition (3.4.13) of λ and identity (3.5.6) defining the proportional aggregator g. To compute the ratio c_t/W_t, we use the identity $U(c) = \lambda W$ from Lemma 3.5.3 together with the definitions of x and h, and the utility recursion :

$$\frac{c_t}{W_t} = \frac{c_t g_t'(x_t)}{U_t(c)} = \frac{c_t g_t'(x_t)}{\upsilon_t(U_{t+1}(c))\, g_t(x_t)} = h_t(x_t).$$

□

Another transformation of g_t that is going to be useful in the theory of SI pricing to follow is defined as

$$q_t(x) \equiv \frac{g_t(x) - x g_t'(x)}{g_t'(x)} \quad x \in (0,\infty). \tag{3.5.10}$$

The following lemma shows that the function q_t arises naturally in computing the growth of the gradient of an SI recursive utility.

LEMMA 3.5.8. *Given the same assumptions as those of Proposition 3.4.12, suppose further that U is SI recursive utility with proportional aggregator g, and let q_t be defined by (3.5.10). Then the gradient $\pi \equiv \nabla U_0(c)$ solves the recursion*

$$\pi_0 = \lambda_0, \quad \frac{\pi_t}{\pi_{t-1}} = q_{t-1}(x_{t-1})\, \kappa_t(U_t(c))\, \lambda_t, \quad t = 1, \ldots, T, \tag{3.5.11}$$

where x_t is defined in (3.5.9), and λ is given in terms of x in (3.5.8) with $\lambda_T = 1$.

PROOF. The claim is a corollary of Proposition 3.4.12 and the observation that

$$q_t(x) = \frac{\partial f_t(c,\upsilon)/\partial \upsilon}{\partial f_t(c,\upsilon)/\partial c}, \quad x \equiv \frac{c}{\upsilon} \in (0,\infty).$$

□

The elasticity of g should not be confused with the elasticity of intertemporal substitution, or EIS for short, a term that is widely used in the literature. In the current context, the **EIS** is defined as the inverse of the elasticity of q:

$$\frac{1}{\mathrm{EIS}_t(x)} \equiv \frac{d\log q_t(x)}{d\log x}, \quad x \in (0,\infty). \tag{3.5.12}$$

The following example shows that a constant EIS corresponds to the SI aggregator form (3.5.2) of Example 3.5.1.

EXAMPLE 3.5.9. Suppose the proportional aggregator g implies a constant EIS and let

$$\delta \equiv \frac{1}{\mathrm{EIS}} \in \mathbb{R}.$$

Define the parameter β by

$$1-\beta \equiv g_t'(1) = h_t(1) \in (0,1).$$

Integrating (3.5.12) and using $h_t(x) = 1/(1+q_t(x)/x)$, we have

$$q_t(x) = \frac{\beta x^{\delta}}{1-\beta} \quad \text{and} \quad h_t(x) = \frac{(1-\beta)x^{1-\delta}}{(1-\beta)x^{1-\delta}+\beta}. \tag{3.5.13}$$

Finally, integrating (3.5.7) and using the fact that $g_t(1) = 1$ results in the proportional aggregator $g_t(x) \equiv f_t(x,1)$, where f is the aggregator (3.5.2) of the EZW specification of Example (3.5.1). ♢

The constant-EIS proportional aggregator is an example of what we will call a regular proportional aggregator, which is defined below and used in simplifying some of the optimality theory to follow.

DEFINITION 3.5.10. The proportional aggregator g is **regular** if it is differentiable and for every $\omega \in \Omega$, the derivative $g_t'(\omega,\cdot) : (0,\infty) \to (0,\infty)$ is decreasing and satisfies

$$\lim_{x\downarrow 0} g_t'(\omega,x) = \infty \quad \text{and} \quad \lim_{x\to\infty} g_t'(\omega,x) = 0. \tag{3.5.14}$$

In the remainder of this section we assume that g is a regular proportional aggregator and we follow the usual notational convention of omitting the implied state variable ω. In this case, the derivative g_t' is invertible and therefore the corresponding right inverse function $\mathcal{I}_t : \Omega \times (0,\infty) \to (0,\infty)$ is well-defined by

$$g_t'(\mathcal{I}_t(\lambda)) = \lambda, \quad \lambda \in (0,\infty). \tag{3.5.15}$$

Analogously to c and v, we abuse notation by using the same symbol for a process, λ in this case, and a dummy variable representing possible values of the process. The definition of $\mathcal{I}_t$ is motivated by the relationship $\lambda_t = g_t'(x_t)$ of Lemma 3.5.7, which is equivalent to $x_t = \mathcal{I}_t(\lambda_t)$.

The regularity assumption on g implies that g_t is a (strictly) concave function. The **convex dual**[7] of g_t is the function $g_t^* : \Omega \times (0,\infty) \to (0,\infty)$ defined for every $\lambda \in (0,\infty)$ (and implied state) by

$$g_t^*(\lambda) \equiv \max_{x\in(0,\infty)} (g_t(x) - \lambda x) = g_t(\mathcal{I}_t(\lambda)) - \lambda \mathcal{I}_t(\lambda). \tag{3.5.16}$$

To gain some geometric understanding of this quantity, draw the graph of g_t on the real plain and a line of slope λ that intersects the graph. Raise the line as high as possible while intersecting the graph and maintaining the slope λ. At the highest such point, the line is tangent to the graph and intersects the vertical axis at $g_t^*(\lambda)$.

Note that after the change of variables

$$\lambda \equiv g_t'\left(\frac{c}{v}\right) \quad \text{and} \quad x \equiv \frac{c}{v},$$

$$\frac{\partial f_t(c,v)}{\partial v} = g_t^*(\lambda) \quad \text{and} \quad q_t(x) = \frac{g_t^*(\lambda)}{\lambda}. \tag{3.5.17}$$

An observation that will be useful later on is that the last equation uniquely determines $q_t(x)$ given λ:

LEMMA 3.5.11. *Assuming the proportional aggregator g is regular, for every time $t < T$ and $q \in (0,\infty)$, there exists a unique $\lambda \in (0,\infty)$ such that $q = g_t^*(\lambda)/\lambda$.*

PROOF. Plot the graph of the concave function g_t and note that as the slope $\lambda = g_t'(x)$ of the tangent line at x decreases, the intercept $g_t^*(\lambda)$ with the vertical axis increases. Therefore, g_t^* is a (strictly) decreasing function. Letting the tangent line's slope go to zero, observe that $g_t^*(0+) = g_t(\infty) > g_t(1) = 1$. Consider now the graph of g_t^* on the plane and, for any given $q \in (0,\infty)$, the line $L = \{(\lambda, \lambda q) : \lambda \in (0,\infty)\}$. The graph of g_t^* is downward sloping, it is above L near the vertical axis and therefore crosses L at exactly one point, which defines the unique $\lambda \in (0,\infty)$ such that $q = g_t^*(\lambda)/\lambda$. □

3.6. Equilibrium with scale invariant recursive utility

Proposition 3.4.12 gives the formula $\mathcal{E}\lambda$ for the gradient of recursive utility at a given reference consumption plan c. If c is optimal for some agent given a market, then $\pi \equiv \mathcal{E}\lambda$ is also an equilibrium state-price density (SPD). Since an SPD prices all traded assets, we have a link between consumption and market prices, which is the basis for so-called consumption-based asset pricing models. A useful way of thinking about equilibrium price restrictions is in terms of the period-t **intertemporal marginal rate of substitution (IMRS)** π_t/π_{t-1}, where π is the utility gradient at a given reference consumption plan c. If c is optimal, the IMRS places recursive restrictions on

[7] In more classical terms, if we define $f_t(x) \equiv -g_t(x)$ and $f_t^*(x^*) \equiv g_t^*(-x^*)$, then f_t^* is the Legendre transform of f_t, also known as the convex conjugate of f_t.

every traded (or synthetic) contract. For example, as we saw in Section 2.3, assuming the contract contract (δ, V) has well-defined returns $R_t = V_t / (V_{t-1} - \delta_{t-1})$, its pricing by π can be stated as

$$\mathbb{E}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}} R_t\right] = 1. \tag{3.6.1}$$

Applied to a traded money-market account with rate process $r \in \mathcal{P}_0$, we have the IMRS mean restriction

$$\mathbb{E}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}}\right] = \frac{1}{1 + r_t}.$$

Given the latter, the Hansen-Jagannathan bound (2.3.9) implies a lower bound on the IMRS variance.

Remark 3.6.1. In interpreting the IMRS, it is important to keep in mind that it is defined relative to the underlying probability P. If π is an SPD under P and $\hat{P}$ is another full-support probability, then the corresponding SPD under $\hat{P}$ is given, through Lemma 2.4.4, as $\hat{\pi}_t = \pi_t \mathbb{E}_t[d\hat{P}/dP]$, where $\mathbb{E}$ denotes expectation under P. $\diamondsuit$

In the remainder of this section we elaborate on equilibrium pricing restrictions under the assumption of SI recursive utility, which is consistent with the representative-agent argument of Example 3.2.9. The theory presented is agnostic toward this interpretation—the consumption plan c can be either an individual investor's plan or a representative agent's plan. Expectations, the SPD property and related quantities whose definition depends on an implicit underlying probability are all relative to a given full-support probability P.

A formula for the equilibrium IMRS for SI recursive utility was established in Lemma 3.5.8. We now show how the IMRS can be determined by taking as input consumption growth only. As a preliminary example, consider an agent maximizing the additive utility of Example 3.4.9 under the assumption of scale invariance, which implies the EZW specification of Example 3.5.1 with $\gamma = \delta$. The utility gradient at a reference consumption plan c can be computed spot by spot, resulting in the IMRS

$$\frac{\pi_t}{\pi_{t-1}} = \beta \left(\frac{c_t}{c_{t-1}}\right)^{-\delta}, \qquad \text{(assuming } \gamma = \delta\text{)}. \tag{3.6.2}$$

The IMRS in this case is entirely determined by the consumption growth process on the information tree. The following proposition shows that for a more general SI recursive utility the mapping from consumption growth to IMRS is less direct, requiring the solution of a backward recursion on the information tree. Although not explicitly stated, the proof shows that the process x corresponds to the consumption-to-CE ratio of equation (3.5.9).

PROPOSITION 3.6.2 (IMRS and consumption growth). *Suppose U is SI recursive utility with differentiable proportional aggregator g and CE υ with derivative κ. Given a consumption plan c, let the process x be defined by the backward recursion*

$$x_{t-1} = \upsilon_{t-1}\left(\frac{g_t(x_t)}{x_t}\frac{c_t}{c_{t-1}}\right)^{-1}, \quad t = 1,\dots,T; \quad x_T = 1. \tag{3.6.3}$$

Then the process $\pi \in \mathcal{L}_{++}$ is the gradient of U_0 at c if and only if $\pi_0 = g_0'(x_0)$ and

$$\frac{\pi_t}{\pi_{t-1}} = q_{t-1}(x_{t-1})\,\kappa_t\left(\frac{g_t(x_t)}{x_t}\frac{c_t}{c_{t-1}}\right)g_t'(x_t), \quad t = 1,\dots,T, \tag{3.6.4}$$

where the functions q_t are defined by (3.5.10).

PROOF. With some abuse of notation, let $U \equiv U(c)$. The definition $x_t \equiv c_t/\upsilon_t(U_{t+1})$ and the recursion defining the utility function imply

$$U_t = \upsilon_t(U_{t+1})\,g_t(x_t) = c_{t-1}\frac{g_t(x_t)}{x_t}\frac{c_t}{c_{t-1}}.$$

Substituting the last expression for U_t in $x_{t-1} = c_{t-1}/\upsilon_{t-1}(U_t)$ and using the homogeneity of υ_{t-1} to cancel out c_{t-1} results in recursion (3.6.3). The same expression for U_t and the fact that κ_t is homogeneous of degree zero (Lemma 3.5.2) result in the identity

$$\kappa_t(U_t) = \kappa_t\left(\frac{g_t(x_t)}{x_t}\frac{c_t}{c_{t-1}}\right). \tag{3.6.5}$$

Lemma 3.5.8 completes the proof. □

EXAMPLE 3.6.3. For the constant-CRRA CE (3.5.1), Example 3.4.11 and equation (3.6.3) imply that the middle factor of the IMRS expression (3.6.4) can be written as

$$\kappa_t\left(\frac{g_t(x_t)}{x_t}\frac{c_t}{c_{t-1}}\right) = \left(\frac{x_t}{x_{t-1}}\right)^{\gamma}\left(g_t(x_t)\frac{c_t}{c_{t-1}}\right)^{-\gamma}.$$

♢

For SI recursive utility, pricing in terms consumption growth is closely related to pricing in terms of the market return. To elaborate, we henceforth fix a reference SI recursive utility U with a regular proportional aggregator (Definition 3.5.10). Given a consumption plan $c \in \mathcal{L}_{++}$, suppose π is the gradient of U_0 at c. The following proposition relates the corresponding IMRS to the quantity

$$M_t \equiv \frac{W_t}{W_{t-1} - c_{t-1}}, \quad \text{where} \quad W_t \equiv \mathbb{E}_t\left[\sum\nolimits_{s=t}^{T}\frac{\pi_s}{\pi_t}c_s\right]. \tag{3.6.6}$$

In the representative-agent pricing context of Example 3.2.9, c is the aggregate endowment and M is the market return process.

PROPOSITION 3.6.4 (IMRS and market returns). *Suppose U is SI recursive utility whose CE υ has derivative κ and whose proportional aggregator g is regular, with convex dual g^* as defined in (3.5.16). Given any $c, \pi \in \mathcal{L}_{++}$, let the process M be defined by (3.6.6). Then the process $\lambda \in \mathcal{L}_{++}$ is uniquely defined as the solution to the following backward recursion, which also defines the process q along the way:*

$$q_{t-1} = \frac{g^*_{t-1}(\lambda_{t-1})}{\lambda_{t-1}} = \frac{1}{\upsilon_{t-1}(\lambda_t M_t)}, \quad \lambda_T = 1. \tag{3.6.7}$$

Finally, π is a gradient of U_0 at c if and only if

$$\frac{\pi_t}{\pi_{t-1}} = q_{t-1}\kappa_t(\lambda_t M_t)\lambda_t, \quad t = 1, \ldots, T, \quad \pi_0 = \lambda_0. \tag{3.6.8}$$

PROOF. Suppose π is the gradient of U_0 at c, and let $U \equiv U(c)$. By Lemma 3.5.3 and the definition of M_t,

$$U_t = \lambda_t W_t = (W_{t-1} - c_{t-1})\lambda_t M_t. \tag{3.6.9}$$

By Lemma 3.5.7, $\lambda_t = g'_t(x_t)$, where x_t denotes the consumption-to-CE ratio (3.5.9). Given this fact, the combination of identities (3.5.10) and (3.5.17) proves the first equality in (3.6.7) with $q_t = q_t(x_t)$. The remainder of (3.6.7) is a consequence of the following string of equalities

$$q_{t-1} = x_{t-1}\frac{1 - h_{t-1}(x_{t-1})}{h_{t-1}(x_{t-1})} = \frac{c_{t-1}}{\upsilon_{t-1}(U_t)}\frac{W_{t-1} - c_{t-1}}{c_{t-1}} = \frac{1}{\upsilon_{t-1}(\lambda_t M_t)}.$$

The first equality follows from the definition of h_t and q_t in (3.5.7) and (3.5.10), the second equality follows from the identity $h_t(x_t) = c_t/W_t$ of Lemma 3.5.7, and the last equality follows by inserting expression (3.6.9) for U_t and simplifying using the homogeneity of υ_{t-1}. That the process λ uniquely solves (3.6.7) is shown in Lemma 3.5.11.

The IMRS expression (3.6.8) follows from Lemma 3.5.8, expression (3.6.9) for U_t, and the fact that κ_t is homogeneous of degree zero (Lemma 3.5.2). Conversely, the IMRS expression uniquely specifies the process π, which must therefore be equal to the (unique) gradient of U_0 at c. □

EXAMPLE 3.6.5. In addition to the assumptions of Proposition 3.6.4, suppose that $\upsilon_t = u_\gamma^{-1}\mathbb{E}_t u_\gamma$, where u_γ is the power or logarithmic function (3.5.1) for some CRRA $\gamma > 0$. The CE derivative calculation of Example 3.4.11 and equation (3.6.7) imply that $\kappa_t(\lambda_t M_t) = (q_{t-1}\lambda_t M_t)^{-\gamma}$. IMRS expression (3.6.8) in this case reduces to

$$\frac{\pi_t}{\pi_{t-1}} = (q_{t-1}\lambda_t)^{1-\gamma} M_t^{-\gamma}, \tag{3.6.10}$$

with q and λ given by (3.6.7). Note that

$$\gamma = 1 \quad \text{implies} \quad \frac{\pi_t}{\pi_{t-1}} = \frac{1}{M_t}. \tag{3.6.11}$$

♢

Consider now the consumption growth to market return ratio

$$\Phi_t \equiv \frac{c_t}{c_{t-1}}\frac{1}{M_t} = \left(\frac{1}{\varrho_{t-1}} - 1\right)\varrho_t, \quad \varrho_t \equiv \frac{c_t}{W_t}, \tag{3.6.12}$$

where the second equation follows from the definition of M in (3.6.6). For a regular proportional aggregator g, the processes x and λ are related by $\lambda_t = g_t'(x_t)$ and $x_t = \mathcal{I}_t(\lambda_t)$. By Lemma 3.5.7, $\varrho_t = h_t(x_t)$ and therefore the ratio Φ is entirely determined by either x or λ. Therefore, given x or λ, consumption growth and market returns carry the same information, which explains why their role is interchangeable in Propositions 3.6.2 and 3.6.4. In both cases, x or λ is determined by a backward recursion. For EZW utility with non-unit EIS, the fact that we can invert the identity $\varrho_t = h_t(x_t)$ means that the IMRS expression can be computed jointly in terms of market returns and consumption growth, without having to solve a backward recursion. The following proposition shows that the resulting IMRS expression[8] is the geometric average of expression (3.6.2) for the additive ($\gamma = \delta$) case, and expression (3.6.11) for the unit CRRA ($\gamma = 1$) case.

PROPOSITION 3.6.6 (EZW pricing with non-unit EIS). *Suppose U is the EZW utility of Example 3.5.1 for some CRRA γ and inverse-EIS parameter $\delta \neq 1$. Fix any consumption plan c and let M_t be defined by (3.6.6). Then $\pi \in \mathcal{L}_{++}$ is the gradient of U_0 at c if and only if*

$$\frac{\pi_t}{\pi_{t-1}} = \left(\beta\left(\frac{c_t}{c_{t-1}}\right)^{-\delta}\right)^{\phi}\left(\frac{1}{M_t}\right)^{1-\phi}, \quad \phi \equiv \frac{1-\gamma}{1-\delta}. \tag{3.6.13}$$

PROOF. Once again, we apply Lemma 3.5.8 and we show that the IMRS expression (3.5.11) reduces to the claimed expression. By Lemma 3.5.7, $\varrho_t \equiv c_t/W_t = h_t(x_t)$, where x_t is the consumption-to-CE ratio. Here h_t is given by equation (3.5.13), and we can therefore invert the last equation to compute x_t as a function of ϱ_t:

$$x_t = \left(\frac{\beta}{1-\beta}\frac{\varrho_t}{1-\varrho_t}\right)^{1/(1-\delta)}.$$

Using the expression for q_t in (3.5.13), we then compute

$$q_t(x_t) = \left(\frac{\beta}{1-\beta}\right)^{1/(1-\delta)}\left(\frac{\varrho_t}{1-\varrho_t}\right)^{\delta/(1-\delta)}. \tag{3.6.14}$$

Similarly, using the fact that $g_t(x) = \left((1-\beta)x^{1-\delta} + \beta\right)^{1/(1-\delta)}$, we find

$$\lambda_t = g_t'(x_t) = (1-\beta)^{1/(1-\delta)}\left(\frac{1}{\varrho_t}\right)^{\delta/(1-\delta)}. \tag{3.6.15}$$

[8]The proposition's proof shows the more general IMRS expression (3.6.17), which applies for any, not necessarily additive, differentiable SI CE. Another generalization is given in Skiadas [2009] for a constant CRRA CE but any regular proportional aggregator g for which h is invertible.

The last two centered equations and equation (3.6.12) together imply

$$q_{t-1}(x_{t-1})\lambda_t = \beta^{1/(1-\delta)}\Phi_t^{-\delta/(1-\delta)}. \tag{3.6.16}$$

Finally, we claim that

$$\kappa_t(U_t(c)) = \kappa_t(\lambda_t M_t) = \kappa_t\left(\Phi_t^{-\delta/(1-\delta)}M_t\right).$$

Since κ_t is homogeneous of degree zero (Lemma 3.5.2), the first equation follows from equation (3.6.9) (just as in the proof of Proposition 3.6.4), and the second equation follows from equation (3.6.16). The last two displays and IMRS expression (3.5.11) result in

$$\pi_0 = \lambda_0 \quad \text{and} \quad \frac{\pi_t}{\pi_{t-1}} = \beta^{1/(1-\delta)}\Phi_t^{-\delta/(1-\delta)}\kappa_t\left(\Phi_t^{-\delta/(1-\delta)}M_t\right). \tag{3.6.17}$$

We will also use the identity

$$\upsilon_{t-1}\left(\Phi_t^{-\delta/(1-\delta)}M_t\right) = \beta^{-1/(1-\delta)}, \tag{3.6.18}$$

To show it, let $q_t = q_t(x_t)$ and recall from Proposition 3.6.4 (and its proof) that $q_{t-1} = 1/\upsilon_{t-1}(\lambda_t M_t)$ and therefore $\upsilon_{t-1}(q_{t-1}\lambda_t M_t) = 1$ (since υ is SI). Substituting expression (3.6.16) for $q_{t-1}\lambda_t$ results in (3.6.18).

All results so far apply for any differentiable SI CE. For a constant CRRA CE, $\kappa_t(z) = (z/\upsilon_{t-1}(z))^{-\gamma}$. Applying this expression with $z = \Phi_t^{-\delta/(1-\delta)}M_t$ in the IMRS expression (3.6.17) and using (3.6.18) the claimed IMRS expression (3.6.13) follows. □

Example 3.6.7 (unit EIS). The key to the proof of the preceding proposition is the invertibility of the identity $\varrho_t = h_t(x_t)$ of Lemma 3.5.7, which does not hold in the unit-EIS case $g_t(x) = x^{1-\beta}$, $\beta \in (0,1)$, where $h_t(x) = 1-\beta$. Equation (3.6.12) in this case gives

$$\frac{c_t}{c_{t-1}} = \beta M_t \qquad \text{(assuming unit EIS).} \qquad \diamondsuit$$

This section's results attribute all IMRS variability to stochastic consumption growth and/or market returns. Another source of IMRS variability can be hard-wired into preferences, for example, through stochastic parameters β, γ, δ, or beliefs (Remark 3.6.1). There are many other possible sources of IMRS variability that do not fit this section's formalism, including agent heterogeneity, non-tradeability of labor income due to moral hazard concerns, collateral constraints and associated leverage dynamics, other institutional constraints, market panics or runs, limited ability to model the future leading to model revisions violating dynamic consistency, and trading patterns driven by narrative and influence dynamics that are hard to explain in a Bayesian framework.

3.7. Optimal consumption and portfolio choice

Consider an agent $(\mathcal{D}, e)$ with access to an arbitrage-free market X. In abstract terms, in this section we discuss, under special assumptions, the problem of finding a trade $x \in X$ such that $e + x$ is optimal for $\mathcal{D}$ given X. We specialize this problem by assuming that preferences have an SI recursive utility representation,

$$e \equiv w1_{\{0\}\times\Omega}, \quad w \in (0, \infty)\,, \tag{3.7.1}$$

and the market X is implemented by $1 + J$ contracts as specified in Section 1.7, where V_t^j is strictly positive for every time $t > 0$ and contract j. Since the market is arbitrage-free, S_{t-1}^j and $R_t^j \equiv V_t^j / S_{t-1}^j$ are also strictly positive. Contract zero is a money-market account (MMA) with rate process r and return process $R^0 \equiv 1 + r$.

Recall that a trading strategy (θ^0, θ) defines in (1.7.6) a corresponding portfolio allocation policy $\psi = \left(\psi^1, \ldots, \psi^J\right)$, and equation (1.7.7) defines the notation R^ψ in terms of ψ and the contract returns R^j. The agent enters period t with financial wealth W_{t-1}, consumes c_{t-1} and invests the remainder according to the allocation ψ_t earning a return R_t^ψ for the period. The wealth process W must therefore satisfy

$$W_0 = w, \quad W_t = (W_{t-1} - c_{t-1})\, R_t^\psi, \quad W_T = c_T. \tag{3.7.2}$$

The pair (c, W) is related to the synthetic contract $(\delta^\theta, V^\theta)$ by

$$S_{t-1}^\theta = W_{t-1} - c_{t-1}, \quad W_t = V_t^\theta, \quad c_t = \delta_t^\theta, \quad t = 1, \ldots, T.$$

The middle equation in (3.7.2) is therefore simply the claim $R^\theta = R^\psi$ of equation (1.7.7). The agent effectively spends $w - c_0$ at time zero to purchase the synthetic contract $(\delta^\theta, V^\theta)$ and subsequently consumes all dividends. Since in an arbitrage-free market every synthetic contract is traded, the contract $(\delta^\theta, V^\theta)$ is priced by every SPD π, which is condition (3.5.5) and the reason we have used the same notation W in both instances.

More parsimoniously, we take as primitive the contract returns and the agent's initial wealth and express the optimal decision using the following terminology and notation.

Definition 3.7.1. A **consumption allocation policy** is a $(0, 1)$-valued adapted process ϱ such that $\varrho_T = 1$. A **portfolio allocation policy** is a process $\psi = \left(\psi^1, \ldots, \psi^J\right) \in \mathcal{P}_0^{1\times J}$. An **allocation policy** is a pair (ϱ, ψ) of a consumption policy and a trading policy. A **wealth process** is a strictly positive adapted process. The allocation policy (ϱ, ψ) **generates** the wealth process W defined recursively by

$$W_0 = w, \quad W_t = W_{t-1}\,(1 - \varrho_{t-1})\, R_t^\psi, \tag{3.7.3}$$

in which case the allocation policy (ϱ, ψ) is said to **finance** the consumption plan $c \equiv \varrho W$. An allocation policy is **optimal** if it finances an optimal consumption plan.

The section's main result follows. It applies to a concave SI recursive utility, where concavity is a consequence of the assumption that both the CE and the proportional aggregator are concave (via Lemma 3.5.6). For simplicity, we assume a differentiable CE v and that the period-t portfolio allocation can be any member of the set

$$\Psi_t \equiv \left\{ \psi_t \in L_{t-1}^{1\times J} \mid R_t^{\psi} \text{ is strictly positive} \right\}. \tag{3.7.4}$$

The result remains true[9] if v is only assumed to be concave and Ψ_t is an arbitrary spot-dependent nonempty convex subset of the set in (3.7.4).

THEOREM 3.7.2. *Assume the reference agent has endowment* (3.7.1) *and preferences represented by an SI recursive utility U with a regular proportional aggregator g and a concave differentiable CE v. Let h, $\mathcal{I}$ and g^* be defined in* (3.5.7), (3.5.15) *and* (3.5.16), *respectively. An allocation policy (ϱ, ψ) is optimal if and only if it can be computed by the following recursive procedure:*

(1) *(Initialization) Set $\lambda_T = \varrho_T = 1$ and $t = T$.*
(2) *(Recursive step) Given λ_t, select $\psi_t \in \Psi_t$ such that*

$$v_{t-1}(\lambda_t R_t^{\psi}) = \max_{p\in\Psi_t} v_{t-1}\left(\lambda_t R_t^{p}\right), \tag{3.7.5}$$

let $\lambda_{t-1} \in L_{t-1}^{++}$ be the unique solution to

$$\frac{\lambda_{t-1}}{g_{t-1}^*\left(\lambda_{t-1}\right)} = v_{t-1}(\lambda_t R_t^{\psi}), \tag{3.7.6}$$

and set $\varrho_{t-1} = h_{t-1}\left(\mathcal{I}_{t-1}\left(\lambda_{t-1}\right)\right)$.
(3) *(Loop) While $t > 1$, decrease t by one and repeat (2).*

Assuming (ϱ, ψ) is optimal and finances c, the process λ generated by this algorithm is the marginal price of wealth process defined by (3.4.13), *and $U(c) = \lambda W$, where W is the wealth process generated by (ϱ, ψ).*

PROOF. Let us first observe that condition (3.7.5) is equivalent to

$$\mathbb{E}_{t-1}\left[\kappa_t(\lambda_t R^{\psi})\lambda_t\left(R_t^j - R_t^0\right)\right] = 0, \quad j = 1, \ldots, J, \tag{3.7.7}$$

where κ is the derivative of v. The condition sets to zero the partial derivatives of the value being maximized in (3.7.5) with respect to each α_t^i. Since the CE v is assumed to be concave, the resulting optimality condition (3.7.7) is necessary and sufficient for (3.7.5).

We henceforth assume that the allocation policy (ϱ, ψ) generates the wealth process W and finances the consumption plan $c \equiv \varrho W$, which by the budget equation in the form (3.7.2) implies that

$$R_t^{\psi} = \frac{W_t}{W_{t-1} - c_{t-1}}. \tag{3.7.8}$$

[9]For details and a proof of this claim see Skiadas [2013a].

Necessity: Suppose c is optimal with corresponding marginal value of wealth process λ. Then $\pi \equiv \nabla U_0(c)$ is an SPD, and therefore

$$\mathbb{E}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}} R_t^j\right] = 1, \quad j = 0, 1, \ldots, J, \tag{3.7.9}$$

which in turn implies

$$\mathbb{E}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}} R_t^\psi\right] = 1, \quad t = 1, \ldots, T. \tag{3.7.10}$$

Substituting expression (3.7.8) for R^ψ into (3.7.10) shows that π prices the contract (W, c), except for the irrelevant technicality that $c_0 \neq 0$, which is easily finessed by redefining the time zero contract value to be $W_0 - c_0$. The result is that W satisfies (3.5.5), and we can therefore apply Proposition 3.6.4 with $M = R^\psi$. Therefore λ uniquely solves recursion (3.7.6) and π_t/π_{t-1} is given by equation (3.6.8), which in conjunction with the state-pricing condition (3.7.9) implies the optimality condition (3.7.7), and hence the optimal-portfolio equation (3.7.5). The expression for ϱ follows from Lemma 3.5.7.

Sufficiency: Conversely, suppose that the processes λ and q are constructed by the Theorem's algorithm and let π be defined by the recursion (3.6.8) but with R^ψ in place of M. By the construction of q and π, and Lemma 3.5.2, we have

$$1 = q_{t-1}\upsilon_{t-1}(\lambda_t R_t^\psi) = q_{t-1}\mathbb{E}_{t-1}\left[\kappa_t(\lambda_t R_t^\psi)\right] = \mathbb{E}_{t-1}\left[\frac{\pi_t}{\pi_{t-1}} R_t^\psi\right].$$

Therefore, equation (2.4.2) is satisfied, which as we saw earlier implies that W satisfies (3.5.5) and Proposition 3.6.4 applies with $M = R^\psi$. It follows that λ is the marginal value of wealth process (since it uniquely solves (3.7.6)), $\pi = \nabla U_0(c)$ and $U(c) = \lambda W$ (by Lemma 3.5.3). Since (3.7.7) is satisfied by construction, so is the first equation in (3.7.9). The latter together with (3.7.7) implies the second equation in (3.7.9). Since π prices every contract implementing the market, it is an SPD, which proves the optimality of c. □

Example 3.7.3 (Unit EIS and optimal consumption policy). Suppose that the proportional aggregator of the unit-EIS form of Example 3.5.9 with $\delta = 1$. In this case, $h_t(x) = 1 - \beta$ for all x, and therefore the optimal consumption-to-wealth ratio is also a constant, $\varrho_t = 1 - \beta$, for every specification of the conditional CE. As we saw in Example 3.5.1, if the conditional CE is an expected utility one with unit CRRA, then the recursive utility is ordinally equivalent to expected discounted logarithmic utility. By moving away from the expected-discounted utility framework, we can vary risk aversion while retaining the simplifying assumption of a constant consumption-to-wealth ratio. ♢

EXAMPLE 3.7.4 (Markovian formulation). Suppose that the contracts implementing the market have the Markovian structure of Section 2.6, which is defined in terms of an underlying martingale basis with stochastically independent increments. In this case, the process λ and the optimal allocation policy can be expressed as functions of the Markov state: $\lambda_t = \lambda(t, Z_t)$, $\varrho_t = \varrho(t, Z_t)$ and $\psi_t = \psi(t, Z_{t-1})$, with the abuse of notation defined in Section 2.6. Analogously to the arbitrage-pricing application of Section 2.6, the significance of the Markovian formulation is that the recursive formula determining ψ_t and λ_{t-1} in terms of λ_t need be evaluated only for every possible value of the Markov state Z_{t-1} rather than every time-$(t-1)$ spot, which can dramatically reduce the problem's computational complexity. ♢

EXAMPLE 3.7.5 (Deterministic marginal value of wealth). Suppose that the short-rate process r is deterministic and period-t excess returns $R_t^j - R_t^0$ are stochastically independent of $\mathcal{F}_{t-1}$, for every $t > 0$. In this case, a backward induction shows that the marginal-value-of-wealth process λ at the optimum is deterministic. As a consequence, the optimal allocation policy (ϱ, ψ) is also deterministic, with the optimal time-t portfolio weights determined as the solution to the myopic problem $\psi_t \in \arg\max\{\upsilon_{t-1}(R_t^p) : p \in \Psi_t\}$. ♢

The algorithm of Theorem 3.7.2 produces a solution provided there is an optimal portfolio solution to problem (3.7.5). We show below that this is indeed the case for an expected-utility SI CE with CRRA $\gamma > 0$. Problem (3.7.5) in this case can be equivalently expressed as

$$\psi_t \in \arg\max_{p \in \Psi_t} \mathbb{E}_{t-1}^Q \left[\frac{(R_t^p)^{1-\gamma} - 1}{1-\gamma} \right], \tag{3.7.11}$$

provided we define the probability Q to have the conditional density process ξ determined by the recursion

$$\frac{\xi_t}{\xi_{t-1}} = \frac{\lambda_t^{1-\gamma}}{\mathbb{E}_{t-1}[\lambda_t^{1-\gamma}]}, \quad \xi_0 = 1. \tag{3.7.12}$$

The reason for this is the change-of-measure formula of Lemma 2.4.10.

EXAMPLE 3.7.6 (Unit CRRA and optimal portfolio choice). The optimal portfolio problem with unit CRRA ($\gamma = 1$) reduces to

$$\psi_t \in \arg\max_{p \in \Psi_t} \mathbb{E}_{t-1} \log(R_t^p).$$

In this case, the optimal portfolio weights are the same as for a myopic agent who maximizes conditional expected logarithmic utility over a single period. In contrast to Example 3.7.5, the myopic portfolio rule is optimal even if the marginal-value-of-wealth process is stochastic. Referring back to Example 3.7.3, recall that a unit EIS implies a myopic consumption policy. The intersection of that example and the present

one is the case of expected discounted logarithm utility, in which case the entire optimal policy (ϱ, ψ) is myopic. $\diamondsuit$

PROPOSITION 3.7.7. *For an SI CE with positive CRRA γ, the optimal portfolio problem has a solution and therefore the algorithm of Theorem 2.4.10 produces a solution.*

PROOF. The argument applies over a single period conditionally on each time-$(t-1)$ spot. We therefore assume, without loss of generality, that $T = 1$ and the underlying probability P coincides with the probability Q specified above. We omit time subscripts and let $u(x) \equiv (x^{1-\gamma} - 1)/(1-\gamma)$. Suppose first that $\gamma \geq 1$. Given any SPD $(1, \pi)$, define the compact set $A \equiv \{z \in L^{++} \mid \mathbb{E}u(z) \geq u(1+r),\ \mathbb{E}[\pi z] = 1\}$ and the closed set $B \equiv \left\{R^p \mid p \in \mathbb{R}^{1\times J}\right\}$. Since $1+r \in B$ and $\mathbb{E}[\pi z] = 1$ for all $z \in B$, the optimal portfolio problem is equivalent to maximizing the continuous function $\mathbb{E}u$ over the compact set $A \cap B$. Existence of a maximum follows by Proposition B.2.6. If $\gamma \in (0,1)$, A as defined above is not compact. We instead define A to be the set of every $[0,\infty)$-valued random variable z such that $\mathbb{E}[\pi z] = 1$. Arguing as above, we conclude there is an optimal portfolio choice, provided we verify that the optimizing value of z is strictly positive. This follows from the fact that the marginal value $z^{-\gamma}$ becomes infinite at zero and therefore a sufficiently small deviation away from zero is always improving. $\square$

3.8. Recursive utility and optimality in continuous time

We conclude with an introduction to recursive utility and optimal consumption/portfolio choice in the Brownian setting of Section 2.8. Technicalities aside, the solution is simpler than last section's discrete counterpart thanks to the quadratic approximations encoded in the Ito calculus. We omit several mathematical details in order to convey the essential argument without requiring knowledge of probability theory beyond what has already been discussed in this text. As in Section 2.8, the underlying filtration is generated by a single SBM B over the time horizon $[0,T]$, defined on some underlying state space (say, the set of continuous paths), with some probability (specifying the likelihood of every path) with expectation operator $\mathbb{E}$. The extension to a filtration generated by multiple independent SBMs is straightforward and mainly a matter of introducing appropriate vector notation.

A **consumption plan** is a strictly positive adapted process c (satisfying a suitable integrability restriction for utility processes and related quantities to be well defined). As in the discrete case, we differentiate between consumption over a single period, which in the current context is $c_t dt$ over $[t, t+dt]$, and terminal lump-sum consumption c_T. State price densities and utility gradients are defined in terms of the inner

product

$$\langle x \mid y \rangle = \mathbb{E}\left[\int_0^T x_t y_t dt + x_T y_T\right].$$

To formulate recursive utility in continuous time, we fix a reference consumption plan c and, abusing notation, we let $U \equiv U(c)$ denote the corresponding normalized utility process with Ito decomposition

$$dU_t = \alpha_t^U dt + \beta_t^U dB_t, \quad U_T = c_T.$$

As we saw in Sections 2.5 and 2.8, expressing α_t^U as a function of $\left(U_t, \beta_t^U\right)$ can be thought of as a backward in time recursion, which is formally expressed as a backward stochastic differential equation (BSDE) to be solved jointly in $\left(U_t, \beta_t^U\right)$. This is the way we can specify recursive utility directly in continuous time.

The heuristic argument that follows can be applied to a broader class of recursive utilities, but for concreteness we assume the expected-utility CE $v_t = u^{-1}\mathbb{E}_t u$ for some continuous and increasing function $u : (0, \infty) \to \mathbb{R}$, with Arrow-Pratt coefficient of absolute risk aversion $A \equiv -u''/u'$. Ito's lemma implies

$$u(U_{t+dt}) = u(U_t) + u'(U_t)\left(\alpha_t^U - \frac{A(U_t)}{2}\left(\beta_t^U\right)^2\right)dt + u'(U_t)\, dB_t.$$

Apply the conditional expectation $\mathbb{E}_t$ on both sides, which eliminates the Brownian term, and then apply u^{-1} on both sides, followed by a first-order Taylor expansion on the right-hand side. The result is the **Arrow-Pratt CE approximation**[10]

$$v_t(U_{t+dt}) = U_t + \left(\alpha_t^U - \frac{A(U_t)}{2}\left(\beta_t^U\right)^2\right)dt.$$

We can heuristically rewrite the Arrow-Pratt approximation as

$$v_t(U_{t+dt}) = \mathbb{E}_t U_{t+dt} - \frac{A(U_t)}{2}\mathrm{var}_t\left[U_{t+dt}\right].$$

As we will see, this is the key in concluding that if a myopic optimal portfolio in a scale-invariant formulation is justified, the optimal portfolio has to be mean-variance efficient in the maximum-Sharpe-ratio sense of Section 2.2 over every infinitesimal time interval.

The aggregator of the recursive utility can also be specified more generally, but again for concreteness, we assume

$$u_\delta(U_t) = \left(1 - e^{-\beta dt}\right)u_\delta(c_t) + e^{-\beta dt}u_\delta\left(v_t(U_{t+dt})\right), \quad U_T = c_T, \tag{3.8.1}$$

where $\beta > 0$ is a constant (not to be confused with β^U). For now, $u_\delta : (0, \infty) \to \mathbb{R}$ can be any increasing and continuously differentiable function. Inserting the Arrow-Pratt CE approximation in this recursion

[10] Named after the contributions of Arrow [1965, 1971] and Pratt [1964].

and using a first-order Taylor expansion results in an expression for α_t^U in terms of $\left(U_t, \beta_t^U\right)$ corresponding to the BSDE

$$dU_t = -\left(f\left(c_t, U_t\right) - \frac{A\left(U_t\right)}{2}\left(\beta_t^U\right)^2\right)dt + \beta_t^U dB_t, \quad U_T = c_T, \tag{3.8.2}$$

where

$$f\left(c, U\right) \equiv \beta \frac{u_\delta\left(c\right) - u_\delta\left(U\right)}{u_\delta'\left(U\right)}.$$

We proceed assuming the BSDE has a unique solution,[11] and the solution is increasing in c. Analogously to the discrete case, this section's entire analysis applies with only cosmetic changes if the parameters f and A are state and time dependent (provided $f\left(\omega, t, c, U\right)$ and $A\left(\omega, t, U\right)$ as functions of (ω, t) are predictable processes).

REMARK 3.8.1. The quadratic term in the drift of BSDE (3.8.2) can be eliminated by passing to the ordinally equivalent utility $V_t = u\left(U_t\right)$. Letting $\varphi \equiv u_\delta \circ u^{-1}$, recursion (3.8.1) can be restated as

$$\varphi\left(V_t\right) = \left(1 - e^{-\beta dt}\right)\varphi\left(u\left(c_t\right)\right) + e^{-\beta dt}\varphi\left(\mathbb{E}_t\left(V_{t+dt}\right)\right), \quad V_T = u\left(c_T\right),$$

This is of the same form as (3.8.1) after substituting $u\left(c_t\right)$ for c_t, φ for u_δ, V for $u\left(U_t\right)$ and finally the risk-neutral CE $\mathbb{E}_t$ for v_t. The argument leading to (3.8.2) in this case leads to the BSDE[12]

$$dV_t = -\phi\left(c_t, V_t\right)dt + \beta_t^V dB_t, \quad V_T = u\left(c_T\right), \tag{3.8.3}$$

where $\phi\left(c, v\right) \equiv \beta\left(\varphi\left(c\right) - \varphi\left(v\right)\right)/\varphi'\left(v\right)$. Alternatively, the equivalence of BSDEs (3.8.2) and (3.8.3) can be shown as an application of Ito's lemma. Assuming sufficient integrability conditions, BSDE (3.8.3) can equivalently be expressed in the form

$$V_t = \mathbb{E}_t\left[\int_t^T \phi\left(c_s, V_s\right)ds + V_T\right],$$

which should be thought of as a fixed-point problem in V. ♢

In the remainder of this section, we concentrate on the special case of an SI U. Since U is normalized, that means U is homogeneous of degree one, and f must also have this property: $f\left(c, U\right) = Ug\left(c/U\right)$. The function g is the continuous-time counterpart of the proportional aggregator. Thanks to Theorem A.4.3, additivity of the aggregator

[11] The existence/uniqueness theory for nonlinear BSDEs begins, independently, with Duffie and Epstein [1992a] (who treated BSDEs of the reduced form of Remark 3.8.1) and Pardoux and Peng [1990] (who allowed for a more general dependence on the volatility). These papers imposed conditions that are violated in this section's main application with EZW utility. For the latter, the relevant BSDE foundations were developed by Schroder and Skiadas [1999] and Xing [2017].

[12] In the continuous-time counterpart of Exercise 6, given in Skiadas [1998], the concavity or convexity of ϕ corresponds to monotonicity of preferences for information, or what in a different setting Kreps and Porteus [1978] termed preferences for the timing for resolution of uncertainty.

coupled with scale invariance means that there exist constants β and δ (the inverse of the EIS) such that

$$g = \beta u_\delta \quad \text{where} \quad u_\delta(x) \equiv \frac{x^{1-\delta} - 1}{1-\delta}, \quad \text{with } u_1 = \log. \tag{3.8.4}$$

Similarly, the assumption of an SI expected-utility CE implies that $u = u_\gamma$ for a constant coefficient of relative risk aversion γ, that is, $A(U)\,U = \gamma$. Preference monotonicity implies $\beta > 0$. We also assume strict concavity of both u_δ and u_γ, and therefore $\gamma, \delta > 0$. Letting $\beta_t^U = \sigma_t^U U_t$, the utility BSDE becomes

$$\frac{dU_t}{U_t} = -\left(g\left(\frac{c_t}{U_t}\right) - \frac{\gamma}{2}\left(\sigma_t^U\right)^2\right)dt + \sigma_t^U dB_t, \quad U_T = c_T. \tag{3.8.5}$$

For concreteness, we proceed under the assumption that g is defined by (3.8.4), corresponding to the continuous-time version of the EZW utility of Example 3.5.1. Once again, the analysis that follows applies more generally to a (sufficiently regular) increasing and concave g, and both g and γ can be state and time dependent (subject to the usual adaptedness restriction). Assuming that g and γ are deterministic allows us to make the claim that g determines and is determined by the utility of deterministic consumption plans, and given g, increasing γ increases risk aversion.

The following is a continuous-time analog to Example 3.5.1.

EXAMPLE 3.8.2. For any $b \in \mathbb{R}$ and $\beta \in (0, \infty)$, let

$$\tilde{U}_t(c) \equiv \mathbb{E}_t\left[\int_t^T e^{-b(s-t)} u_\delta(c_s)\,ds + \frac{e^{-b(T-t)}}{\beta} u_\delta(c_T)\right]. \tag{3.8.6}$$

One way of applying this section's theory to the dynamic utility $\tilde{U}$ is to allow g to be time dependent, which does not affect any of the arguments that follow. Specifically, suppose the dynamic utility U is defined by BSDE (3.8.5) with

$$g \equiv \beta_t u_\delta \quad \text{where} \quad \frac{1}{\beta_t} \equiv \int_t^T e^{-b(s-t)} ds + \frac{e^{-b(T-t)}}{\beta}.$$

An exercise in Ito's lemma then shows that $\tilde{U}_t = \beta_t^{-1} u_\delta(U_t)$.

Alternatively, suppose $\alpha \in \mathbb{R}$ satisfies $b = \beta - (1-\delta)\,\alpha$ and the dynamic utility $\bar{U}$ is defined by the BSDE

$$\frac{d\bar{U}_t}{\bar{U}_t} = -\left(\alpha + \beta u_\delta\left(\frac{c_t}{\bar{U}_t}\right) - \frac{\gamma}{2}\left(\sigma_t^U\right)^2\right)dt + \sigma_t^U dB_t, \quad \bar{U}_t = c_T. \tag{3.8.7}$$

Another exercise in Ito's lemma shows that

$$\tilde{U}_t = \frac{u_\delta\left(\bar{U}_t\right)}{\beta} - \frac{\alpha}{\beta}\int_t^T e^{-b(s-t)} ds,$$

and therefore $\bar{U}_t$ is ordinally equivalent to the additive utility (3.8.6). Strictly speaking, $\bar{U}$ is of the EZW form if and only if $\alpha = 0$, or equivalently $b = \beta$, but the remainder of this section applies just as well with $g \equiv \alpha + \beta u_\delta$ for any α. Alternatively, the EZW form is recovered after a suitable change of the unit account. Analogously to Example 3.5.1, we change the unit of account from a "bushel" to a "dollar," where at time t one bushel equals $\mathbf{u}_t \equiv e^{\alpha t}$ dollars. A consumption plan c in bushels is consumption plan $c^{\mathbf{u}} \equiv c\mathbf{u}$ in dollars, and $U^{\mathbf{u}}(c^{\mathbf{u}}) \equiv \mathbf{u}\bar{U}(c)$ defines a dynamic utility $U^{\mathbf{u}}$ representing the same preferences over consumption plans in dollars as $\bar{U}$ does over consumption plans in bushels. Integration by parts shows that BSDE (3.8.7) is equivalent to BSDE

$$\frac{dU_t^{\mathbf{u}}}{U_t^{\mathbf{u}}} = -\left(\beta u_\delta\left(\frac{c_t^{\mathbf{u}}}{U_t^{\mathbf{u}}}\right) - \frac{\gamma}{2}\left(\sigma_t^U\right)^2\right)dt + \sigma_t^U dB_t, \quad U_t^{\mathbf{u}} = c_T^{\mathbf{u}},$$

which is the normalized EZW form. As we will see in Remark 3.8.4 below, on the market side, changing units from bushels to dollars corresponds to adjusting the short rate process from r to $r + \alpha$. ♢

Consider now an agent[13] whose preferences are represented by the continuous-time SI recursive utility just defined, who has some initial financial wealth $w > 0$ and no other income, and who has access to the financial market of Section 2.8, but with r, a, y, σ, and therefore μ, representing general predictable processes, rather than constants (subject to omitted technical integrability restrictions). Allowing more general market parameters will be useful in our discussion of how optimal portfolios relate to mean-variance efficiency. The agent can use w to purchase an initial portfolio in the MMA and the stock and can rebalance the account over time, provided the account balance W_t stays positive at every time t. Unlike the trading strategies of Section 2.8, which were assumed to be self-financing, the agent can withdraw cash from the account at a time-t rate c_t for all $t < T$, followed by a terminal lump-sum payment $c_T = W_T$, thus converting the initial wealth $w = W_0$ to a consumption plan c. The agent seeks to do so in a way that maximizes the utility $U_0(c)$. By dynamic consistency, maximization of time-zero utility implies the maximization of utility at every future time. It is therefore sufficient to solve the agent's problem from the perspective of time zero. A useful way of thinking of the agent's consumption and trading strategy is as the time-zero purchase of the

[13]The rest of this section is based on Schroder and Skiadas [2003], which includes trading constraints, more general recursive utilities, as well as multiple assets and sources of risk, all in the context of an SI formulation with possibly incomplete markets but tradeable income. Non-tradeable income is inconsistent with scale invariance. Schroder and Skiadas [2005] give a version of the argument that includes non-tradeable income based on translation invariance (which removes wealth effects).

synthetic contract (c, W), where c is the dividend process generated by the contract and W is the contract's cum-dividend price process.

The reader would have no difficulty extending the self-financing budget equation of Section 2.8 to include this type of contract. Instead, we equivalently describe the budget equation in terms of wealth allocations, analogously to (3.7.3). An **allocation policy** is a pair (ϱ, ψ) of predictable processes, where for every time $t < T$, $\varrho_t \in (0, \infty)$ represents the consumption rate as a proportion of wealth and ψ_t is the proportion of wealth that is allocated to the stock, with the remainder allocated to the MMA. We also assume, as part of an allocation policy definition, that $\varrho_T \equiv 1$ (corresponding to the assumption $c_T = W_T$) and that the following version of the budget equation is a well-formed SDE uniquely determining the strictly positive **wealth process** W **generated** by (ϱ, ψ):

$$W_0 = w, \quad \frac{dW_t}{W_t} = (r_t - \varrho_t)\, dt + \psi_t \left(\frac{dS_t}{S_t} + y_t dt \right). \tag{3.8.8}$$

The resulting consumption plan **financed** by the allocation policy (ϱ, ψ) is $c \equiv \varrho W$. We call a consumption plan **feasible** if it is financed by some allocation policy, and **optimal** if it maximizes U_0 among all feasible consumption plans. We wish to determine an allocation policy that is **optimal** in that it finances an optimal consumption plan.

Given the reference consumption plan c, the relevant market $X(c)$ is the set of all x such that $c + x$ is feasible consumption plan. As in Section 2.8, let the strictly positive Ito process π be defined, for arbitrary $\pi_0 > 0$, by the SDE

$$\frac{d\pi}{\pi} = -r_t dt - \eta_t dB, \quad \text{where} \quad \eta_t \equiv \frac{\mu_t - r_t}{\sigma_t}. \tag{3.8.9}$$

In a variant of Proposition 2.8.1, we will show that π is an SPD for $X(c)$ by analyzing the linear BSDE

$$dW_t = -\left(c_t - r_t W_t - \eta_t \sigma^W\right) dt + \sigma_t^W dB_t, \quad W_T = c_T, \tag{3.8.10}$$

which results from entering $c = \varrho W$ into the budget equation (3.8.8) and using the return dynamics (2.8.1) and the definition of η in (3.8.9). BSDE (3.8.10) expresses the heuristic backward recursion

$$W_t = \mathbb{E}_t \left[c_t + \frac{\pi_{t+dt}}{\pi_t} W_{t+dt} \right], \quad W_T = c_T.$$

In the finite model, such a recursion states that the present-value function represented by π prices the contract (c, W) and W_t is the time-t present value of c. The latter conclusion is not guaranteed in continuous time. As we saw in the last chapter, in continuous time, we have to entertain the possibility that the synthetic contract (c, W) is implemented by a "doubling" strategy (in a generalized sense), where wealth can be created from nothing, or a reverse doubling strategy,

where wealth is guaranteed to be destroyed. The fact that W is required to stay positive precludes doubling strategies, but not reverse doubling strategies. As a consequence W_t should be at least as great as the time-t present value of c. A condition that limits in a suitable sense how big W can get rules out reverse doubling strategies, allowing the conclusion that W_t equals the time-t present value of c.

LEMMA 3.8.3. *Suppose* $\left(W, \sigma^W\right)$ *solves BSDE* (3.8.10). *Then*

$$W_t \geq \frac{1}{\pi_t}\mathbb{E}_t\left[\int_t^T \pi_s c_s ds + \pi_T c_T\right], \quad t \in [0,T]. \tag{3.8.11}$$

Moreover, if $\mathbb{E}\left[\sup_t \pi_t W_t\right] < \infty$, *the inequality holds as an equality.*

PROOF. Given the dynamics (2.8.6) for π and (3.8.10) for W, integration by parts implies that $d\left(\pi_t W_t\right) = -\pi_t c_t dt + dM_t$, where M is a local martingale. Let τ_n be a sequence of stopping times such that $\tau_n \uparrow T$ with probability one and M is a martingale up to time τ_n, and therefore $M_t = \mathbb{E}_t M_{\tau_n}$ on $\{\tau_n \geq t\}$, which in turn implies

$$\pi_t W_t = \mathbb{E}_t\left[\int_0^{\tau_n} \pi_s c_s ds + \pi_{\tau_n} W_{\tau_n}\right] \quad \text{on } \{\tau_n \geq t\}.$$

As $n \to \infty$, the term $\mathbb{E}_t\left[\int_t^{\tau_n} \pi_s c_s ds\right]$ converges to $\mathbb{E}_t\left[\int_t^T \pi_s c_s ds\right]$ by the monotone convergence theorem (for conditional expectations). We also know that $\pi_{\tau_n} W_{\tau_n}$ converges to $\pi_T W_T = \pi_T c_T$. Inequality (3.8.11) follows by Fatou's lemma (for conditional expectations). If we further know that $\mathbb{E}\left[\sup_t \pi_t W_t\right] < \infty$, then $\lim_n \mathbb{E}_t\left[\pi_{\tau_n} W_{\tau_n}\right] = \mathbb{E}_t\left[\pi_T W_T\right]$ and (3.8.11) becomes an equality thanks to Lebesgue's dominated convergence theorem (for conditional expectations), another fundamental result in the theory of integration.[14] □

Preference monotonicity implies that if the consumption plan c is a candidate to be optimal, it should not be financed by a reverse doubling strategy. A sufficient condition for this to be true is last lemma's integrability condition, which we would have to verify in a mathematically complete formulation. The integrability condition allows us to confirm the SPD property of π as we set out to do.

LEMMA 3.8.4. *Suppose an allocation policy generates the wealth process* W *and finances the consumption plan* c. *If* $\mathbb{E}\left[\sup_t \pi_t W_t\right] < \infty$, *then* π *is an SPD relative to* $X(c)$, *that is,* $\langle \pi \mid x \rangle \leq 0$ *whenever* $c + x$ *is a feasible consumption plan.*

PROOF. Suppose $c+x$ is a feasible consumption. By the last lemma, we have $\pi_0 w \geq \langle \pi \mid c + x \rangle$ and $\pi_0 w = \langle \pi \mid c \rangle$. Subtracting the latter from the former, we find $0 \geq \langle \pi \mid x \rangle$. □

[14] For all the referenced conditional expectation limit results, see, for example, Chapter 10 of Dudley [2002].

Given the last lemma, our optimality verification argument will be exactly as in Proposition 3.3.4, based on concavity and the SPD property of the utility gradient.

REMARK 3.8.5 (Change of unit of account). As in Example 3.8.2, consider a change of the unit of account where consumption plan c in "bushels" is consumption plan $c^{\mathbf{u}} = c\mathbf{u}$ in "dollars." If π is an SPD in bushels, then $\mathbf{u}\pi$ is the same SPD in dollars. Maximizing $U_0(c)$ subject to $\langle \pi \mid c \rangle \leq w\pi_0$ (as we are effectively doing in the solution method that follows) is equivalent to maximizing $U_0^{\mathbf{u}}(c^{\mathbf{u}})$ subject to $\langle \mathbf{u}\pi \mid c^{\mathbf{u}} \rangle \leq w(\mathbf{u}\pi)_0$, where $U^{\mathbf{u}}(c^{\mathbf{u}}) \equiv U(c)$. In Example 3.8.2, $\mathbf{u}_t \equiv e^{\alpha t}$ and the only difference between π and $\pi^{\mathbf{u}}$ is in their respective implied short rate processes r and $r+\alpha$. To see that apply integration by parts to $\pi\mathbf{u}$ and compare the resulting Ito expansion to (3.8.9). $\diamondsuit$

Technical aspects aside, the utility gradient calculation is analogous to Proposition 3.3.4 for the discrete case. Given any feasible direction x, the idea is to use the utility BSDEs to write the Ito expansion of $U(c+\epsilon x) - U(c)$, and then linearize the Ito coefficients, assuming small ϵ, resulting in a linear BSDE of the form we used earlier to price the contract (c, W). The following remark, which can be skipped on a first reading, outlines the general pattern and associated gradient inequality.[15]

REMARK 3.8.6. Omitting technical details, suppose that for a differentiable function F, the process $Y \equiv Y(c)$ solves the BSDE

$$dY_t = -F_t(c_t, Y_t, Z_t)\,dt + Z_t dB_t, \quad Y_T = F_T(c_T). \tag{3.8.12}$$

Fixing c, we use the notation $(Y, Z) \equiv (Y(c), Z(c))$ and

$$\lambda_t = \frac{\partial F_t(c_t, Y_t, Z_t)}{\partial c}, \; F_Y(t) \equiv \frac{\partial F(c_t, Y_t, Z_t)}{\partial Y}, \; F_Z(t) \equiv \frac{\partial F(c_t, Y_t, Z_t)}{\partial Z}.$$

Then (omitting technical requirements),

$$\lim_{\epsilon \downarrow 0} \frac{Y_t(c+\epsilon x) - Y_t(c)}{\epsilon} = \mathbb{E}_t\left[\int_t^T \frac{\mathcal{E}_s}{\mathcal{E}_t}\lambda_s x_s ds + \frac{\mathcal{E}_T}{\mathcal{E}_t}\lambda_T x_T\right], \tag{3.8.13}$$

where $\mathcal{E}$ solves the SDE

$$\frac{d\mathcal{E}_t}{\mathcal{E}_t} = F_Y(t)\,dt + F_Z(t)\,dB_t, \quad \mathcal{E}_0 = 1.$$

[15] The recursive utility gradient calculation and its use in this chapter originated in my doctoral thesis Skiadas [1992]. At that time, my advisor Darrell Duffie had just pioneered continuous-time recursive utility with Larry Epstein in Duffie and Epstein [1992a], and asset pricing applications using dynamic programming Markovian methods in Duffie and Epstein [1992b]. My idea was to characterize optimality in a static way in terms of gradients without a Markovian structure. These initial results appeared in Duffie and Skiadas [1994]. I continued this work at Northwestern with Mark Schroder, who was a doctoral student at the time, leading to the theory on which this section is based.

Assuming F is concave, we outline a proof of the gradient inequality

$$Y_0(c+x) \leq Y_0(c) + \langle \mathcal{E}\lambda \mid x \rangle. \tag{3.8.14}$$

Let $y \equiv Y(c+x) - Y(c)$ and $z \equiv Z(c+x) - Z(c)$. Omitting time indices, the gradient inequality for F gives

$$0 \leq D \equiv -F(c+x, Y+y, Z+z) + F(c, Y, Z) + \lambda x + F_Y y + F_Z z.$$

Subtracting the BSDE for $Y(c)$ from the BSDE for $Y(c+x)$, we get

$$dy = -(\lambda x - D + F_Y y + F_Z z)\, dt + z dB, \quad y_T = x_T.$$

This is a linear BSDE of the form (3.8.10), with $(y, F_c x - D, F_Y, F_Z)$ corresponding to $(W, c, -r, -\eta)$. Given sufficient regularity, the argument of Lemma 3.8.3 in this context implies that $y_0 = \langle \mathcal{E} \mid \lambda x - D \rangle$. Since both $\mathcal{E}$ and D are positive, $y_0 \leq \langle \mathcal{E} \mid \lambda x \rangle = \langle \mathcal{E}\lambda \mid x \rangle$, which is the claimed gradient inequality. ♢

We continue taking as given the reference consumption plan c and associated utility process $U = U(c)$ determined by BSDE (3.8.5) with g defined in (3.8.4). Although we have used x for a variety of roles, in the remainder of this section it represents the consumption to utility ratio, in terms of which we define the process λ:

$$x_t \equiv \frac{c_t}{U_t} \text{ and } \lambda_t \equiv g'(x_t) = \beta x_t^{-\delta} \text{ for } t < T; \quad \lambda_T = 1. \tag{3.8.15}$$

The convex dual g^* is defined by (3.5.16), and therefore

$$g^*(\lambda_t) = g(x_t) - g'(x_t)\, x_t. \tag{3.8.16}$$

Remark 3.8.6 applies to BSDE (3.8.5) by letting $(Y, Z) = \left(U, U\sigma^U\right)$ and $F(c, Y, Z) \equiv Y g(c/Y) - (\gamma/2) Y (Z/Y)^2$. As a consequence, the gradient of U_0 is $\mathcal{E}\lambda$, where $\mathcal{E}$ solves the SDE

$$\frac{d\mathcal{E}_t}{\mathcal{E}_t} = \left(g^*(\lambda_t) + \frac{1}{2}\gamma \left(\sigma_t^U\right)^2 \right) dt - \gamma \sigma_t^U dB_t, \quad \mathcal{E}_0 = 1.$$

The assumed concavity of g implies the concavity of F, as can be seen by applying Lemma 3.5.6 to each of the additive terms defining F. This leads to the gradient inequality and the following verification argument.

LEMMA 3.8.7. *Suppose the allocation policy (ϱ, ψ) finances the plan c and $\pi \equiv \mathcal{E}\lambda$ is an SPD for the market $X(c)$. Then (ϱ, ψ) is optimal.*

PROOF. Suppose the plan $c + x$ is feasible. The gradient inequality (3.8.14) with $Y = U$ and $\pi \equiv \mathcal{E}\lambda$ gives $U_0(c+x) \leq U_0(c) + \langle \pi \mid x \rangle$. The fact that $x \in X(c)$ implies $\langle \pi \mid x \rangle \leq 0$. The two inequalities together imply that $U_0(c+x) \leq U_0(c)$. □

To apply this lemma, suppose (ϱ, ψ) finances c and $\pi \equiv \mathcal{E}\lambda$ solves SDE (3.8.9) with $\pi_0 = \lambda_0$, and is therefore an SPD for $X(c)$ by Lemma 3.8.4. The Euler equation for homogeneous functions gives

$$U_t = \lim_{\epsilon \downarrow 0} \frac{U_t(c + \epsilon x) - U_t(c)}{\epsilon} = \frac{\lambda_t}{\pi_t} \mathbb{E}_t \left[\int_t^T \pi_s c_s ds + \pi_T c_T \right]. \tag{3.8.17}$$

The first equation follows from the fact U_t is homogeneous of degree one (so ϵ cancels out) and the second equation follows from identity (3.8.13) with $Y = U$ and $\mathcal{E}\lambda = \pi$. Provided $\mathbb{E}[\sup_t \pi_t W_t] < \infty$, Lemma 3.8.3 implies that the right-hand side in (3.8.17) equals $\lambda_t W_t$. Using integration by parts, we therefore have the key identities

$$U = \lambda W \quad \text{and} \quad \frac{dU}{U} = \frac{d\lambda}{\lambda} + \frac{dW}{W} + \frac{d\lambda}{\lambda}\frac{dW}{W}. \tag{3.8.18}$$

The first equation and (3.8.15) allow us to compute the optimal consumption policy as a function of λ_t:

$$\varrho_t \equiv \frac{c_t}{W_t} = \lambda_t \frac{c_t}{U_t} = \lambda_t x_t = \beta^{1/\delta} \lambda_t^{1-1/\delta}, \quad t < T. \tag{3.8.19}$$

As in the discrete case, the determination of the optimal portfolio allocation policy can be performed jointly with a backward recursion that determines λ. To see how, we start with the notation

$$\frac{d\lambda_t}{\lambda_t} \equiv \mu_t^\lambda dt + \sigma_t^\lambda dB_t.$$

The budget equation (3.8.8) and (3.8.18) imply

$$\sigma^U = \sigma^\lambda + \psi\sigma. \tag{3.8.20}$$

On the other hand, integration by parts applied to $\pi = \mathcal{E}\lambda$ gives

$$\frac{d\pi}{\pi} = \left(g^*(\lambda_t) + \frac{1}{2}\gamma \left(\sigma_t^U\right)^2 + \mu_t^\lambda - \gamma \sigma_t^U \sigma^\lambda \right) dt - \left(\gamma \sigma_t^U - \sigma_t^\lambda\right) dB_t.$$

Matching coefficients with the Ito decomposition (3.8.9) for $d\pi/\pi$ and substituting expression (3.8.20) for σ^U, we find

$$-\mu^\lambda = r + g^*(\lambda) + \frac{\gamma}{2}\left((\psi\sigma)^2 - \left(\sigma^\lambda\right)^2 \right), \quad \psi\sigma = \frac{1}{\gamma}\left(\eta + (1-\gamma)\sigma^\lambda\right).$$

These two equations are the counterpart of the joint recursive determination of λ and ψ in the second step of the algorithm of Theorem 3.8.21. The Arrow-Pratt CE approximation leads to the closed form solution for $\psi\sigma$, which can now be substituted into the expression for μ^λ. The result is a stand-alone BSDE for λ as summarized in the following solution method. The claimed optimal allocation rule is obtained by substituting $\eta = (\mu - r)/\sigma$ in the above expression for $\psi\sigma$, and EIS $\equiv 1/\delta$ in expression (3.8.19) for ϱ. Optimality of the proposed solution can be verified by essentially reversing our earlier steps and applying Lemma 3.8.7.

Solution method: Determine $\left(\lambda, \sigma^{\lambda}\right)$ by solving the BSDE

$$\frac{d\lambda_t}{\lambda_t} = -\left(r + g^*\left(\lambda_t\right) - \frac{\gamma}{2} Q_t\left(\sigma_t^{\lambda}\right)\right) dt + \sigma_t^{\lambda} dB_t, \quad \lambda_T = 1, \tag{3.8.21}$$

where $Q_t(z) \equiv z^2 - \left(\gamma^{-1}\eta_t - \left(1-\gamma^{-1}\right) z\right)^2$. The optimal allocation policy (ϱ, ψ) is given by $\varrho_T = 1$ and, for $t < T$,

$$\varrho_t = \beta^{\text{EIS}} \lambda_t^{1-\text{EIS}} \quad \text{and} \quad \psi_t = \frac{1}{\gamma} \frac{\tilde{\mu}_t - r_t}{\sigma_t^2}, \tag{3.8.22}$$

where $\tilde{\mu}_t \equiv \mu_t + (1-\gamma)\,\sigma_t^{\lambda}\sigma_t$.

EXAMPLE 3.8.8 (Unit EIS). Assuming unit EIS ($\delta = 1$), (3.8.22) implies the optimal consumption allocation rate $\varrho_t = \beta$. To relate this to Example 3.7.3, let us write $\bar{\beta}$ and $\bar{\varrho}$ for the parameters β and ϱ in last section's discrete context. If we heuristically think of the infinitesimal time interval $[t, t+dt]$ as a single discrete period, then $\bar{\beta} = e^{-\beta dt} = 1 - \beta dt$ and $\bar{\varrho}_t = \varrho_t dt$. Therefore, the conclusion $\bar{\varrho}_t = 1 - \bar{\beta}$ of Example 3.7.3 corresponds to $\varrho_t = \beta$ in the current context. ♢

EXAMPLE 3.8.9 (Unit CRRA). Assuming $\gamma = 1$, $\tilde{\mu}_t = \mu_t$ and the optimal portfolio allocation in (3.8.22) reduces to $\psi_t = \left(\mu_t - r_t\right)/\sigma_t^2$. This is a mean-variance efficient portfolio in the sense of Section 2.2 (with $\Sigma_t = \sigma_t^2$), whose single-period analysis can be heuristically applied over an infinitesimal time interval $[t, t+dt]$. In the discrete Example 3.7.6, we concluded that the optimal portfolio is myopic. Here we can add the insight of mean-variance efficiency thanks to the Arrow-Pratt approximation of a logarithmic CE. ♢

We encountered another example of a myopic optimal portfolio in Example 3.7.5, which in the current context would correspond to deterministic r, μ and σ. For simplicity, in the following example we take these parameters to be constant and derive a closed-form solution. As in the last example, the combination of a myopic portfolio rule and the Arrow-Pratt CE approximation lead to a mean-variance efficient optimal portfolio.

EXAMPLE 3.8.10 (Constant investment opportunity set). Suppose that $r_t = r$, $\mu_t = \mu$, $\sigma_t = \sigma$, and therefore $\eta_t = \eta \equiv (\mu - r)/\sigma$, are all deterministic constants. BSDE (3.8.21) has a deterministic solution $\left(\sigma^{\lambda} = 0\right)$, implying the mean-variance efficient optimal portfolio[16]

$$\psi_t = \frac{\eta}{\gamma\sigma} = \frac{1}{\gamma}\frac{\mu - r}{\sigma^2}.$$

[16]This optimal portfolio allocation (with extensions) was first shown for the additive case ($\gamma = \delta$) by Merton [1969, 1971], whose work has been seminal for the theory of dynamic optimal consumption and portfolio choice. Merton used the Hamilton-Jacobi-Bellman approach, which is also applicable in the current setting.

For $\delta \neq 1$, the optimal ϱ in (3.8.22) is

$$\varrho_t = p\left(1 + \left(p\beta^{-1/\delta} - 1\right) e^{-p(T-t)}\right)^{-1}, \quad t < T, \quad \varrho_T = 1,$$

where $p \equiv \left(1 - \delta^{-1}\right)\left(r + (2\gamma)^{-1}\eta^2\right) + \delta^{-1}\beta$. To show this, set $\sigma^\lambda = 0$ in BSDE (3.8.21), thus reducing it to an ordinary differential equation (ODE), which after the change of variable $z_t \equiv (\lambda_t/\beta)^{(1-\delta)/\delta}$ becomes $dz_t = (pz_t - \beta)\,dt$ or $d\left(e^{-pt}z_t\right) = (\beta/p)\,de^{-pt}dt$. Integrating from t to T, results in

$$\lambda_t = \left(\frac{\beta^{1/\delta}}{p} + \left(1 - \frac{\beta^{1/\delta}}{p}\right) e^{-p(T-t)}\right)^{\delta/(1-\delta)}.$$

For $\delta = 1$, the optimal consumption policy is given in Example 3.8.8 and, like ψ, does not depend on λ. One may still wish to compute λ in order to verify the solution and also determine $U = \lambda W$. This is easily done since $\log \lambda_t$ satisfies an easy-to-solve linear ODE.

Finally, note that the above solution applies to the expected discounted utility of Example (3.8.2), and more generally to any recursive utility specified by BSDE (3.8.5) with $g = \alpha + \beta u_\delta$, by simply replacing r with $r + \alpha$ while keeping η and $\psi = \eta/\gamma\sigma$ the same (implying that μ changes to $\mu + \alpha$). This can be seen either directly by tracing the role of adding α to g, or through Remark 3.8.5 after the change of the unit of account described in Example (3.8.2). $\diamondsuit$

The more general expression for the optimal portfolio ψ_t in (3.8.22) allows for deviations from mean-variance efficiency due to a stochastically varying shadow-price-of-wealth process. One way of understanding its structure in terms of the discrete theory is through the observation that for the CE $v_t = u_\gamma^{-1}\mathbb{E}_t u_\gamma$, the optimal portfolio rule can be expressed as (3.7.11), where Q is the probability with the conditional density process ξ defined by recursion (3.7.12). Heuristically, if the discrete time period is the time interval $[t, t + dt]$, it is not hard to see, using Ito's lemma, that recursion (3.7.12) corresponds to $d\xi_t/\xi_t = -(1-\gamma)\,\sigma_t^\lambda dB_t$. As we saw in Section 2.8, Girsanov's theorem implies that the drift of the stock's cum-dividend return, which is μ under P, becomes $\tilde{\mu}_t \equiv \mu_t + (1-\gamma)\,\sigma_t^\lambda\sigma_t$ under Q. The claimed optimal portfolio in (3.8.22) is therefore mean-variance efficient under the probability Q. Just as in the last two examples under P, this is a consequence of the Arrow-Pratt CE approximation in the myopic optimal portfolio problem (3.7.11) under Q.

Examples in which σ^λ is stochastic can be formulated in terms of Markovian formulations (more naturally with multiple sources of risk), where the BSDE solutions can be related to corresponding PDE solutions, in a generalization of the argument that related the linear BSDE (2.8.15) to PDE (2.8.19).

3.9. Exercises

Exercise 1 Assume the context of the CAPM equilibrium of Example (3.2.8), including preference transitivity. Provide a dual proof that the equilibrium is effective complete by showing that there exists a complete market $\bar{X} \supseteq X$ such that $(\bar{X}, c)$ is an equilibrium and applying Corollary 3.2.7.
Hint: Define the complete market $\bar{X}$ by setting to zero the present value of all time-one payoffs that are orthogonal to all traded time-one payoffs.

Exercise 2 This exercise outlines a variant of the representative-agent argument of Example 3.2.9 that includes an example of a CAPM equilibrium. To do so, we modify definition of a consumption set and Definition 3.2.9 of a preference correspondence by weakening the monotonicity requirement to introduce an upper bound on allowable consumption: For every preference correspondence $\mathcal{D}$, there exists a consumption plan $b \in \mathcal{L}$ such that $\operatorname{dom}(\mathcal{D}) = \{c \in \mathcal{L} \mid c < b\}$, and for every arbitrage cash flow y, if $x = 0$ or $x \in \mathcal{D}(c)$ then $x + y \in \mathcal{D}(c)$, provided $x + y < b$. (The notation $c < b$ means $c(\omega, t) < b(\omega, t)$ for all (ω, t).) The technical motivation behind this definition is to allow for a quadratic utility representation, as in this exercise's final part, which implies variance aversion and the CAPM pricing equation. Quadratic utility is increasing only up to a maximum. The idea is to only consider equilibria in which the constraint that consumption is below that maximum is non-binding and therefore the nature of the utility form above the maximum is irrelevant. We proceed with a more general specification that emphasizes the aggregation argument.

Agent preferences are specified in terms of a given scale-invariant preference $\mathcal{D}^0$ with $\operatorname{dom}(\mathcal{D}^0) \equiv \{c \in \mathcal{L} \mid c < 0\}$. For every $i \in \{1, \ldots, I\}$, the preference correspondence of agent $(\mathcal{D}^i, e^i)$ is specified in terms of $b^i \in \mathcal{L}$ by $\operatorname{dom}(\mathcal{D}^i) \equiv \{c \in \mathcal{L} \mid c < b^i\}$ and $\mathcal{D}^i(c) \equiv \mathcal{D}^0(c - b^i)$. Further assume that for all i, there exist $v^i, w^i \in \mathbb{R}$ such that

$$e^i - w^i 1_{\Omega \times \{0\}} \in X, \quad b^i - v^i 1_{\Omega \times \{0\}} \in X \quad \text{and} \quad w^i < v^i,$$

and let $b \equiv \sum_i b^i$, $e \equiv \sum e^i$, $v \equiv \sum_i v^i$, $w \equiv \sum_i w^i$. The allocation $c \equiv \left(c^1, \ldots, c^I\right)$ is defined by

$$c^i \equiv b^i - \frac{v^i - w^i}{v - w}(b - e), \quad i = 1, \ldots, I.$$

Assume that $e < b$ and therefore $c^i < b^i$ for all i by construction.

(a) Show that the allocation c is market-clearing and X-feasible, assuming X is arbitrage-free.

(b) Show that (X, c) is an equilibrium if and only if e is an optimal consumption plan for the representative agent $(\mathcal{D}, e)$, defined by $\operatorname{dom}(\mathcal{D}) \equiv \{c \in \mathcal{L} \mid c < b\}$ and $\mathcal{D}(c) \equiv \mathcal{D}^0(c - b)$.

(c) Show that if (X, c) is an equilibrium and $\mathcal{D}(e)$ is convex, then (X, c) is an effectively complete market equilibrium.

(d) Specialize the setting by assuming that $T = 1$ and $\mathcal{D}^0$ has the quadratic utility representation $U^0(c) = -c_0^2 - \beta \mathbb{E}[c_1^2]$, where $\mathbb{E}$ is expectation under a given full-support probability and β is a positive scalar. Assume further that the parameters $b^1, \ldots, b^I$ are (deterministic) constants. Verify that the agent preferences are variance averse in the sense of Example 3.2.8. Adding the assumptions of a traded money-market account and the regularity condition $\operatorname{var}[e_1] > 0$ and $e_0 \neq w$, all the CAPM assumptions of Example 3.2.8 are satisfied. In this context, give an alternative derivation by the CAPM beta-pricing equation (3.2.4) by using Proposition 3.3.4 and the optimality of the aggregate endowment for the representative agent.
Hint: Construct a SPD whose time-one value is an affine function of the market return.

Exercise 3 This exercise outlines a variant of the representative-agent pricing argument of Example 3.2.9, where preferences are assumed to be translation invariant instead of scale invariant. (As shown in Appendix A, an additive utility representing a translation-invariant preference correspondence necessarily takes an exponential form, which in the expected-utility case corresponds to constant absolute risk aversion.) Every preference correspondence in this exercise is assumed to have as its domain the entire consumption set $\mathcal{L}$. Agents are specified in terms of a fixed reference agent $(\mathcal{D}^0, e^0)$. The key assumption is that $\mathcal{D}^0$ is **translation invariant**:

$$\mathcal{D}^0(c) = \mathcal{D}^0(c + \theta) \text{ for all } \theta \in \mathbb{R}.$$

For $i = 1, \ldots, I$, agent $(\mathcal{D}^i, e^i)$ is specified in terms of the parameters $(\alpha^i, \theta^i, x^i) \in (0, \infty) \times \mathbb{R} \times X$ by

$$\mathcal{D}^i(c) \equiv \alpha^i \mathcal{D}^0\left(\frac{c}{\alpha^i}\right), \quad e^i \equiv \alpha^i e^0 + \theta^i + x^i.$$

Let $\alpha \equiv \sum_{i=1}^I \alpha^i$, $\theta \equiv \sum_{i=1}^I \theta^i$, $x \equiv \sum_{i=1}^I x^i$, $e \equiv \sum_{i=1}^I e^i$, and define the **representative agent** $(\mathcal{D}, e)$ by

$$\mathcal{D}(c) = \alpha \mathcal{D}^0\left(\frac{c}{\alpha}\right), \quad e = \alpha e^0 + \theta + x.$$

The allocation $c = \left(c^1, \ldots, c^I\right)$ is defined by

$$c^i = \theta^i + \frac{\alpha^i}{\alpha}(e - \theta). \tag{3.9.1}$$

(a) Show that (X, c) is an equilibrium if and only if the consumption plan e is optimal for the **representative agent** given the market X, that is, $X \cap \mathcal{D}(e) = \emptyset$.

(b) Suppose (X, c) is an equilibrium and $\mathcal{D}(e)$ is convex. Show that (X, c) is an effectively complete market equilibrium.

(c) Fix an underlying full-support probability and assume that for some $\beta \in (0,1)$, every $\mathcal{D}^i$ has the utility representation

$$U^i(c) = \mathbb{E}\left[-\alpha^i \sum_{t=0}^{T} \beta^t \exp\left(-\frac{c_t}{\alpha^i}\right)\right].$$

Verify that the above preference assumptions are satisfied and specify a state-price density π that corresponds to the utility gradient of the representative agent at the aggregate endowment.

(d) Consider a sequence of the model of part (c) parameterized by the time-horizon $T = 1, 2, \ldots$, and take as given an infinite filtration $\{\mathcal{F}_t \mid t = 1, 2, \ldots\}$ and corresponding adapted process $e = (e_0, e_1, \ldots)$. The information and endowment of the T^{th} model is the restriction of the respective quantity over the time set $\{0, 1, \ldots, T\}$, that is, the filtration is $\{\mathcal{F}_t \mid t = 1, 2, \ldots, T\}$ and the endowment is $(e_0, e_1, \ldots, e_T)$. All other parameters are common across all models. The endowment process is strictly positive and follows the dynamics $e_t = e_{t-1}(1 + g_t)$, where the random variable g_t is stochastically independent of $\mathcal{F}_{t-1}$ and takes the value $\varepsilon \in (0,1)$ with probability $p \in (0.5, 1)$ and $-\varepsilon$ with probability $1-p$. Assume that $m \equiv p\log(1+\varepsilon) + (1-p)\log(1-\varepsilon) > 0$. What is the limit of the equilibrium short rate r_t as $t \to \infty$. You can use the law of large numbers from probability theory, which implies that $t^{-1}\log(e_t/e_0) = t^{-1}\sum_{s=1}^{t}\log(1+g_t)$ converges (with probability one) to the mean of $\log(1+g_t)$, which is $m > 0$, and therefore $e_t \to \infty$ as $t \to \infty$ (with probability one).

Exercise 4 Assume a single period ($T = 1$) and fix an underlying full-support probability throughout. An agent's time-one consumption is restricted to take values in a non-empty open interval $D \subseteq \mathbb{R}$, which is agent specific. Given K states, D^K denotes the set of all D-valued random variables, and $C^2_{++}(D)$ denotes the set of all strictly increasing and twice-continuously differentiable functions of the form $u : D \to \mathbb{R}$, with the first two derivatives denoted u' and u'', respectively. Utilities over time-one consumption are assumed to be of the expected-utility form $\mathbb{E}u(c)$, $c \in D^K$, where $u \in C^2_{++}(D)$. The corresponding **coefficient of absolute risk aversion** is the function $A^u : D \to \mathbb{R}$ defined by

$$A^u(w) \equiv -\frac{u''(w)}{u'(w)}, \quad w \in D.$$

(a) Suppose that $u, \tilde{u} \in C^2_{++}(D)$. Show that $A^u = A^{\tilde{u}}$ if and only if there exist $a \in (0,\infty)$ and $b \in \mathbb{R}$ such that $\tilde{u} = au + b$. Then use Theorem A.2.4 to show that $\tilde{u}$ and u represent the same preference, in the sense that $\mathbb{E}\tilde{u}(x) > \mathbb{E}\tilde{u}(y) \iff \mathbb{E}u(x) > \mathbb{E}u(y)$ for all $x, y \in D^K$, if and only if $A^{\tilde{u}} = A^u$.

(b) We call u **HARA with coefficients** $(\alpha, \beta) \in \mathbb{R}$ if $u \in C^2_{++}(D)$ with

$$D \equiv \{w \mid \alpha + \beta w > 0\} \quad \text{and} \quad A^u(w) \equiv \frac{1}{\alpha + \beta w}.$$

The parameter β is the **coefficient of cautiousness.** (For $\beta < 0$, consumption is bounded above, just as in Exercise 3.9.) Show that u is HARA with coefficients (α, β) if and only if there exist $a \in (0, \infty)$ and $b \in \mathbb{R}$ such that

$$au(w) + b = \begin{cases} \frac{1}{\beta - 1}(\alpha + \beta w)^{(\beta-1)/\beta}, & \text{if } \beta \neq 0 \text{ and } \beta \neq 1; \\ \log(\alpha + w), & \text{if } \beta = 1; \\ -\alpha \exp(-w/\alpha), & \text{if } \beta = 0 \text{ and } \alpha > 0. \end{cases} \tag{3.9.2}$$

(c) Explain why u HARA implies that u is strictly concave.

Exercise 5 This exercise, which has Exercise 4 as a prerequisite, formulates and analyzes a representative-agent equilibrium for agents who maximize HARA expected utility. The formulation nests Example 3.2.9 and Exercises 2 and 3 for the expected utility case. [17]

Assume $T = 1$ and fix an underlying full-support probability over K states. For simplicity, agents can only trade in a forward market to modify their time-one consumption. Taken as given is a column vector $V \equiv (V_1, \ldots, V_J)'$, where $V_j \in \mathbb{R}^K$ represents the time-one value of some asset that can be traded in a forward market. A **portfolio** is a row vector $\theta \equiv (\theta_1, \ldots, \theta_J)$, where $\theta_j \in \mathbb{R}$ represents a number of forward contracts in asset j. Given a **forward-price vector** $f \equiv (f_1, \ldots, f_J)' \in \mathbb{R}^{J \times 1}$, the portfolio θ generates the cash flow $(0, \theta(V - f))$. Since all time-zero cash flows are zero, we define the **market** (given f) by

$$X_f \equiv \left\{\theta(V - f) \mid \theta \in \mathbb{R}^{1 \times J}\right\}.$$

Each agent $i \in \{1, \ldots, I\}$ has a (time-one) endowment of the form

$$e^i \equiv a^i + b^i V, \quad a^i \in \mathbb{R}, \quad b^i \in \mathbb{R}^{1 \times J},$$

and maximizes the utility $\mathbb{E}u_i$ over time-one consumption, where u_i is HARA with coefficients (α^i, β). Note that $\mathbb{E}$ and β are the same across agents. The set of admissible (time-one) consumption plans for agent i is D_i^K, where $D_i \equiv \{w \mid \alpha^i + \beta w > 0\}$. The consumption plan $c^i \in D_i^K$ is **optimal** for agent i given f if there is no $x \in X_f$ such that $c^i + x \in D_i^K$ and $\mathbb{E}u_i(c^i + x) > \mathbb{E}u_i(c^i)$. An **allocation** is any element of $D_1^K \times \cdots \times D_I^K$. The allocation $c \equiv (c^1, \ldots, c^I)$ is said to be f**-feasible** if $c^i - e^i \in X_f$ for all i, and **market-clearing** if $\sum_i c^i = e$. An **equilibrium** is a pair (f, c) of a forward-price vector f and an f-feasible and market-clearing allocation c such that for all $i \in \{1, \ldots, I\}$, c^i is optimal for agent i given f.

[17] An extension with asymmetric information is given in DeMarzo and Skiadas [1998, 1999].

Define the aggregate parameters

$$\alpha \equiv \sum_{i=1}^{I} \alpha^i, \quad a \equiv \sum_{i=1}^{I} a^i, \quad b \equiv \sum_{i=1}^{I} b^i, \quad e \equiv \sum_{i=1}^{I} e^i = a + bV.$$

The **representative agent** is endowed with the aggregate endowment e and maximizes the expected utility $\mathbb{E}u$, where u is HARA with coefficients (α, β). The set of admissible consumption plans for the representative agent is therefore D^K, where $D \equiv \{w \mid \alpha + \beta w > 0\}$. Assume that for every agent i, $e^i \in D_i^K$ and therefore $e \in D^K$.

(a) Fix any forward-price vector f such that X_f is arbitrage-free (meaning $X_f \cap \mathbb{R}_+^K = \{0\}$). Show that

$$c^i \equiv a^i + b^i f + \frac{\alpha^i + \beta\left(a^i + b^i f\right)}{\alpha + \beta\left(a + bf\right)} b\left(V - f\right), \quad i = 1, \ldots, I, \tag{3.9.3}$$

defines an allocation $c \in D_1^K \times \cdots \times D_I^K$, which is f-feasible and market-clearing, and there exist scalars $\lambda_f^i > 0$ such that

$$\lambda_f^i u_i'\left(c^i\right) = u'\left(e\right), \quad i = 1, \ldots, I.$$

(b) Show that for all $f \in \mathbb{R}^{J \times 1}$, the aggregate endowment e is optimal for the representative agent given f (that is, there is no $x \in X_f$ such that $e + x \in D^K$ and $\mathbb{E}u\left(e + x\right) > \mathbb{E}u\left(e\right)$) if and only if

$$f = \frac{1}{\mathbb{E}u'\left(e\right)} \mathbb{E}\left[u'\left(e\right) V\right]. \tag{3.9.4}$$

(c) Show that equations (3.9.4) and (3.9.3) define an equilibrium (f, c).

(d) Show that the equilibrium of part (c) is an effectively complete market equilibrium: for all $x^1, \ldots, x^I$ such that $c^i + x^i \in D_i^K$, if $\mathbb{E}u_i\left(c^i + x^i\right) \geq \mathbb{E}u_i\left(c^i\right)$ for all i and $\mathbb{E}u_i\left(c^i + x^i\right) > \mathbb{E}u_i\left(c^i\right)$ for some i, then $\sum_{i=1}^{I} x^i \notin X_f$. You can base your argument on Proposition 3.3.8, but you should give a proof from first principles.

(e) Show that the equilibrium of part (c) is unique. *Hint:* If f is part of some equilibrium and the market-clearing allocation c is Pareto optimal and f-feasible, then (f, c) is an equilibrium.

(f) Assume $\beta \equiv -1$. Given the equilibrium of part (c), call R a **traded forward return** if there exists a portfolio θ such that $\theta f \neq 0$ and $R = \theta V / \theta f$. Assume that $a + bf \neq 0$ and $\operatorname{var}\left[e\right] > 0$, and define the forward return of the aggregate endowment:

$$R^m \equiv \frac{a + bV}{a + bf} = \frac{e}{\text{forward price of } e}.$$

Assuming that R^0 is a forward traded return that is uncorrelated to R^m, show that every traded forward return R satisfies the CAPM equation

$$\mathbb{E}\left[R - R^0\right] = \frac{\operatorname{cov}\left[R, R^m\right]}{\operatorname{var}\left[R^m\right]} \mathbb{E}\left[R^m - R^0\right].$$

Exercise 6 (Preferences for the Timing of Resolution of Uncertainty) In this course preferences are defined over consumption plans taking the information tree (filtration) as given. In settings in which the information structure is not fixed, it is natural to consider preferences not only over consumption but also over information.[18] In particular, an agent may have preferences for earlier or later resolution of uncertainty about future consumption. In the context of Example 3.3.11, given the consumption plan b, an agent with preferences for early resolution of uncertainty would prefer to have all T coin tosses announced at time zero, rather than wait to find out the outcome of the t^{th} coin toss at time t.

More formally, let C be a (non-empty) set of consumption plans and let Φ be the set of every filtration $\{\mathcal{F}_t : 0, \ldots, T\}$ satisfying $\mathcal{F}_0 = \{\Omega, \emptyset\}$ and $\mathcal{F}_T = 2^{\Omega}$. We consider a (non-normalized) utility function $V_0(\cdot)$ over the set of all pairs $(c, \{\mathcal{F}_t\}) \in C \times \Phi$ such that c is adapted to $\{\mathcal{F}_t\}$. Taking as primitive an underlying probability with expectation operator $\mathbb{E}$ and the functions $F_t : \mathbb{R}^2 \to \mathbb{R}$, $t = 0, \ldots, T-1$ and $F_T : \mathbb{R} \to \mathbb{R}$, assume that $V_0(c, \{\mathcal{F}_t\})$ is the initial value of the process V that solves the backward recursion

$$V_t = F_t(c_t, \mathbb{E}[V_{t+1} \mid \mathcal{F}_t]), \quad t = 0, \ldots, T-1; \quad V_T = F_T(c_T).$$

We say that the utility function $V_0(\cdot)$ expresses preferences for earlier resolution of uncertainty if for all $(c, \{\mathcal{F}_t^1\})$ and $(c, \{\mathcal{F}_t^2\})$ in its domain,

$$\mathcal{F}_t^1 \subseteq \mathcal{F}_t^2 \text{ for all } t \quad \text{implies} \quad V_0\left(c, \left\{\mathcal{F}_t^1\right\}\right) \leq V_0\left(c, \left\{\mathcal{F}_t^2\right\}\right).$$

Preferences for late resolution of uncertainty are defined analogously, with the last inequality reversed.

(a) Use Jensen's inequality to show that if the functions F_t are convex (resp. concave) in their second (utility) argument for all $t < T$, then $V_0(\cdot)$ expresses preferences for earlier (resp. later) resolution of uncertainty.

(b) Suppose the choice of F_t corresponds to the functional form of expected discounted utility. What does that imply about preferences over information?

(c) Suppose the utility process associated with $(c, \{\mathcal{F}_t\})$ is of the EZW form of Example (3.5.1), with CRRA γ and inverse EIS δ. Characterize preferences for the timing of resolution of uncertainty in terms of the relative value of the parameters γ and δ. You can use without proof the fact that if u is of the HARA form (3.9.2) and $\beta \in (0, 1)$, then the function $u^{-1}((1-\beta)u(x) + \beta u(y))$ is concave if u is concave, and

[18] The notion of preferences for the timing of resolution of uncertainty appears in Kreps and Porteus [1978] using a formalism based on preferences over probability distribution. The approach without probabilities in this exercise is from Skiadas [1998].

convex if u is convex. (This follows by the argument of Example 3.2.9 for the power-or-logarithmic case and a similar argument applies for the exponential case.)

Exercise 7 In this problem assume scale-invariant recursive utility with a proportional aggregator that takes the unit-EIS form $g_t(x) = x^{1-\beta}$ for some $\beta \in (0,1)$.

(a) Specialize the result of Proposition 3.6.2 (pricing and consumption growth) to this case.

(b) Specialize the result of Proposition 3.6.4 (pricing and market returns) to this case.

(c) Show that the procedures for computing π in parts (a) and (b) can be easily derived from each other, given the relationship between consumption growth and market returns of Example 3.6.7 (unit EIS).

(d) Further specialize the results of parts (a) and (b) by assuming unit-EIS Epstein-Zin-Weil utility. (Provide as specific expressions as you can.)

Exercise 8 An alternative approach to proving Theorem 3.7.2 on optimal consumption and portfolio choice is based on the ideas of dynamic programming.[19] While it is not difficult to show the complete Theorem 3.7.2, in this problem you are only asked to use a dynamic programming argument to prove the sufficiency part: If (ϱ, ψ) is constructed according to the Theorem's algorithm, then (ϱ, ψ) finances an optimal consumption plan. Also, for notational brevity, assume the proportional aggregator $g_t = g$ is time independent.

Define a value function to be a mapping $\mathcal{V} : (0,\infty) \to \mathcal{L}_{++}$ that assigns to each $w \in (0,\infty)$ a strictly positive adapted process $\mathcal{V}(w)$. For each spot (F,t), think of $\mathcal{V}(w)(F,t)$ as the optimal utility level at spot (F,t) given total wealth w at that spot. (In a typical application, the dependence on (F,t) is through the value $Z(F,t)$ of some Markov process Z on that spot, and the value function is therefore defined as a function of wealth and the Markov state. Here, $Z(F,t)$ can be thought of as being the spot itself, which can be identified with the entire history leading to that spot.)

Write $\mathcal{V}_t(w)$ for the random variable that takes the value $\mathcal{V}(w)(F,t)$ on F for every spot (F,t). The corresponding **Bellman equation** is

$$\mathcal{V}_{t-1}(w) = \max_{\varrho_{t-1},\psi_t} f\left(\varrho_{t-1} w, \upsilon_{t-1}\left(\mathcal{V}_t\left(\left(1-\varrho_{t-1}\right) w R_t^{\psi}\right)\right)\right),$$

Proceed in the following steps.

[19] Richard Bellman coined the term "dynamic programming" in 1950 at the RAND corporation as a marketing ploy in getting his research agenda past his boss, and later secretary of defense, Wilson. As recounted by Bellman's colleague and coauthor Dreyfus [2002], Wilson was no fan of mathematical research.

(a) What exactly is the mathematical content of the Bellman equation at each spot? What is the set over which maximization occurs at each spot? Be as precise and specific as you can.

(b) Conjecture the functional form of $\mathcal{V}(w)$ and explain your reasoning.

(c) Show that the algorithm of Theorem 3.7.2 produces a solution to the Bellman equation.

(d) Provide what is known as a verification argument, that is, show that given the right terminal condition, a solution to the Bellman equation defines an optimal policy. You should explain how the Bellman equation defines the optimal policy and then prove optimality using only the budget equation, the utility definition, and the Bellman equation.

APPENDIX A

Additive Utility Representations

This appendix presents some basic results on the representation of preferences by additive utility functions. Simple ordinal assumptions, mainly separability and continuity, are enough for the existence of an additive utility representation. Additional ordinal restrictions, like preference convexity, scale invariance or translation invariance, are shown to have strong implications for the structure of an additive utility representation. The appendix concludes with a discussion of risk aversion in the context of expected utility.

A.1. Utility representations of preferences

In this appendix we discuss preferences over the set $C \equiv (\ell, \infty)^N$ for some integer $N \geq 2$ and constant $\ell \in [-\infty, 1)$. We treat a constant $\alpha \in (\ell, \infty)$ as an element of C by identifying it with $(\alpha, \alpha, \ldots, \alpha)$. From Chapter 3, we adopt Definition 3.1.1 of a preference correspondence and Definition 3.3.1 of a utility (but without the restriction $C = \mathcal{L}_{++}$). A preference correspondence $\mathcal{D}$ defines a corresponding binary relation $\succ$ on C by letting $a \succ b \iff a - b \in \mathcal{D}(b)$. (Formally speaking, a **binary relation** $\succ$ on C is a subset of $C \times C$ and the notation $a \succ b$ means $(a, b) \in \succ$.) We henceforth use the term **preference** to mean any binary relation $\succ$ on C defined as above by some preference correspondence $\mathcal{D}$, in which case we say that a utility $U : C \to \mathbb{R}$ **represents** $\succ$ if it represents $\mathcal{D}$, that is, for all $a, b \in C$,

$$a \succ b \quad \iff \quad U(a) > U(b).$$

All preferences in this appendix are assumed to have a utility representation, but we use the preference notation $\succ$ where we wish to emphasize that a property is ordinal and therefore not dependent on any particular choice of a utility representation. In the remainder of this section, we review necessary and sufficient conditions for a preference to admit a utility representation. The argument is simplified by our standing assumption of monotone preferences.[1]

Let the relation $\succeq$ on C be defined by $a \succeq b \iff \text{not}\, b \succ a$. Clearly, $\succ$ admits a utility representation U if and only if $\succeq$ admits a utility representation U in the sense (A.1.1) of the following result.

[1] Debreu [1983] (based on a 1954 working paper) shows the existence of a continuous utility representation of a continuous, complete and transitive binary relation, without monotonicity.

THEOREM A.1.1. *For any binary relation $\succeq$ on C, there exists a utility function $U : C \to \mathbb{R}$ such that*

$$a \succeq b \iff U(a) \geq (b). \tag{A.1.1}$$

if and only if $\succeq$ is

- ***total:*** *for all $a, b \in C$, either $a \succeq b$ or $b \succeq a$.*
- ***transitive:*** *$a \succeq b$ and $b \succeq c$ implies $a \succeq c$.*
- ***increasing:*** *for all $a, b \in C$, if $b \neq a \geq b$ then not $b \succeq a$.*
- ***continuous:*** *for every sequence $\{(a_n, b_n)\}$ such that $a_n \succeq b_n$, if $\lim_{n\to\infty}(a_n, b_n) = (a, b) \in C \times C$, then $a \succeq b$.*

PROOF. The "only if" part is immediate from the definitions. Conversely, suppose that $\succeq$ is total, transitive, increasing and continuous. On C, define the function $U(c) \equiv \inf\{\alpha \in \mathbb{R} \mid \alpha \succeq c\}$ and the relation $a \sim b$ if and only if both $a \succeq b$ and $b \succeq a$. We proceed in six steps.

Step 1. $U(c) \sim c$. Choose any sequence in $\{\alpha_n\}$ in $\mathbb{R}$ that converges to $U(c)$ and $\alpha_n \succeq c$ for all n. Since $\succeq$ is continuous, $U(c) \succeq c$ and $U(c) = \min\{\alpha \in \mathbb{R} \mid \alpha \succeq c\}$. For all $n = 1, 2, \ldots$, it is not the case that $U(c) - n^{-1} \succeq c$, and since $\succeq$ is total, $c \succeq U(c) - n^{-1}$. Since $\succeq$ is continuous, $c \succeq U(c)$. This proves that $U(c) \sim c$.

Step 2. $a \succeq b \iff U(a) \geq U(b)$. Since $\succeq$ is transitive, if $a \succeq b$, then $\{\alpha \in \mathbb{R} \mid \alpha \succeq a\} \subseteq \{\alpha \in \mathbb{R} \mid \alpha \succeq b\}$, and therefore $U(a) \geq U(b)$ by the definition of U. Conversely, since $\succeq$ is monotone, $U(a) \geq U(b)$ implies that $U(a) \succeq U(b)$. By Step 1, $a \succeq U(a)$ and $U(b) \succeq b$. Since $\succeq$ is transitive, it follows that $a \succeq b$.

Step 3. U is increasing. Suppose $a, b \in C$ and $b \neq a \geq b$. Since $\succeq$ is total and increasing, $a \succeq b$ and not $b \succeq a$. By Step 2, $U(a) \geq U(b)$ and not $U(b) \geq U(a)$, and therefore $U(a) > U(b)$.

Step 4. For all $u \in (\ell, \infty)$, $U(u) = u$. Since $\succeq$ is total and increasing, for all $\alpha, u \in (\ell, \infty)$, $\alpha \succeq u$ if and only if $\alpha \geq u$. The claim now follows by the definition of U.

Step 5. For all $u \in \mathbb{R}$, if $u \sim c$, then $u = U(c)$. By Step 1, $c \sim U(c)$. Since $\succeq$ is transitive, $u \sim U(c)$. By Steps 2 and 4, it then follows that $u = U(c)$.

Step 6. U is continuous. Consider any sequence $\{c_n\}$ in C converging to $c \in C$. Suppose first that $\{U(c_n)\}$ converges to some u. Since $U(c_n) \sim_n c_n$ for all n and $\succeq$ is continuous, it follows that $u \sim c$ and therefore $u = U(c)$ by Step 5. Consider now the general case, where we do not know a-priori that $\{U(c_n)\}$ converges. Since $c_n \to c$, there exist $a, b \in (\ell, \infty)$ such that c and all components of c_n are valued in the interval $[a, b]$ and therefore $U(c)$ and $U(c_n)$ are all valued in the compact set $[U(a), U(b)]$ (which equals $[a, b]$, although that does not matter here). It follows that every subsequence of $\{U(c_n)\}$ has a convergent further subsequence, which as shown earlier must converge to $U(c)$. This proves that $\{U(c_n)\}$ also converges to $U(c)$. □

A.2. Additive utility representations

The rest of this appendix discusses preferences on $(\ell,\infty)^N$, where $\ell \in [-\infty, 1)$, that admit an additive utility representation in the following sense.

DEFINITION A.2.1. A utility $U : (\ell,\infty)^N \to \mathbb{R}$ is **additive** if there exist functions $U_n : (\ell,\infty) \to \mathbb{R}$ such that

$$U(x) = \sum\nolimits_{n=1}^{N} U_n(x_n), \quad x \in (\ell,\infty)^N. \tag{A.2.1}$$

Given any $x, y \in (\ell,\infty)^N$ and $A \subseteq \{1,\ldots,N\}$, $x_A y$ denotes the element of $(\ell,\infty)^N$ defined by

$$(x_A y)_n = \begin{cases} x_n, & \text{if } n \in A; \\ y_n, & \text{if } n \notin A. \end{cases}$$

DEFINITION A.2.2. A preference $\succ$ on $(\ell,\infty)^N$ is **separable** if

$$x_A z \succ y_A z \quad \Longleftrightarrow \quad x_A \tilde{z} \succ y_A \tilde{z}, \tag{A.2.2}$$

for all $x, y, z, \tilde{z} \in (\ell,\infty)^N$ and $A \subseteq \{1,\ldots,N\}$.

A preference that admits an additive utility representation is clearly separable. The following remarkable theorem gives a converse. It is a special case of Debreu's additive representation theorem,[2] but it captures the deeper aspects of Debreu's theorem. The proof would take us too far afield and will not be given here.

THEOREM A.2.3 (Existence of Additive Representations). *Suppose $N > 2$ and $\succ$ is a preference on $(\ell,\infty)^N$ that admits some utility representation. Then $\succ$ admits an additive utility representation if and only if it is separable.*

The assumption $N > 2$ is critical; separability is not sufficient for the existence of an additive representation if $N = 2$. The uniqueness part of Debreu's theorem (which also applies if $N = 2$) is stated and proved below. The result has important ramifications for the discussion of the limitations of additive utilities in Chapter 3, as well as our later parametric characterization of scale or translation invariant additive preferences, and the associated foundation of expected utility with constant relative or absolute risk aversion.

[2] Debreu [1983] characterizes continuous additive utility representations in a way that includes Theorems A.2.3 and A.2.4. Debreu's theorems are part of a broader theory of measurement, which is reviewed in the monographs of Krantz et al. [1971] and Narens [1985]. These authors present an algebraic theory that generalizes Debreu's topological results (see also Wakker [1988]). A detailed proof of Debreu's theorem can be found in Wakker [1989].

THEOREM A.2.4 (Uniqueness of Additive Representations). *For all additive utilities U and $\tilde{U}$ on $(\ell,\infty)^N$, where $N \geq 2$, the following two conditions are equivalent:*

(1) *U and $\tilde{U}$ are **ordinally equivalent:** For all $x, y \in (\ell,\infty)^N$,*

$$U(x) > U(y) \iff \tilde{U}(x) > \tilde{U}(y).$$

(2) *U and $\tilde{U}$ are **related by a positive affine transformation:** There exist $a \in (0,\infty)$ and $b \in \mathbb{R}^N$ such that*

$$\tilde{U}_n = aU_n + b_n, \quad n = 1,\dots,N.$$

PROOF. That (2) $\implies$ (1) is immediate. We show the converse, assuming that $\ell = -\infty$. This is without loss of generality: If $\ell > -\infty$, then apply the result for $\ell = -\infty$ to the utility functions that map $z \in \mathbb{R}^N$ to $\sum_n U_n(\ell + e^{z_n})$ and $\sum_n \tilde{U}_n(\ell + e^{z_n})$, respectively.

Consider any ordinally equivalent additive utilities U and $\tilde{U}$ on $\mathbb{R}^N$ that satisfy, for all $n \in \{1,\dots,N\}$,

$$U_n(0) = \tilde{U}_n(0) = 0 \quad \text{and} \quad U_1(1) = \tilde{U}_1(1) = 1. \tag{A.2.3}$$

The claim is that $\tilde{U}_n = U_n$ for all n, which proves (1) $\implies$ (2), since any additive utility on $\mathbb{R}^N$ can be made to satisfy normalization (A.2.3) after a positive affine transformation. Fixing an arbitrary $n \in \{2,\dots,N\}$, $L \in (-\infty,0)$ and scalar Δ such that $L + \Delta > 1$, define the functions $f, g : [0,1] \to \mathbb{R}$ by

$$f(z) \equiv \frac{U_1(L+z\Delta) - U_1(L)}{U_1(L+\Delta) - U_1(L)}, \quad g(z) \equiv \frac{U_n(L+z\Delta) - U_n(L)}{U_1(L+\Delta) - U_1(L)}.$$

Define also $\tilde{f}$ and $\tilde{g}$ by putting a tilde over f, g and every instance of U in the above display. Applying Lemma A.2.5 immediately following this proof to these functions, it follows that $U_1(x) = \tilde{U}_1(x)$ and $U_n(x) = \tilde{U}_n(x)$ for all $x \in [L, L+\Delta]$. This proves that $\tilde{U}_n = U_n$, since every $x \in \mathbb{R}$ is in an interval of the form $[L, L+\Delta] \supset [0,1]$. □

LEMMA A.2.5. *Suppose the functions $f, g, \tilde{f}, \tilde{g} : [0,1] \to \mathbb{R}$ are increasing and continuous, and satisfy*

$$f(0) = g(0) = \tilde{f}(0) = \tilde{g}(0) = 0 \quad \text{and} \quad f(1) = \tilde{f}(1) = 1, \tag{A.2.4}$$

Suppose also that for all $x, y, z, w \in [0,1]$,

$$f(x)+g(y) = f(z)+g(w) \iff \tilde{f}(x)+\tilde{g}(y) = \tilde{f}(z)+\tilde{g}(w). \tag{A.2.5}$$

Then $f = \tilde{f}$ and $g = \tilde{g}$.

PROOF. Let N be any positive integer such that $2^{-N} < g(1)$. Given any $n \in \{N, N+1, \dots\}$, define $x_k^n \in [0,1]$ and $y^n \in (0, g(1))$ by

$$f(x_k^n) = k2^{-n}, \quad k = 0, 1, \dots, 2^n, \quad \text{and} \quad g(y^n) = 2^{-n}. \tag{A.2.6}$$

Note that $x_0^n = 0$ and $x_{2^n}^n = 1$. Since $g(0) = 0$, we have

$$f(x_k^n) + g(0) = f\left(x_{k-1}^n\right) + g(y^n), \quad k = 1, \ldots, 2^n.$$

By assumption (A.2.5), it is also true that

$$\tilde{f}(x_k^n) + \tilde{g}(0) = \tilde{f}\left(x_{k-1}^n\right) + \tilde{g}(y^n), \quad k = 1, \ldots, 2^n. \tag{A.2.7}$$

Since $\tilde{g}(0) = \tilde{f}(0) = 0$, it follows that

$$1 = \tilde{f}(1) = \sum\nolimits_{k=1}^{2^n} \tilde{f}(x_k^n) - \tilde{f}\left(x_{k-1}^n\right) = 2^n \tilde{g}(y^n).$$

This proves that $\tilde{g}(y^n) = 2^{-n}$, which together with (A.2.7) shows that $\tilde{f}(x_k^n) = k2^{-n}$ for $k > 0$. Comparing this conclusion to (A.2.6), we have proved that the functions f^{-1} and $\tilde{f}^{-1}$ are equal on the set $D_n = \{k2^{-n} : k = 0, \ldots, 2^n\}$, for all $n \geq N$. Since the set $\bigcup_{n \geq N} D_n$ is dense in $[0, 1]$ and the functions f^{-1} and $\tilde{f}^{-1}$ are continuous, it follows that $f^{-1} = \tilde{f}^{-1}$ and therefore $f = \tilde{f}$.

To show that $g = \tilde{g}$, we apply the same argument with $(F, G, \tilde{F}, \tilde{G})$ in place of $(f, g, \tilde{f}, \tilde{g})$, where

$$F(z) \equiv \frac{g(z)}{g(1)}, \quad \tilde{F}(z) \equiv \frac{\tilde{g}(z)}{\tilde{g}(1)}, \quad G(z) \equiv \frac{f(z)}{g(1)}, \quad \tilde{G}(z) \equiv \frac{\tilde{f}(z)}{\tilde{g}(1)}.$$

The conclusion $F = \tilde{F}$ implies that $g = a\tilde{g}$ for some $a \in (0, \infty)$. Choose any $\varepsilon, \delta > 0$ such that $f(\varepsilon) = g(\delta)$, and therefore $f(\varepsilon) + g(0) = f(0) + g(\delta)$ (by (A.2.4)). By (A.2.5), it must also be the case that $\tilde{f}(\varepsilon) = \tilde{g}(\delta)$. Since $f = \tilde{f}$ and $\tilde{g} = ag$, this shows that $a = 1$ and therefore $g = \tilde{g}$, completing the proof. □

A.3. Concave additive representations

A preference $\succ$ on $(\ell, \infty)^N$ is **convex** if the set $\{x : x \succ y\}$ is convex for every $y \in (\ell, \infty)^N$. Preference convexity is a key component of this text's theory as it expresses preference for smoothing across time and states. As pointed out in Section 3.3, a convex preference admitting a utility representation may admit no concave utility representation. This is also true of additive representations: $U(x) \equiv e^{x_1} - e^{-x_2}$ defines an additive utility on $(0, \infty)^2$ that represents a convex preference (uniquely up to positive affine transformations). On the other hand, we have the following positive and rather surprising result.[3]

THEOREM A.3.1. *Suppose the preference $\succ$ on $(\ell, \infty)^N$ is convex and it admits an additive utility representation $U : (\ell, \infty)^N \to \mathbb{R}$. Then at least $N - 1$ of the functions U_n defined in (A.2.1) are concave.*

PROOF. The theorem follows from Lemma A.3.3 below. □

[3]The proof of Theorem A.3.1 is based on Yaari [1977], who credited Koopmans with a different proof, as well as Gorman for the twice-differentiable case, in both cases in unpublished papers.

In fact, the argument that follows proves a stronger result in that our standing monotonicity assumption on preferences plays no role and the interval (ℓ, ∞) can be replaced by any other interval. The remainder of this section shows two lemmas that lead to the proof of Theorem A.3.1. These are fairly technical and can be omitted.

The following lemma is stated a little more generally than needed in this section. We will use it again in our discussion of expected utility and risk aversion.

LEMMA A.3.2. *Suppose the continuous function $f : (\ell, \infty) \to \mathbb{R}$ is* not *concave and $p \in (0,1)$. Then there exist some $x^* \in (\ell, \infty)$ and small enough scalar $\varepsilon > 0$ such that for all $\delta \in (0, \varepsilon)$,*

$$f(x^*) < pf(x^* + (1-p)\delta) + (1-p) f(x^* - p\delta). \tag{A.3.1}$$

PROOF. Since f is not concave, there exist $x^0, x^1 \in (\ell, \infty)$ and $\alpha \in (0,1)$ such that $f(x^\alpha) < (1-\alpha) f(x^0) + \alpha f(x^1)$, where $x^\alpha \equiv (1-\alpha) x^0 + \alpha x^1$ and $x^0 < x^1$. Let $h : (\ell, \infty) \to \mathbb{R}$ be the affine function whose graph is the straight line through the points $(x^0, f(x^0))$ and $(x^1, f(x^1))$. Inequality (A.3.1) is satisfied by the given function f if and only if it is satisfied by the function $f - h$. Replacing f by $f - h$, we proceed under the assumption that $f(x^0) = f(x^1) = 0$ and therefore $f(x^\alpha) < 0$. Let $K = \operatorname{argmin}\{f(x) : x \in [x^0, x^1]\}$, a compact set (since f is continuous), and let $x^* = \min K$. Clearly, $x^* \in (x^0, x^1)$ and $f(x^*) < 0$. Choose $\varepsilon > 0$ small enough so that $x^* + (1-p)\varepsilon$ and $x^* - p\varepsilon$ are both valued in $[x^0, x^1]$. The facts that x^* minimizes f on $[x^0, x^1]$ and x^* is the least point in $[x^0, x^1]$ with this property respectively imply, for all $\delta \in (0, \varepsilon)$, the inequalities

$$f(x^*) \leq f(x^* + (1-p)\delta) \text{ and } f(x^*) < f(x^* - p\delta),$$

which in turn imply (A.3.1). □

LEMMA A.3.3. *Suppose that the functions $f, g : (\ell, \infty) \to \mathbb{R}$ are continuous and neither of them is concave. Then there exists some (x^*, y^*) such that the set*

$$\left\{(x,y) \in (\ell, \infty)^2 \mid f(x) + g(y) > f(x^*) + g(y^*)\right\}$$

is not convex (with the convention that the empty set is convex).

PROOF. Let x^* and ε be selected in terms of f as in Lemma A.3.2 with $p = 1/2$, and let y^* and ε be analogous quantities for g. (Note that we can choose the same ε in both cases.) Clearly, the validity of the lemma's claim does not change if f and g are modified by adding a constant. Replacing f with $f - f(x^*)$ and g with $g - g(y^*)$, we proceed under the assumption that

$$f(x^*) = g(y^*) = 0. \tag{A.3.2}$$

Letting $\epsilon \equiv \varepsilon/2$, condition (A.3.1) can be restated as

$$f(x^* + a) + f(x^* - a) > 0 \quad \text{for all } a \in (-\epsilon, \epsilon). \tag{A.3.3}$$

This implies that x^* is not a local maximizer of f. Therefore, the set

$$I_f = \{f(x^* + a) \mid a \in (-\epsilon, \epsilon)\}$$

is an interval (since f is continuous) that includes a subinterval of the form $[0, \delta)$ for some $\delta > 0$. Analogously, we have

$$g(y^* + b) + g(y^* - b) > 0 \quad \text{for all } b \in (-\epsilon, \epsilon), \tag{A.3.4}$$

and I_g includes a subinterval $[0, \delta)$ for some $\delta > 0$. It follows that $I_f \cap I_g$ is nonempty, which is to say that there exist a and b in $(-\epsilon, \epsilon)$ such that $f(x^* + a) = g(y^* + b)$. The last equality in combination with (A.3.3) and (A.3.4) implies

$$f(x^* - a) + g(y^* + b) > 0 \quad \text{and} \quad f(x^* + a) + g(y^* - b) > 0.$$

We have proved that $S \equiv \{(x, y) : f(x) + g(y) > 0\}$ contains the points $(x^* - a, y^* + b)$ and $(x^* + a, y^* - b)$, whose midpoint is (x^*, y^*). By assumption (A.3.2), $(x^*, y^*) \notin S$ and therefore S is not convex. □

A.4. Scale/translation invariant representations

In this section we show that a scale invariance property of an additive utility implies a unique power or logarithmic functional form, while a translation invariance property of an additive utility implies a unique exponential form.[4] As we will see in the following section, these results can be interpreted as providing ordinal foundations for expected utility with constant relative risk aversion in the case of scale-invariant preferences, and constant absolute risk aversion in the case of translation-invariant preferences.

The main argument relies on Theorem A.2.4 on the uniqueness of additive representations and the following technical result.[5]

LEMMA A.4.1. *Suppose that the function $f : \mathbb{R} \to \mathbb{R}$ satisfies*

$$f(x + y) = f(x) + f(y) \quad \textit{for all } x, y \in \mathbb{R}, \tag{A.4.1}$$

and there exists some nonempty open interval on which f is bounded. Then $f(x) = f(1)\, x$ for all $x \in \mathbb{R}$.

PROOF. We are to show that the function $\delta(x) \equiv f(x) - f(1)\, x$ vanishes on the entire real line. Note that $\delta(1) = 0$ and $\delta(x + y) = \delta(x) + \delta(y)$ for all $x, y \in \mathbb{R}$. Iterating the last equation, we have $\delta(1) = \sum_{i=1}^{n} \delta(1/n)$, and therefore $\delta(1/n) = 0$, for every positive integer n. Similarly, for positive integers m and n, we have $\delta(m/n) = \sum_{i=1}^{m} \delta(1/n) = 0$. Therefore, δ vanishes on the set of positive rational numbers. Since $\delta(0 + 0) = \delta(0) + \delta(0)$, it also vanishes at zero, and

[4] To my knowledge, the purely ordinal characterizations of Theorems A.4.3 and A.4.2, which do not assume differentiability of the utility function, first appear in Skiadas [2013b]. An earlier version in Skiadas [2009] assumes that the utility is continuously differentiable in some arbitrarily small neighborhood.

[5] For some historical context and related results, see Aczél [2006].

since $\delta(0) = \delta(r) + \delta(-r)$, it also vanishes on the set of negative rationals. The function δ inherits from f the property that it is bounded on some nonempty open interval (a, b). Given any $x \in \mathbb{R}$, we can find a rational r such that $x + r \in (a, b)$, and since $\delta(x) = \delta(x + r)$, it follows that δ is bounded on the entire real line. Finally, for any real x, the set of all $\delta(nx) = n\delta(x)$ as n ranges over the positives integers remains bounded only if $\delta(x) = 0$. □

The preference $\succ$ on $\mathbb{R}^N$ is said to be **translation invariant** if

$$x \succ y \text{ and } t \in \mathbb{R} \quad \Longrightarrow \quad x + t \succ y + t.$$

THEOREM A.4.2. *Suppose the preference $\succ$ on $\mathbb{R}^N$ admits an additive utility representation (Definition A.2.1). Then $\succ$ is translation invariant if and only if it admits a utility representation of the form*

$$U(x) = \sum\nolimits_{n=1}^{N} w_n \frac{1 - \exp(-\alpha x_n)}{\alpha}, \quad x \in \mathbb{R}^N, \tag{A.4.2}$$

for unique $\alpha \in \mathbb{R}$ and $w_1, \ldots, w_N \in (0, 1)$ such that $\sum_n w_n = 1$. The convention for $\alpha = 0$ is that $(1 - \exp(-\alpha x))/\alpha$ is equal to x, which is the limit as $\alpha \to 0$.

PROOF. Suppose U is an additive utility representing the translation invariant preference $\succ$. Without loss of generality, we assume that $U_n(0) = 0$ for all $n = 1, \ldots, N$ (otherwise, replace U_n by $U_n - U_n(0)$). For any $t \in \mathbb{R}$, the translation invariance of $\succ$ implies that $U(x + t)$ as a function of $x \in (0, \infty)^N$ defines another additive utility representation of $\succ$. Since, by Theorem A.2.4, an additive representation is unique up to a positive affine transformation and we assumed $U_n(0) = 0$, it follows that there exists a function $a : \mathbb{R} \to (0, \infty)$ such that

$$U_n(s + t) = U_n(s)\, a(t) + U_n(t), \quad s, t \in \mathbb{R}, \quad n = 1, \ldots, N. \tag{A.4.3}$$

If a is identically equal to one, then Lemma A.4.1 implies that $U_n(x) = w_n x$, $x \in \mathbb{R}$, for necessarily positive constants w_n, which can be assumed to add up to one after positively scaling the utility, resulting in representation (A.4.2) with $\alpha = 0$. Suppose instead that $a(y) \neq 1$ for some y. Equation (A.4.3) applied with $(s, t) = (x, y)$ and again with $(s, t) = (y, x)$ implies that

$$U_n(x) = w_n(1 - a(x)), \quad x \in \mathbb{R}, \tag{A.4.4}$$

where $w_n \equiv U_n(y)/(1 - a(y)) \neq 0$ (since U_n is increasing). Equation (A.4.3) with $s = 1$ implies that a is the difference of two increasing functions and therefore bounded on some open interval. Substituting expression (A.4.4) for U_n into equation (A.4.3) and simplifying, we find that $\log a$ satisfies the assumptions of Lemma A.4.1. Therefore $\log a(x) = -\alpha x$ for some scalar α, and equation (A.4.4) becomes the claimed representation (A.4.2) after positive scaling. The converse claim is immediate. □

The preference $\succ$ on $(0,\infty)^N$ is **scale invariant** if

$$x \succ y \text{ and } s \in (0,\infty) \quad \Longrightarrow \quad sx \succ sy.$$

THEOREM A.4.3. *Suppose the preference $\succ$ on $(0,\infty)^N$ admits an additive utility representation (Definition A.2.1). Then $\succ$ is scale invariant if and only if it admits a utility representation of the form*

$$U(x) = \sum\nolimits_{n=1}^{N} w_n \frac{x_n^{1-\gamma}-1}{1-\gamma}, \quad x \in (0,\infty)^N, \tag{A.4.5}$$

for unique $\gamma \in \mathbb{R}$ and $w_1,\ldots,w_N \in (0,1)$ such that $\sum_n w_n = 1$. The convention for $\gamma = 1$ is that $(x^{1-\gamma}-1)/(1-\gamma)$ is equal to $\log x$, which is the limit as $\gamma \to 1$.

PROOF. Using the notation $\exp x = (\exp x_1,\ldots,\exp x_N)$, define the preference $\succ^{\exp}$ on $\mathbb{R}^N$ by

$$x \succ^{\exp} y \quad \Longleftrightarrow \quad \exp x \succ \exp y.$$

Note that $\succ^{\exp}$ is translation invariant if and only if $\succ$ is scale invariant. The result follows from Theorem A.4.2 applied to $\succ^{\exp}$. □

A.5. Expected utility representations

The rest of this appendix is on expected utility, which is a special type of additive utility. We henceforth refer to $\{1,\ldots,N\}$ as the **state space** and regard every $x \in (\ell,\infty)^N$ as a (state-contingent) **payoff**. An agent expresses preferences over such payoffs, represented by the preference (relation) $\succ$, not knowing what state will be realized. A strictly positive **probability** is any vector $P = (P_1,\ldots,P_n) \in (0,1)^N$ such that $\sum_n P_n = 1$. An **expected utility** is a pair (P,u) of a strictly positive probability P and an increasing and continuous function $u : (\ell,\infty) \to \mathbb{R}$; it is said to **represent** the preference $\succ$ if $U \equiv \sum_{n=1}^{N} P_n u$ is an (additive) utility representation of $\succ$. Last section's scale or translation invariant additive utilities are examples of expected utility. As in those examples, the preference $\succ$ uniquely determines an expected utility representation up to a positive affine transformation.

THEOREM A.5.1. *If (P,u) and $(\tilde{P},\tilde{u})$ are expected utilities representing the same preference, then $\tilde{P} = P$ and there exist constants $a \in (0,\infty)$ and $b \in \mathbb{R}$ such that $\tilde{u} = au + b$.*

PROOF. By Theorem A.2.4, there exist $a \in (0,\infty)$ and $b_1,\ldots,b_n \in \mathbb{R}$ such that $\tilde{P}_n\tilde{u} = aP_nu + b_n$ for all n. Adding up over all n, we obtain $\tilde{u} = au + b$, where $b \equiv \sum_n b_n$. Therefore, $\tilde{P}_n(au+b) = aP_nu + b_n$ or $(\tilde{P}_n - P_n)au = b_n - \tilde{P}_n b$. The right-hand side of the last equation is a constant but the left-hand side is an increasing function, unless $\tilde{P}_n = P_n$. □

In the remainder of this section we formulate a property of $\succ$ we call state independence and we refine the additive representation Theorem A.2.3 (where $N > 2$) by showing that $\succ$ is both separable and state independent if and only if it admits an expected utility representation.[6] The reader not interested in the details of this claim can safely proceed to the following section. It is worth noting, however, that in this context, last section's results show that state independence is implied by scale or translation invariance.

We formulate the state-independence condition in terms of the **indifference relation** $\sim$ associated with the preference $\succ$, defined by

$$x \sim y \quad \iff \quad (\text{not } x \succ y) \text{ and } (\text{not } y \succ x)\,.$$

We write 1^n for the random variable that is the indicator of $\{n\}$, that is, $1^n_n = 1$ and $1^n_k = 0$ for all $k \neq n$. As usual, we identify a scalar $w \in (\ell, \infty)$ and the constant payoff $(w, w, \ldots, w)$.

Fixing any distinct states $m, n \in \{1, \ldots, N\}$, suppose the agent is indifferent between the constant payoff w and the same payoff but with the contingent value at state m increased by $z \in (0, \infty)$ and the contingent value at state n decreased by $y \in (0, w - \ell)$. We express this indifference with the notation

$$(z; m_+) =_w (y; n_-) \iff w \sim w + z1^m - y1^n.$$

We express the agent's indifference between increasing the constant payoff w by $z \in (0, \infty)$ at state m or by $x \in (0, \infty)$ at state n as

$$(z; m_+) =_w (x; n_+) \iff w + z1^m \sim w + x1^n.$$

We write $(x; n_+) =^m_w (y; n_-)$ if there exists some $z \in (0, \infty)$ such that both $(z; m_+) =_w (y; n_-)$ and $(z; m_+) =_w (x; n_+)$, which provides a sense in which an increase by x at state n is of the same utility magnitude as a decrease by y at state n. We write $(x; n_+) \neq^m_w (y; n_-)$ if there exists some $z \in (0, \infty)$ such that $(z; m_+) =_w (y; n_-)$ but it is *not* the case that $(z; m_+) =_w (x; n_+)$. We call $\succ$ **state independent** if for all $w \in (\ell, \infty)$ and distinct states $m, n, n' \in \{1, \ldots, N\}$, there exist *no*[7] $x \in (0, \infty)$ and $y \in (0, w - \ell)$ such that $\left(x; n'_+\right) =^m_w \left(y; n'_-\right)$ but $(x; n_+) \neq^m_w (y; n_-)$.

[6] Preferences over probability distributions that admit an expected-utility representation were first characterized by von Neumann and Morgenstern [1944]. Incorporating a subjective view of probability in the tradition of Ramsey [1926], Savage [1954] developed an axiomatic foundation for expected utility with a nonatomic probability that is uniquely determined by preferences over mappings from states to a set of consequences. Anscombe and Aumann [1963] offered an alternative foundation for expected utility with subjective beliefs that utilizes objective probabilities to calibrate subjective beliefs. The approach presented here is a variant of that in Wakker [1984, 1988, 1989] and is based on Skiadas [1997, 2009].

[7] Theorem A.5.2 and its proof remain valid if we define state independence to mean that for all distinct $m, n \in \{2, \ldots, N\}$, there exists $\epsilon > 0$ such that for all $x, y \in (0, \epsilon)$, it is not the case that $(x; 1_+) =^m_w (y; 1_-)$ and $(x; n_+) \neq^m_w (y; n_-)$.

THEOREM A.5.2. *Suppose that $N > 2$ and $\succ$ is a preference that admits a utility representation. Then $\succ$ admits an expected utility representation if and only if it is both separable and state independent.*

PROOF. The "only if" part is straightforward. Conversely, suppose $\succ$ is separable and state independent. By Theorem A.2.3, $\succ$ is represented by an additive utility $U = \sum_{n=1}^{N} U_n$. For every state n, define the open interval $I_n \equiv \{U_n(x) \mid x \in (\ell, \infty)\}$ and the function $f_n : I_1 \mapsto I_n$ such that $U_n(x) = f_n(U_1(x))$ for all $x \in (\ell, \infty)$. We will show that each f_n is affine, that is, there exist $a_n \in (0, \infty)$ and $b_n \in \mathbb{R}$ such that $f_n(U_1) = a_n U_1 + b_n$. An expected utility representation (P, u) results by letting $u = U_1$ and $P_n = a_n / \sum_n a_n$.

Note that every f_n is increasing and surjective, and therefore continuous. To show that f_n is affine, we show that for all $\alpha \in I_1$, there exists $\varepsilon > 0$ such that for all $\delta \in (0, \varepsilon)$,

$$f_n(\alpha + \delta) - f_n(\alpha) = f_n(\alpha) - f_n(\alpha - \delta). \tag{A.5.1}$$

By Lemma A.3.2 with $p = 1/2$, this condition implies that both f_n and $-f_n$ are concave, and therefore f_n is affine. Fix any $\alpha \in I_1$ and let $w \in (\ell, \infty)$ be defined by $U_1(w) = \alpha$. Consider any distinct states $m, n \in \{2, \ldots, N\}$. By the utility continuity, there exists a sufficiently small $\varepsilon > 0$ such that for all $\delta \in (0, \varepsilon)$ there exist positive scalars x, y, z, z' that solve the equations

$$\begin{gathered} \delta = U_1(w + x) - U_1(w) = U_1(w) - U_1(w - y), \\ \delta = U_m(w + z') - U_m(w), \\ U_m(w + z) - U_m(w) = U_n(w) - U_n(w - y). \end{gathered}$$

The first three equalities imply that $(x; 1_+) =_w^m (y; 1_-)$, and the last one that $(z; m_+) =_w (y; n_-)$. The state independence assumption then requires that $(z; m_+) =_w (x; n_+)$, and therefore $U_n(w + x) - U_n(w) = U_m(w + z) - U_m(w)$. This proves that $U_n(w + x) - U_n(w) = U_n(w) - U_n(w - y)$, which in turn implies (A.5.1). □

A.6. Expected utility and risk aversion

Concavity of an expected utility corresponds to risk aversion and the more concave the utility is the more risk averse the represented preference. In this section, we make precise and prove this claim. As is customary in the related literature, we use the term "more risk averse" to mean "no less risk averse," and "risk averse" to mean "risk averse or risk-neutral."

Consider a preference $\succ$ on $(\ell, \infty)^N$ with expected utility representation (P, u). The corresponding **certainty equivalent (CE)** v is the unique utility representation of $\succ$ satisfying $v(w) = w$ for all $w \in (\ell, \infty)$. Let $\mathbb{E}$ denote the expectation operator under P, defined by $\mathbb{E}z \equiv \sum_{n=1}^{N} P_n z_n$, $z \in \mathbb{R}^N$. The CE $v : (\ell, \infty)^N \to (\ell, \infty)$ is necessarily

given as $v = u^{-1}\mathbb{E}u$, meaning $v(x) = u^{-1}(\mathbb{E}u(x))$ for all $x \in (\ell, \infty)^N$. The agent is indifferent between x and a constant payoff $v(x)$. For given probability P, a lower CE value $v(x)$ represents higher risk aversion, since, given the same odds, the agent is willing to settle for a lower sure payoff in place of x. An expected utility CE $\tilde{v} = \tilde{u}^{-1}\mathbb{E}\tilde{u}$ is therefore **more risk averse** than v if $\tilde{v}(x) \leq v(x)$ for all $x \in (\ell, \infty)^N$, a condition we abbreviate to $\tilde{v} \leq v$.

THEOREM A.6.1. *Suppose (P, u) and $(P, \tilde{u})$ are expected utilities, with corresponding CEs $v \equiv u^{-1}\mathbb{E}u$ and $\tilde{v} \equiv \tilde{u}^{-1}\mathbb{E}\tilde{u}$. Then $\tilde{v} \leq v$ if and only if $\tilde{u} = f \circ u$ for a concave function $f : u(\ell, \infty) \to \mathbb{R}$.*

PROOF. Define the (increasing) function $f : u(\ell, \infty) \to \mathbb{R}$ such that $\tilde{u} = f \circ u$. If f is concave, then $\tilde{v} \leq v$ by Jensen's inequality:

$$\tilde{u}(v(x)) = f(\mathbb{E}u(x)) \geq \mathbb{E}f \circ u(x) = \tilde{u}(\tilde{v}(x)), \quad x \in (\ell, \infty)^N.$$

Conversely, we assume that $\tilde{v} \leq v$ and f is *not* concave and show a contradiction. Let $p \equiv P_1 > 0$. By Lemma A.3.2, there exists $w \in (\ell, \infty)$ and $\varepsilon > 0$ such that for all $\delta \in (0, \varepsilon)$,

$$f(u(w)) < pf(u(w) + (1-p)\delta) + (1-p)f(u(w) - p\delta). \tag{A.6.1}$$

Pick $\delta > 0$ small enough so that $x \in (\ell, \infty)^N$ is well-defined by

$$u(x_1) = u(w) + (1-p)\delta, \quad u(x_n) = u(w) - p\delta, \quad n \geq 2,$$

which implies that $w = v(x)$, in contradiction to inequality (A.6.1), which states that $\tilde{u}(w) < \mathbb{E}\tilde{u}(x)$ and therefore $w < \tilde{v}(x) \leq v(x)$. □

COROLLARY A.6.2. *For every expected utility (P, u), u is concave if and only if $u^{-1}\mathbb{E}u(x) \leq \mathbb{E}x$ for all $x \in (0, \infty)^N$.*

The corollary's condition states that v is more risk averse than the risk-neutral CE $\mathbb{E}$; it provides an absolute sense in which we can call $\succ$ risk averse. Another reasonable definition of absolute risk aversion is the convexity of the preference $\succ$, which can be thought of as preference for diversification. By Theorem A.3.1, $\succ$ is convex if and only if u is concave, and therefore the two notions of absolute risk aversion coincide. A third notion of absolute risk aversion, which we adopt as a definition, states that whenever $\succ$ rejects adding a risky payoff to a sure payoff, it also rejects adding every scaled-up version of the same risky payoff. We therefore define $\succ$ to be **risk averse** if for all $w \in (\ell, \infty)$ and $x \in (\ell - w, \infty)^N$,

$$w \succ w + x \text{ and } \theta \in (1, \infty) \implies w \succ w + \theta x. \tag{A.6.2}$$

This condition gives a sense of aversion to a specific risk x. Implicit in expected utility is a sense in which risk aversion is risk-source independent,[8] which leads to the equivalence of risk aversion to the concavity

[8]For a specification of scale-invariant preferences with source-dependent risk aversion, see Skiadas [2013b, 2015].

of u, and hence the equivalence of all three notions of absolute risk aversion just introduced: risk aversion, preferences for diversification, and more risk averse than risk neutral.

THEOREM A.6.3. *A preference with expected utility representation (P, u) is risk averse if and only if u is concave.*

PROOF. Suppose u is concave, $w \in (\ell, \infty)$ and $\theta \in (1, \infty)$. Since $w + x$ is a convex combination of w and $w + \theta x$, if $\mathbb{E}u(w + \theta x) \geq u(w)$ then $\mathbb{E}u(w + x) \geq u(w)$, which is the contrapositive of condition (A.6.2) and therefore proves that $\succ$ is risk averse.

Conversely, suppose $\succ$ is risk averse. We show that u is concave assuming $N = 2$, which entails no loss in generality (why?). We do so by confirming that $u^{-1}\mathbb{E}u \leq \mathbb{E}$ and applying Corollary A.6.2. Suppose instead that for some $x \in (\ell, \infty)^2$, $u^{-1}\mathbb{E}u(x) > \mathbb{E}x$. Define $\mu \equiv \mathbb{E}x$ and $\hat{x} \equiv x - \mu$, and therefore

$$\mathbb{E}u(\mu + \hat{x}) > u(\mu) \quad \text{and} \quad \mathbb{E}\hat{x} = 0. \tag{A.6.3}$$

We claim that this condition implies that u is convex on $(\mu - \epsilon, \mu + \epsilon)$ for some $\epsilon > 0$. Suppose not. Then, by Lemma A.3.2 applied to $-u$, for all $n = 1, 2, \ldots$, there exist $\mu_n \in (\mu - 1/n, \mu + 1/n)$ and $\delta_n \in (0, 1)$ such that $u(\mu_n) > \mathbb{E}u(\mu_n + \delta_n \hat{x})$, and $u(\mu_n) > \mathbb{E}u(\mu_n + \hat{x})$ by risk aversion. Since u is continuous and $\lim_{n\to\infty} \mu_n = \mu$, $u(\mu) \geq \mathbb{E}u(\mu + \hat{x})$ in contradiction to (A.6.3). We can therefore select $\epsilon > 0$ such that u is convex on $(\mu - \epsilon, \mu + \epsilon)$. Since u is also (strictly) increasing, it follows[9] that the derivative u' exists and is positive at all but at most countably many points of $(\mu - \epsilon, \mu + \epsilon)$. Since u is continuous, in condition (A.6.3), we can slightly perturb μ if necessary to some value w such that $u'(w)$ exists and is positive, while it is still the case that $\mathbb{E}u(w + \hat{x}) > u(\mu)$. We can then slightly decrease $\hat{x}$ to $\tilde{x}$ so that $\mathbb{E}u(w + \tilde{x}) \geq u(w)$, $\mathbb{E}\tilde{x} < 0$ and $\tilde{x}_n \neq 0$ for every state n. Risk aversion implies that for all $\delta \in (0, 1)$, $\mathbb{E}u(w + \delta\tilde{x}) \geq u(w)$ and therefore

$$\mathbb{E}\left[\frac{u(w + \delta\tilde{x}) - u(w)}{\delta\tilde{x}}\tilde{x}\right] \geq 0.$$

Letting $\delta \downarrow 0$ gives $u'(w)\,\mathbb{E}\tilde{x} \geq 0$, which contradicts the fact that $u'(w) > 0$ and $\mathbb{E}\tilde{x} < 0$ by construction. □

[9] Suppose u is strictly increasing and convex on an interval (a, b) and $w \in (a, b)$. For all small enough $\delta > 0$, define the ordered slopes

$$\Delta_+(\delta) \equiv \frac{u(w + \delta) - u(w)}{\delta} \geq \frac{u(w) - u(w - \delta)}{\delta} \equiv \Delta_-(\delta).$$

As $\delta \downarrow 0$, $\Delta_+(\delta)$ monotonically decreases to a limit defining the right derivative $u'_+(w)$, and $\Delta_-(\delta)$ monotonically increases to a limit defining the left derivative $u'_-(w)$. Clearly, $u'_+(w) > 0$ and $u'_+(w) \geq u'_-(w)$. The intervals $\left(u'_-(w), u'_+(w)\right)$ are non-overlapping as w ranges over (a, b) and therefor at most countably many of them are nonempty. This proves that for all except at most countably many values of $w \in (a, b)$, $u'(w) = u'_-(w) = u'_+(w) > 0$.

Consider an expected utility (P,u), where u belongs to the set C^2 of twice continuously differentiable real-valued functions on the real line. The scale-or-translation invariant cases of Section 3.8 are examples. More generally, every expected utility is "close" to a smooth version in the sense of the following remark (which is not used anywhere else).

REMARK A.6.4. Suppose that a payoff $x \in (\ell,\infty)^N$ is perceived by the agent as $x+\varepsilon$, where ε is an arbitrarily small noise term that is stochastically independent of x and continuously distributed over a compact interval with density $f_\varepsilon \in C^2$. The expected utility of x can be thought of as the reduced form of the expected utility of $x+\varepsilon$:

$$\mathbb{E}u(x) \equiv \mathbb{E}u_\varepsilon(x+\varepsilon) \equiv \sum_{n=1}^{N} P_n \int_{\mathbb{R}} u_\varepsilon(x_n+y)\, f_\varepsilon(y)\, dy.$$

We can therefore define the function $u:(\ell,\infty)\to\mathbb{R}$ as

$$u(z) \equiv \int_{\mathbb{R}} u_\varepsilon(z+y)\, f_\varepsilon(y)\, dy = \int_{\mathbb{R}} u_\varepsilon(x)\, f_\varepsilon(x-z)\, dx.$$

For continuous but otherwise arbitrary u_ε, the fact that $f_\varepsilon \in C^2$ has a compact support implies that $u \in C^2$. $\diamondsuit$

The **Arrow-Pratt coefficient of absolute risk aversion** associated with $u \in C^2$ is the function

$$A^u \equiv -\frac{u''}{u'}. \tag{A.6.4}$$

Section 3.8 presents our central motivation for introducing this function, which is the approximation of the expected utility of small Brownian risks. Just as the CE $u^{-1}\mathbb{E}u$, A^u is invariant to positive affine transformations of u and determines u up to a positive affine transformation. (To see the latter claim, write (A.6.4) as $A^u(x)\,dx = -d\log u'(x)$ and integrate twice.) Given P, the function A^u is therefore uniquely associated with the corresponding CE $u^{-1}\mathbb{E}u$ and hence the preference relation represented by (P,u). Note that the TI formulation of Theorem A.6.4 corresponds to the constant absolute risk aversion assumption $A^u(x) \equiv \alpha$, $x \in \mathbb{R}$, and the SI formulation of Theorem A.6.4 corresponds to the constant relative risk aversion assumption $xA^u(x) \equiv \gamma$, $x \in (0,\infty)$. The function $x \mapsto xA^u(x)$ is known as the **coefficient of relative risk aversion**.

By Theorem A.6.4, for $u \in C^2$, the expected utility (P,u) is risk averse if and only if $u'' \le 0$ if and only if $A^u \ge 0$. The inequalities are to be interpreted as holding on the entire domain, for example, $A^u \ge 0$ means $A^u(x) \ge 0$ for all $x \in (\ell,\infty)$. By Theorem A.6.4, for $u, \tilde{u} \in C^2$, the expected utility $(P,\tilde{u})$ is more risk averse than the expected utility (P,u) (in the sense that $\tilde{u}^{-1}\mathbb{E}\tilde{u} \le u^{-1}\mathbb{E}u$) if and only if $A^{\tilde{u}} \ge A^u$. To confirm this claim, note that if $\tilde{u} = f \circ u$ then $A^{\tilde{u}} = A^u + \left(A^f \circ u\right) u'$ and f is concave if and only if $A^f \ge 0$.

APPENDIX B

Elements of Convex Analysis

This appendix reviews some basic linear algebra and convex analysis concepts. The emphasis is on the finite-dimensional case, but from a perspective that is amenable to infinite-dimensional extensions.[1]

B.1. Inner product space

The analysis of this appendix is all carried out relative to an inner product space, which is a vector space together with an inner product. In this section we define these terms and discuss some of their basic properties.

All vector spaces in this text are defined relative to the field $\mathbb{R}$ of the real numbers, also referred to as **scalars**. A **vector** or **linear** space is a triple of a set X, whose elements are called **vectors** or **points**, a **scaling** operation assigning a vector αx to each $(\alpha, x) \in \mathbb{R} \times X$, and an **addition** operation assigning a vector $x+y$ to each $(x, y) \in X \times X$, provided the following restrictions hold for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{R}$: $(\alpha + \beta)\, x = \alpha x + \beta x$ and $\alpha\,(x + y) = \alpha x + \alpha y$; $(\alpha\beta)\, x = \alpha\,(\beta x)$ and $1x = x$; $x+y = y+x$ and $(x + y)+z = x+(y + z)$; there is a (necessarily unique) vector 0, called the **zero vector**, such that $0 + x = x$ for all x, and for each vector x there is a (necessarily unique) vector $-x$ such that $x + (-x) = 0$. The difference between two vectors is defined as $x - y \equiv x + (-y)$. (As in the main text, $\equiv$ means equal by definition.)

It is customary to refer to a vector space X, while it is implied that scaling and addition operations are also specified along with X. Two vector spaces $X$and $\tilde{X}$ are **isomorphic** if there is a bijection (one-to-one and onto mapping) $\phi : X \to \tilde{X}$ that respects the addition and scaling operations: $\phi\,(x + y) = \phi\,(x) + \phi\,(y)$ and $\phi\,(sx) = s\phi\,(x)$ for all $x, y \in X$ and $s \in \mathbb{R}$. Note that addition and scaling on the left-hand side of these equations refer to operations in X and those on the right-hand side refer to operations in $\tilde{X}$. One can think of $\tilde{X}$ as being the same vector space as X, but with each vector x given a unique new label $\phi\,(x)$. We refer to ϕ as a **vector space isomorphism.**

[1] Treatments of finite-dimensional convex optimization theory include Rockafellar [1970], Bertsekas [2003] and Boyd and Vandenberghe [2004], the latter providing an introduction to modern computational optimization algorithms. Infinite-dimensional extensions can be found in Bonnans and Shapiro [2000], Dunford and Schwartz [1988], Ekeland and Témam [1999] and Luenberger [1969].

A canonical example of a vector space is **d-dimensional Euclidean space**, for positive integer d, defined as the Cartesian product $\mathbb{R}^d$ with coordinate-wise scaling and addition operations: $(\alpha x + y)_i = \alpha x_i + y_i$ for every $i \in 1, \ldots, d$, $\alpha \in \mathbb{R}$ and $x, y \in \mathbb{R}^d$. The set L of all random variables on some state space Ω is another example of a vector space, where a random variable is defined as a real-valued function on Ω and scaling and addition are defined state-wise: $(\alpha x + y)(\omega) = \alpha x(\omega) + y(\omega)$ for all $\omega \in \Omega$, $\alpha \in \mathbb{R}$ and $x, y \in L$. If $\Omega = \{\omega_1, \ldots, \omega_d\}$, then L is isomorphic to $\mathbb{R}^d$.

We henceforth fix a reference vector space X. Unless otherwise specified, a vector is an element of X. A **linear subspace** of X is any subset Y of X that is closed with respect to scaling and addition: $\alpha x + y \in Y$ for all $\alpha \in \mathbb{R}$ and $x, y \in Y$. Equivalently, a linear subspace of X is a subset of X that is a vector space in its own right, with the suitably restricted addition and scaling operations inherited from X.

A **linear combination** of the vectors $x_1, \ldots, x_n$ is any vector of the form $\alpha_1 x_1 + \cdots + \alpha_n x_n$ for scalar α_i. (The term **linear combination** will always refer to the linear combination of finitely many vectors.) If all the α_i are zero, we call the linear combination **trivial.** The **linear span** of a set of vectors S, denoted $\operatorname{span}(S)$, is the intersection of all linear subspaces that include S, which is the same as the set of all linear combinations of elements of S. We say that $\operatorname{span}(S)$ is the linear subspace **generated** or **spanned** by S or the elements of S. If $S = \{x_1, \ldots, x_n\}$, we also write $\operatorname{span}(x_1, \ldots, x_n)$ for $\operatorname{span}(S)$.

A set of vectors S is **linearly independent** if every linear combination of elements of S that is also in S is trivial, a condition that is equivalent to the unique representation of each element of $\operatorname{span}(S)$ as a linear combination of elements of S, and is also equivalent to the nonexistence of an element x of S that can be expressed as a linear combination of elements in $S \setminus \{x\}$.

A **basis** of X is a linearly independent set of vectors that spans X. A vector space is **finite-dimensional** if it has a finite basis and **infinite-dimensional** otherwise. Every basis of a finite-dimensional vector space has the same number of elements, called the space's **dimension.** The vector space $\{0\}$ has, by definition, dimension zero.

If X has finite dimension d, we represent a basis $\{B_1, \ldots, B_d\}$ of X as a column matrix $B \equiv (B_1, \ldots, B_d)'$ and we write σ^x for the row vector in $\mathbb{R}^d$ that **represents** the point x in X relative to the given basis B :

$$x = \sigma^x B \equiv \sum_{i=1}^{d} \sigma_i^x B_i.$$

The mapping $x \mapsto \sigma^x$ defines a vector space isomorphism from X to $\mathbb{R}^d$, which shows that every finite-dimensional vector space is isomorphic to a Euclidean space.

A **functional** is a function of the form $f : X \to \mathbb{R}$. The functional f is **linear** if $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$ for all $x, y \in X$ and $\alpha, \beta \in \mathbb{R}$. Given a finite basis $B = (B_1, \ldots, B_d)'$ and a linear functional f, we use the notation $f(B) \equiv (f(B_1), \ldots, f(B_d))'$. The single vector $f(B)$ in $\mathbb{R}^d$ determines the entire function f, since $x = \sigma^x B$ implies $f(x) = \sigma^x f(B) = \sigma^x \cdot f(B)$, where the dot of the last expression denotes the **Euclidean inner product:**

$$x \cdot y = \sum_{i=1}^{d} x_i y_i, \quad x, y \in \mathbb{R}^d. \tag{B.1.1}$$

The mapping $(x, y) \mapsto x \cdot y$ is bilinear (linear in each of x and y keeping the other fixed), symmetric ($x \cdot y = y \cdot x$), and positive definite ($x \cdot x \geq 0$, with equality holding if and only if $x = 0$). Abstracting away these properties, we define a (real) **inner product** $\langle \cdot \mid \cdot \rangle$ on the vector space X as a mapping that assigns to each $(x, y) \in X \times X$ a real number $\langle x \mid y \rangle$ such that for all $x, y, z \in X$ and $\alpha \in \mathbb{R}$,

- $\langle \alpha x + y \mid z \rangle = \alpha \langle x \mid z \rangle + \langle y \mid z \rangle$.
- $\langle x \mid y \rangle = \langle y \mid x \rangle$.
- $\langle x \mid x \rangle \geq 0$, with equality holding if and only if $x = 0$.

An **inner product space** is a vector space together with an inner product on this space. Two inner product spaces X and $\tilde{X}$ (each with an implied inner product) are said to be **isomorphic** if there is a vector space isomorphism $\phi : X \to \tilde{X}$ that preserves inner products: $\langle x \mid y \rangle = \langle \phi(x) \mid \phi(y) \rangle$, where the left-hand side refers to the X inner product and the right-hand side refers to the $\tilde{X}$ inner product.

Example B.1.1. Every symmetric positive definite matrix $Q \in \mathbb{R}^{d \times d}$ defines an inner product on $\mathbb{R}^d$ by letting $\langle x \mid y \rangle = xQy'$, where $x, y \in \mathbb{R}^d$ are treated as row vectors. The Euclidean inner product is obtained if Q is the identity matrix. $\diamondsuit$

Example B.1.2. Suppose X is any finite-dimensional vector space and B is any basis of X. Then an inner product is defined by letting $\langle x \mid y \rangle = \sigma^x \cdot \sigma^y$, where $x = \sigma^x B$ and $y = \sigma^y B$. Later, we will establish the more interesting fact that given any inner product on X, we can select the basis B so that this example's representation is valid. $\diamondsuit$

Example B.1.3. Let $\mathbb{E}$ denote the expectation operator relative to a probability on a finite state space Ω, where every state is assigned a nonzero probability. On the vector space L of all random variables on Ω, the inner product $\langle x \mid y \rangle = \mathbb{E}[xy]$ renders L an inner product space that is isomorphic to a Euclidean inner product space (why?). Note that for zero-mean $x, y \in L$, $\langle x \mid y \rangle = \operatorname{cov}[x, y]$, but covariance is not an inner product on all of L (why?). $\diamondsuit$

The **norm induced** by the inner product $\langle \cdot \mid \cdot \rangle$ is the function $\|\cdot\| : X \to \mathbb{R}$ defined by

$$\|x\| = \sqrt{\langle x \mid x \rangle}, \quad x \in X. \tag{B.1.2}$$

Here $\|x\|$ denotes the **norm of** x, which is the value the norm $\|\cdot\|$ assigns to x. Note that an inner-product isomorphism also preserves norms, a property known as **isometry**. Conversely, every isometry also preserves inner products, since

$$\|x + y\|^2 = \|x\|^2 + 2\langle x \mid y \rangle + \|y\|^2. \tag{B.1.3}$$

Vectors x and y are said to be **orthogonal** if $\langle x \mid y \rangle = 0$. Inspection of identity (B.1.3) shows that the vectors x and y are orthogonal if and only if they satisfy the **Pythagorean identity**

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 .$$

We have so far treated inner products purely algebraically. For geometric intuition, we visualize $\|x\|$ as the length of the vector x, and the orthogonality condition $\langle x \mid y \rangle = 0$ as x and y forming a right angle. For any vector x, we write $\hat{x} \equiv \|x\|^{-1}x$ for the unique positively scaled version of x whose norm is one. The inner product $\langle \hat{x} \mid y \rangle$ is the value of the scalar α that minimizes $\|y - \alpha\hat{x}\|$, a condition that defines $\bar{y} \equiv \langle \hat{x} \mid y \rangle \hat{x}$ as the (orthogonal) **projection** of the point y onto $\text{span}\,(x) = \text{span}\,(\hat{x})$. One can visualize $\bar{y}$ as the point of the line $\text{span}\,(x)$ that is closest to the point y. To justify this claim, use identity (B.1.3) and the fact that $\|\hat{x}\| = 1$ to compute

$$\|y - \alpha\hat{x}\|^2 = \|y\|^2 - \|\bar{y}\|^2 + (\alpha - \langle \hat{x} \mid y \rangle)^2 .$$

The quadratic is clearly minimized for $\alpha = \langle \hat{x} \mid y \rangle$ and the corresponding minimum is equal to $\|y - \bar{y}\|^2 = \|y\|^2 - \|\bar{y}\|^2$. The latter is a Pythagorean identity that is equivalent to the orthogonality of $y - \bar{y}$ to $\bar{y}$ and therefore to x.

The very useful Cauchy-Schwarz inequality results from the simple observation that the nonzero vectors $\hat{x}$, $\hat{y}$ and $\langle \hat{x} \mid \hat{y} \rangle \hat{x}$ form an orthogonal triangle whose hypotenuse $\hat{y}$ has unit length, and therefore each of its other sides has length less than one. In particular, $|\langle \hat{x} \mid \hat{y} \rangle| \leq 1$. The following proposition elaborates on this conclusion. Two vectors x and y are said to be **collinear** if either $x = \alpha y$ or $y = \alpha x$ for some $\alpha \in \mathbb{R}$.

PROPOSITION B.1.4. *All vectors x, y satisfy the* ***Cauchy-Schwarz inequality***

$$|\langle x \mid y \rangle| \leq \|x\|\|y\|,$$

and the ***triangle inequality***

$$\|x + y\| \leq \|x\| + \|y\| . \tag{B.1.4}$$

The Cauchy-Schwarz inequality holds as an equality if and only if x and y are collinear.

PROOF. Omitting trivial cases, we assume x and y are nonzero. Direct calculation using the bilinearity of an inner product and the fact that $\langle \hat{y} \mid \hat{y} \rangle = 1$ shows that $\hat{x} - \langle \hat{x} \mid \hat{y} \rangle \hat{y}$ is orthogonal to $\hat{y}$. Using the positive definiteness of inner products and the Pythagorean identity, we conclude that $0 \leq \|\hat{x} - \langle \hat{x} \mid \hat{y} \rangle \hat{y}\|^2 = 1 - \langle \hat{x} \mid \hat{y} \rangle^2$, with equality holding if and only if $\hat{x} = \langle \hat{x} \mid \hat{y} \rangle \hat{y}$ if and only if x and y are collinear. We showed $\langle \hat{x} \mid \hat{y} \rangle^2 \leq 1$, which is equivalent to the Cauchy-Schwarz inequality. Identity (B.1.3) and the Cauchy-Schwarz inequality imply the triangle inequality. □

EXAMPLE B.1.5. In Example B.1.3, consider the covariance inner product on the vector space $\{x \in L \mid \mathbb{E}x = 0\}$. The norm of x is its standard deviation and the Cauchy-Schwarz inequality corresponds to the fact that the correlation coefficient of two random variables lies in the interval $[-1, 1]$. ♢

We henceforth take as given the inner product $\langle \cdot \mid \cdot \rangle$, rendering X an inner product space. Every $x \in X$ defines a corresponding linear functional $f(y) = \langle x \mid y \rangle$, $y \in X$, in which case, we say that x is the (necessarily unique) **Riesz representation** of f.

In the finite-dimensional case, the inner product and Riesz representations can be expressed concretely in terms of a linear basis. To simplify notation, we extend the inner products to matrices of vectors, using the usual matrix addition and multiplication rules. In particular, given a column matrix of vectors $B = (B_1, \ldots, B_d)'$ and any vector x, we write $\langle x \mid B' \rangle \equiv (\langle x \mid B_1 \rangle, \ldots, \langle x \mid B_d \rangle)$, and $\langle B \mid B' \rangle$ for the **Gram matrix** of B, defined as the $d \times d$ matrix whose (i, j) entry is $\langle B_i \mid B_j \rangle$.

PROPOSITION B.1.6. *Suppose $B = (B_1, \ldots, B_d)'$ is a basis of X and for every $x \in X$, the row vector σ^x is defined by $x = \sigma^x B$. Then the Gram matrix $\langle B \mid B' \rangle$ is symmetric and positive definite and the inner product in X can be expressed as*

$$\langle x \mid y \rangle = \sigma^x \langle B \mid B' \rangle \sigma^{y\prime}, \quad x, y \in X. \tag{B.1.5}$$

The Riesz representation of every linear functional f exists and is given by $f(B)' \langle B \mid B' \rangle^{-1} B$, and therefore $\sigma^x = \langle x \mid B' \rangle \langle B \mid B' \rangle^{-1}$ for all $x \in X$.

PROOF. Bilinearity of the inner product implies the inner product representation (B.1.5). By the symmetry of the inner product, $\langle B \mid B' \rangle$ is a symmetric matrix. By the positive definiteness of the inner product, for every row vector $\alpha \in \mathbb{R}^d$, $\alpha \langle B \mid B' \rangle \alpha' = \langle \alpha B \mid \alpha B \rangle \geq 0$, with equality holding if and only if $\alpha B = 0$ if and only if $\alpha = 0$ (since B is a basis). This proves that $\langle B \mid B' \rangle$ is a positive definite matrix (and therefore invertible). If f is a linear function, then $f(y) = f(B)' \sigma^{y\prime}$ and $\langle x \mid y \rangle = \sigma^x \langle B \mid B' \rangle \sigma^{y\prime}$. Therefore, x is the Riesz representation of f if and only if $\sigma^x = f(B)' \langle B \mid B' \rangle^{-1}$. In this case, $\sigma^x = \langle x \mid B' \rangle \langle B \mid B' \rangle^{-1}$, since $f(B_i) = \langle x \mid B_i \rangle$ for all i. □

A corollary is that every d-dimensional inner product space is isomorphic to the inner product space of Example B.1.1. The isomorphism is the mapping $x \mapsto \sigma^x$, and the matrix Q of Example B.1.1 is the Gram matrix $\langle B \mid B' \rangle$. The basis B is said to be **orthonormal** if the corresponding Gram matrix is an identity matrix, which means the basis elements all have unit norm and are pairwise orthogonal. The last paragraph of Section B.4 shows that every finite-dimensional inner product space admits an orthonormal basis and is therefore isomorphic to a Euclidean inner-product space.

B.2. Basic topological concepts

Throughout this section X is an inner product space with induced norm $\|\cdot\|$, which we use to define convergence and related basic topological concepts. The function $\|\cdot\| : X \to \mathbb{R}$ is positively homogeneous: $\|\alpha x\| = |\alpha| \, \|x\|$ for all $\alpha \in \mathbb{R}$ and $x \in X$; positive definite: $\|x\| \geq 0$, with equality holding if and only if $x = 0$; and satisfies the triangle inequality (B.1.4), which can be equivalently stated as

$$\big| \|x\| - \|y\| \big| \leq \|x - y\| \quad \text{for all } x, y \in X. \tag{B.2.1}$$

These three properties more broadly define what is known as a **norm** on a vector space, which may or may not be induced by an inner product. For every $r \in \mathbb{R}_+$ and $x \in X$, the **open ball** with center x and radius r is the set

$$B(x; r) \equiv \{ y \in X \mid \|y - x\| < r \}.$$

A sequence $\{x_n\} = (x_1, x_2, \dots)$ of points in X **converges** to the **limit** $x \in X$ if for every $\varepsilon > 0$, there exists an integer N_ε such that $n > N_\varepsilon$ implies $x_n \in B(x; \varepsilon)$. In this case, the sequence $\{x_n\}$ is said to be **convergent.** A function $f : D \to \mathbb{R}$, where $D \subseteq X$, is **continuous at** $x \in D$ if for any sequence $\{x_n\}$ in D converging to x, the sequence $\{f(x_n)\}$ converges to $f(x)$. Note that $f : D \to \mathbb{R}$ is continuous at $x \in D$ if and only if for all $\varepsilon > 0$, there exists some $\delta > 0$ (depending on ε) such that $y \in D \cap B(x; \delta)$ implies $|f(y) - f(x)| < \varepsilon$. The function f is **continuous** if it is continuous at every point of its domain D.

Inequality (B.2.1) shows that the underlying norm is a continuous function. The inner product $\langle \cdot \mid \cdot \rangle$ is also continuous, in the following sense.

PROPOSITION B.2.1. *Suppose $\{x_n\}$ and $\{y_n\}$ are sequences in X converging to x and y, respectively. Then $\{\langle x_n \mid y_n \rangle\}$ converges to $\langle x \mid y \rangle$.*

PROOF. By the triangle and Cauchy-Schwarz inequalities, we have

$$\begin{aligned} |\langle x_n \mid y_n \rangle - \langle x \mid y \rangle| &= |\langle x - x_n \mid y - y_n \rangle + \langle x_n - x \mid y \rangle + \langle x \mid y_n - y \rangle| \\ &\leq \|x - x_n\| \, \|y - y_n\| + \|x_n - x\| \, \|y\| + \|x\| \, \|y_n - y\|. \end{aligned}$$

Letting n go to infinity completes the proof. □

In a finite-dimensional space, a linear functional f on X is necessarily continuous, since it has a Riesz representation z and therefore $|f(x) - f(y)| \leq \|z\| \|x - y\|$ by the Cauchy-Schwarz inequality. In contrast, the following example shows that a linear functional on an infinite-dimensional space may not be continuous.

EXAMPLE B.2.2. Let X be the inner product space of every sequence $\{x_n\}$ in $\mathbb{R}$ such that $x_n = 0$ for all but finitely many values of n, with the inner product $\langle x \mid y \rangle = \sum_n x_n y_n$. The functional $f(x) = \sum_n n x_n$ is linear but not continuous (why?). $\diamondsuit$

A sequence $\{x_n\}$ of points in X is **Cauchy** if for every $\varepsilon > 0$, there exists an integer N such that $m > n > N$ implies $x_m \in B(x_n; \varepsilon)$. A subset S of X is **complete** if every Cauchy sequence in S converges to a limit that is an element of S. A subset of a complete set is complete if and only if it is closed. The triangle inequality implies that every convergent sequence is Cauchy. The following example shows that the converse need not be true for an arbitrary inner product space. Intuitively, a Cauchy sequence should converge to something, but if that something is not within the space X, then the sequence is not convergent. As we will see shortly, difficulties of the sort do not arise in finite-dimensional spaces.

EXAMPLE B.2.3. Let $X = C[0,1]$ be the (infinite-dimensional) vector space of all continuous functions on the unit interval with the inner product $\langle x \mid y \rangle = \int_0^1 x(t)\, y(t)\, dt$. The sequence $\{x_n\}$ defined by $x_n(t) = 1/(1 + nt)$, $t \in [0,1]$, is Cauchy but does not converge in X. $\diamondsuit$

If X is finite-dimensional, the convergence or Cauchy property of a sequence is equivalent to the respective property of the sequence's coordinates relative to any given basis.

PROPOSITION B.2.4. *Suppose $B = (B_1, \ldots, B_d)'$ is a basis of X and $\sigma_n = \left(\sigma_n^1, \ldots, \sigma_n^d\right) \in \mathbb{R}^d$ for $n = 1, 2, \ldots$ The sequence $\{\sigma_n B\}$ is Cauchy (resp. converges to σB) if and only if the scalar sequence $\{\sigma_n^i\}$ is Cauchy (resp. converges to σ_i) for every coordinate $i \in \{1, \ldots, d\}$.*

PROOF. Suppose $\{\sigma_n^i\}$ is Cauchy for each i. The triangle inequality implies

$$\|\sigma_n B - \sigma_m B\| \leq \sum\nolimits_i \left|\sigma_n^i - \sigma_m^i\right| \|B_i\|,$$

from which it follows that $\{\sigma_n B\}$ is Cauchy. Conversely, suppose $\{\sigma_n B\}$ is Cauchy. Noting the identity $\sigma_n = \langle \sigma_n B \mid y \rangle$, where $y = B' \langle B \mid B' \rangle^{-1}$, apply the Cauchy-Schwarz inequality to obtain

$$\left|\sigma_n^i - \sigma_m^i\right| = |\langle \sigma_n B - \sigma_m B \mid y_i \rangle| \leq \|\sigma_n B - \sigma_m B\| \|y_i\|.$$

Therefore $\{\sigma_n^i\}$ is Cauchy. The claims in parentheses follow by the same argument, with σ in place of σ_m. $\square$

A **Hilbert space** is any inner product space that is complete (relative to the norm induced by the inner product). One of the fundamental properties of the real line is that it is complete. Given this fact, the last proposition implies that *every finite-dimensional inner product space is a Hilbert space.* Another consequence of Proposition B.2.4 is that for a finite-dimensional vector space, the convergence or Cauchy property of a sequence, and the continuity of a function are all properties that do not depend on the choice of an underlying inner product (or norm, as we will see shortly).

A subset S of X is (sequentially) **compact** if every sequence in S has a subsequence that converges to a vector in S. This definition immediately implies that a compact set is complete and therefore closed. A compact set S is also **bounded**, meaning that $\sup\{\|s\| : s \in S\} < \infty$. If S were unbounded, there would exist a sequence $\{s_n\}$ in S such that $\|s_n\| > n$ for all n, which precludes the existence of a convergent subsequence. On the real line, a closed bounded interval is compact since it can be sequentially subdivided in halves that contain infinitely many points of a given sequence, leading to the selection of subsequence that is Cauchy and therefore convergent (by the completeness of the real line). By virtue of Proposition B.2.4, this type of argument can be extended to $\mathbb{R}^d$ by applying it to each coordinate sequentially. Since every finite-dimensional space is isomorphic to $\mathbb{R}^d$, we have

PROPOSITION B.2.5. *In a finite-dimensional space, a set is compact if and only if it is closed and bounded.*

The claim is not true if the assumption of finite dimensionality is removed. For example, consider l_2, the Hilbert space of all square-summable sequences of real numbers with the inner product $\langle x \mid y\rangle = \sum_{n=1}^{\infty} x(n)\, y(n)$. The sequence $\{e_n\}$ in l_2, where $e_n(n) = 1$ and $e_n(k) = 0$ for all $k \neq n$, satisfies $\|e_n\| = 1$ for all n, but $\{e_n\}$ has no convergent subsequence. In fact, it can be shown that the closure of the unit ball is not compact in every infinite-dimensional normed space.

Our interest in compactness is mainly due to the following result, showing that a continuous function achieves its supremum and infimum over a compact set.

PROPOSITION B.2.6. *Suppose that S is a compact subset of X and the function $f : S \to \mathbb{R}$ is continuous. Then there exist $s^*, s_* \in S$ such that $f(s^*) \geq f(s) \geq f(s_*)$ for all $s \in S$.*

PROOF. Let $\{s_n\}$ be a sequence such that $\lim_n f(s_n) = \sup f$. By the compactness of S, there exists a subsequence of $\{s_n\}$ converging to some $s^* \in S$. Since f is continuous, $f(s^*) = \sup f$ and therefore $f(s^*) \geq f(s)$ for all $s \in S$. The same argument applied to $-f$ completes the proof. □

Proposition B.2.6 implies that in a finite-dimensional space all norms are topologically equivalent, in the following sense. Suppose X is finite-dimensional and consider any other norm $\|\cdot\|_*$ on X (not necessarily induced by any inner product). Since $\|\cdot\|_*$ is convex, it is continuous on the open ball $B(0;2)$ and therefore achieves a minimum and a maximum over the closure of $B(0;1)$. By the homogeneity of $\|\cdot\|_*$, there exist constants k and K such that $k\|x\| \leq \|x\|_* \leq K\|x\|$ for all $x \in X$. Therefore, for a finite-dimensional vector space, all norms define the same notion of convergence and hence the same compact sets. The situation is quite different with infinite-dimensional spaces, where different norms can define dramatically different notions of convergence.

We conclude this section with an overview of associated topological concepts and notation. Let S be any subset of X. The **closure** of S is the set of

$$\bar{S} = \{x \mid B(x;\varepsilon) \cap S \neq \emptyset \text{ for all } \varepsilon > 0\},$$

which is the same as the set of every vector x for which there exists some sequence in S that converges to x. The set S is **closed** if $S = \bar{S}$, or equivalently, if every convergent sequence in S converges to a point in S. The **interior** of S is the set

$$S^0 = \{x \mid B(x;\varepsilon) \subset S \text{ for some } \varepsilon > 0\}.$$

The **boundary** of S is the set $\bar{S} \setminus S^0$. The set S is **open** if $S = S^0$, or equivalently, if its complement $X \setminus S$ is closed. The set of all open subsets of X is known as the (norm) **topology** of X. The following properties of open and closed sets can be verified as an exercise. The empty set and X are both open and closed. The union of finitely many closed sets is closed, and the intersection of finitely many open sets is open. Arbitrary intersections of closed sets are closed, and arbitrary unions of open sets are open. The closure of a set is the intersection of all its closed supersets, and the interior of a set is the union of all its open subsets. Continuity of a function or compactness of a set are topological properties, in the sense that they depend on the underlying norm only through the topology they define. This fact will not be used or shown here, but can be found in textbooks that include an introduction to general topology including normed spaces (for example, Dudley [2002]).

B.3. Convexity

The purpose of this section is to introduce the central notions of convexity and concavity that underly the analysis of the rest of this appendix, as well as some basic topological implications of convexity assumptions. As always, the reference inner product space X is taken as given.

A set $C \subseteq X$ is **convex** if the line segment connecting any two points in C lies entirely within C: for all $x, y \in C$, $\alpha \in (0,1)$ implies

$\alpha x + (1-\alpha)\, y \in C$. Two important special types of convex set that arise in this text are convex cones and linear manifolds. A subset C of X is a **cone** if for all $x \in C$, $\alpha \in \mathbb{R}_+$ implies $\alpha x \in C$. A cone C is convex if and only if $x, y \in C$ implies $x+y \in C$ (and a convex cone C is a linear subspace if and only if $x \in C$ implies $-x \in C$). A set $M \subseteq X$ is a **linear manifold** if the line through any two points in M lies entirely within M: for all $x, y \in M$ and $\alpha \in \mathbb{R}$ implies $\alpha x + (1-\alpha)\, y \in M$. Clearly, every linear manifold is convex. Given any subset S of X and vector x, we write $x + S = S + x = \{x + s \mid s \in S\}$ to denote the **translation** of S by x. Note that if M is a linear manifold and x is any point of M, then $M - x$ is a linear subspace. Conversely, the translation of any linear subspace is a linear manifold. Therefore $M \subseteq X$ is a linear manifold if and only if it takes the form $M = x + L$ for some vector x and linear subspace L. In this case, the **dimension** of M is the dimension of L. If L is finite-dimensional, it is spanned by some basis $B = (B_1, \ldots, B_d)'$ and $M = \{x \in X \mid \langle B \mid x \rangle = b\}$ for some $b \in \mathbb{R}^d$. Conversely, every set of this form is a linear manifold. A convex set may or may not be closed, but the situation is less clear with linear manifolds. By Lemma B.2.4, a finite-dimensional linear subspace is necessarily complete and therefore closed if X is complete. In contrast, an infinite-dimensional linear subspace may not be closed even if the space X is complete.[2]

A function $f : D \to \mathbb{R}$, where $D \subseteq X$, is **concave** if D is a convex set and for all $x, y \in D$, $\alpha \in (0,1)$ implies $f(\alpha x + (1-\alpha)\, y) \geq \alpha f(x) + (1-\alpha)\, f(y)$. This property means that on a two-dimensional plot of f along the line segment connecting x and y, the graph of f lies above the line segment, and this is true for all $x, y \in D$. The function f is **convex** if $-f$ is convex. Note that a functional $f : X \to \mathbb{R}$ is linear if and only if it is both concave and convex. We saw in the last section that a linear functional is necessarily continuous if X is finite-dimensional. This fact generalizes as follows.

THEOREM B.3.1. *Suppose that X is finite dimensional, and $x \in X$ is in the interior of the domain of a concave function f. Then f is continuous at x.*

PROOF. Since every finite-dimensional inner product space is isomorphic to a Euclidean space, we assume that $X = \mathbb{R}^d$ with the Euclidean inner product. Choose $\alpha, \beta \in \mathbb{R}^d$ such that $\alpha_i < x_i < \beta_i$ for all i and $[\alpha, \beta] \equiv \prod_{i=1}^d [\alpha_i, \beta_i]$ is included in the domain of f. A point in $[\alpha, \beta]$ is **extreme** if all its coordinates are of the form α_i or β_i. (For example, for $d = 2$, $[\alpha, \beta]$ is a rectangle and its extreme points are the

[2]Let X is be the space of every sequence $x = \{x_n\}$ that is square-summable (that is, $\sum_n x_n^2 < \infty$) with the inner product $\langle x, y \rangle = \sum_n x_n y_n$. This space can be shown to be complete, but the linear subspace of every sequence x such that $x_n = 0$ for all but finitely many values of n is not closed.

corners.) Concavity of f implies that for every $y \in [\alpha, \beta]$, there exists an extreme point $\bar{y} \in [\alpha, \beta]$ such that $f(y) \geq f(\bar{y})$. (To see that, think of a plot of $f(y)$ on $[\alpha_i, \beta_i]$ as a function of its i^{th} coordinate and note that the minimum is achieved at one of the endpoints.) Since $[\alpha, \beta]$ has finitely many extreme points, f is minimized by one of them. Let b denote the minimum value, and let $r > 0$ be small enough so that $B(x; r) \subseteq [\alpha, \beta]$. Fixing any $y \in B(x; r)$, let $u \equiv (y - x) / \|y - x\|$ and $\phi(\alpha) \equiv f(x + \alpha u)$ for all $\alpha \in [-r, r]$. We then have the inequalities

$$\begin{aligned} K \equiv \frac{\phi(0) - b}{r} &\geq \frac{\phi(0) - \phi(-r)}{r} \\ &\geq \frac{\phi(\|y - x\|) - \phi(0)}{\|y - x\|} \geq \frac{\phi(r) - \phi(0)}{r} \geq \frac{b - \phi(0)}{r} = -K. \end{aligned}$$

The fact that ϕ is bounded below by b justifies the first and last inequalities. The three middle expressions represent slopes that decrease from left to right since $\phi : [-r, r] \to \mathbb{R}$ is concave. This proves that $|f(y) - f(x)| \leq K \|y - x\|$ for all $y \in B(x; r)$. □

Convexity assumptions in combination with completeness can play the role that compactness did in the last section, even in the infinite-dimensional case, where compact sets are harder to come by. Given a sequence $\{x_n\}$, we write $\text{conv}(x_n, x_{n+1}, \dots)$ for the set of all finite linear combinations of the form $\sum_i \alpha_i x_{n_i}$, where $\alpha_i > 0$, $\sum_i \alpha_i = 1$ and $n_i \geq n$ for all i. Equivalently, $\text{conv}(x_n, x_{n+1}, \dots)$ is the intersection of all convex sets that contain the points $x_n, x_{n+1}, \dots$.

LEMMA B.3.2. *Suppose $\{x_n\}$ is a bounded sequence that lies in a complete convex subset of X. Then there exists a convergent sequence $\{y_n\}$ such that $y_n \in \text{conv}(x_n, x_{n+1}, \dots)$ for all n.*

PROOF. Since the sequence $\{x_n\}$ is bounded,

$$L_n \equiv \inf \{\|y\| \mid y \in \text{conv}(x_n, x_{n+1}, \dots)\}$$

is finite, increasing in n, and converges to the finite limit $M \equiv \sup_n L_n$ as $n \to \infty$. Choose $y_n \in \text{conv}(x_n, x_{n+1}, \dots)$ such that $\|y_n\| < L_n + 1/n$ for all n. We will show that the sequence $\{y_n\}$ is Cauchy and hence convergent, since $\{x_n\}$ is assumed to lie in a complete convex set. Given any $\epsilon > 0$, choose N such that $1/N < \epsilon$ and $L_n > M - \epsilon$ for all $n > N$. For all $m > n > N$, we have $(1/2)(y_m + y_n) \in \text{conv}(x_n, x_{n+1}, \dots)$ and therefore $\|(1/2)(y_m + y_n)\| > M - \epsilon$ and

$$\begin{aligned} \|y_m - y_n\|^2 &= 2\|y_m\|^2 + 2\|y_n\|^2 - \|y_m + y_n\|^2 \\ &< 2(L_m + 1/m)^2 + 2(L_n + 1/n)^2 - 4(M - \epsilon)^2 \\ &< 4(M + \epsilon)^2 - 4(M - \epsilon)^2 = 16M\epsilon. \end{aligned}$$

The Cauchy property of $\{y_n\}$ follows. □

The following result is a corollary of Proposition B.2.6 in the finite-dimensional case, but is remarkable in that it applies equally well to the infinite-dimensional case and relies on the last lemma rather than compactness.

PROPOSITION B.3.3. *Suppose the nonempty set S is convex, complete and bounded, and the function $f : X \to \mathbb{R}$ is concave over S and continuous at every point of S. Then there exists $\bar{y} \in S$ such that $f(y) = \max\{f(s) \mid s \in S\}$.*

PROOF. Choose the sequence $\{x_n\}$ in S so that $\{f(x_n)\}$ converges to $M \equiv \sup\{f(x) \mid x \in S\}$. Let $\{y_n\}$ be as in the last lemma, converging to $y \in S$ (since S is closed). Given any $\epsilon > 0$, choose N large enough so that $n > N$ implies $f(x_n) > M - \epsilon$. Fixing $n > N$, suppose that $y_n = \sum_i \alpha_i x_{n_i}$ for finitely many $\alpha_i > 0$ such that $\sum_i \alpha_i = 1$ and $n_i \geq n$. Concavity of f implies that

$$M \geq f(y_n) \geq \sum_i \alpha_i f(x_{n_i}) > \sum_i \alpha_i (M - \epsilon) = M - \epsilon.$$

This shows that the sequence $\{f(y_n)\}$ converges to M and also converges to $f(y)$, since f is continuous at y. Therefore, $M = f(y)$. □

B.4. Projections on convex sets

In Section B.1 we characterized the projection of a vector onto a line by an orthogonality condition. As the geometric intuition suggests, this argument generalizes to projections on convex sets. The vector x_S is defined to be a **projection** of a vector x on a set $S \subseteq X$ if $x_S \in S$ and $\|x - s\| \geq \|x - x_S\|$ for all $s \in S$. In words, x_S is a point of S that minimizes the distance between x and every point in S. A projection x_S on S may not exist (for example, take X to be the real line, $S = (0, 1)$ and $x = 2$), but as shown in this section, assuming S is convex, if x_S exists it is unique, and if S is complete (Cauchy sequences in S converge in S) then x_S does exist. Our central concern is the dual characterization of a projection, which makes use of the following condition: The vector x **supports** the set S at $\bar{s} \in X$ if

$$\langle x \mid \bar{s} \rangle = \inf\{\langle x \mid s \rangle \mid s \in S\}. \tag{B.4.1}$$

This definition does not require that $\bar{s} \in S$, which is useful in a later section, but if $\bar{s} \in S$, then x supports S at $\bar{s}$ if and only if $\langle x \mid s - \bar{s} \rangle \geq 0$ for all $s \in S$. The latter inequality can be visualized as the requirement that the vectors x and $s - \bar{s}$ form an acute angle.

The central result on projections on convex sets follows. Note that the third part on existence applies in particular if X is a Hilbert space and S is closed, since then S is necessarily complete.

THEOREM B.4.1. *(Projection Theorem) Suppose S is a convex subset of the inner product space X. Then the following are true for all $x, y \in X$.*

(1) The vector x_S is a projection of x on S if and only if $x_S \in S$ and $x_S - x$ supports S at x_S.
(2) If x_S is a projection of x on S and y_S is a projection of y on S, then $\|x_S - y_S\| \leq \|x - y\|$. Therefore, if a projection of x onto S exists, it is unique.
(3) If S is complete, then the projection of x on S exists.

PROOF. 1. Fix $x_S, s \in S$ and define $x^\alpha = x_S + \alpha(s - x_S) \in S$ for all $\alpha \in [0,1]$. Then $\|x - x^\alpha\|^2 = \|x - x_S\|^2 + \alpha^2 \|s - x_S\|^2 + 2\alpha \langle x_S - x \mid s - x_S \rangle$. The claim follows from the observation that as α ranges over $[0,1]$, this quadratic is minimized at $\alpha = 0$ if and only if $\langle x_S - x \mid s - x_S \rangle \geq 0$ (why?).

2. Let $\delta \equiv y - x$ and $\delta_S \equiv y_S - x_S$. By part 1, $\langle x_S - x \mid \delta_S \rangle \geq 0$ and $\langle y - y_S \mid \delta_S \rangle \geq 0$. Adding up the two inequalities, we conclude $\langle \delta - \delta_S \mid \delta_S \rangle \geq 0$ and therefore

$$\|\delta\|^2 = \|\delta - \delta_S\|^2 + \|\delta_S\|^2 + 2\langle \delta - \delta_S \mid \delta_S \rangle \geq \|\delta_S\|^2 .$$

Applying this argument with $x = y$ proves the uniqueness claim.

3. The projection of x on S exists if and only if the projection of x on the intersection of S and a sufficiently large closed ball around x exists. We can therefore assume that S is bounded and the claim follows from Proposition B.3.3. □

The remainder of this section elaborates on the important special case of projections on linear manifolds. A vector x is said to be **orthogonal** to a linear manifold M if x is orthogonal to $y - z$ for all $y, z \in M$. The **orthogonal to M subspace**, denoted by $M^\perp$, is the linear subspace of all vectors that are orthogonal to M. Note that $M^\perp = (x + M)^\perp$ for every $x \in X$, and a vector supports M at some point if and only if it is orthogonal to M. Theorem B.4.1 applied to linear manifolds gives

COROLLARY B.4.2. *(Orthogonal Projection Theorem) Suppose M is a linear manifold in X and $x \in X$. A point x_M is the projection of x on M if and only if $x_M \in M$ and $x - x_M \in M^\perp$. Assume further that M is a complete (for example, finite-dimensional). Then x has a unique decomposition of the form $x = x_M + n$, $x_M \in M$, $n \in M^\perp$. Finally, if $M = x + L$ for a (necessarily complete) linear subspace L, then*[3] *$M^{\perp\perp} = L$.*

PROOF. We show the last claim only, since the rest is immediate from Theorem B.4.1. Also immediate is the fact that $M^\perp = L^\perp$ and $L \subseteq L^{\perp\perp}$ (whether L is complete or not). Assuming that L is complete,

[3] Although not needed in this text, it is not hard to show that for an arbitrary linear subspace L of a Hilbert space, $L^{\perp\perp} = \bar{L}$.

we now show that $L^{\perp\perp} \subseteq L$. Consider any $x \in L^{\perp\perp}$ and let $x = x_L + n$, where $x_L \in L$ and $n \in L^{\perp}$, and therefore $\langle x \mid n \rangle = \langle x_L \mid n \rangle + \langle n \mid n \rangle$ and $\langle x_L \mid n \rangle = 0$. Since $x \in L^{\perp\perp}$, we also have $\langle x \mid n \rangle = 0$. The last three equalities imply that $\langle n \mid n \rangle = 0$ and therefore $x = x_L \in L$. □

The following example demonstrates what can go wrong if M is not complete (and therefore infinite-dimensional).

EXAMPLE B.4.3. The space $C[0,1]$ of all continuous functions of the form $x : [0,1] \to \mathbb{R}$ is a Hilbert space under the inner product $\langle x \mid y \rangle = \int_0^1 x(t)\, y(t)\, dt$. It it existed, the projection of the zero vector onto the linear manifold $M \equiv \{x \in C[0,1] \mid x(0) = 1\}$ would minimize $\int_0^1 x(t)^2\, dt$ subject to the constraint $x(0) = 1$. Such a minimum does not exist within $C[0,1]$. ♢

Corollary B.4.2 implies the following useful relationship between orthogonal projections and Riesz representations.

PROPOSITION B.4.4. *Suppose that $f(y) = \langle x \mid y \rangle$ for all $y \in X$ and f_L is the restriction of f on the linear subspace L. The vector x_L is the Riesz representation of f_L in L if and only if it is the projection of x on L.*

The existence of orthogonal projections on complete linear subspaces implies that in any Hilbert space the Riesz representation of a continuous linear functional exists. The argument is redundant in the finite-dimensional case, since the claim was established in Proposition B.1.6, but still worth reviewing.

THEOREM B.4.5. *In a Hilbert space, a linear functional is continuous if and only if it has a (necessarily unique) Riesz representation.*

PROOF. Suppose f is a continuous linear functional. The null subspace $N = \{x : f(x) = 0\}$ is closed and therefore complete (since X is assumed complete). If $X = N$, the claim is trivial. Otherwise, pick any $z \in X$ such that $f(z) \neq 0$, let z_N be the projection of z on N and define $y = f(z - z_N)^{-1}(z - z_N)$. Then y is orthogonal to N and satisfies $f(y) = 1$. For any $x \in X$, the fact that $x - f(x)\, y \in N$ implies that $\langle x - f(x)\, y \mid y \rangle = 0$, and therefore $f(x) = \langle x \mid y \rangle / \langle y \mid y \rangle$. The vector $\langle y \mid y \rangle^{-1} y$ is the Riesz representation of f. □

Projections on finite-dimensional subspaces can be expressed by simple formulas in terms of a given basis.

PROPOSITION B.4.6. *Suppose L is a linear subspace of X and $B = (B_1, \ldots, B_d)'$ is a basis of L. For every $x \in X$, the vector*

$$x_L = \langle x \mid B' \rangle \langle B \mid B' \rangle^{-1} B \tag{B.4.2}$$

is the projection of x on L and $x - x_L$ is the projection of x on

$$L^{\perp} = \{y \in X \mid \langle B \mid y \rangle = 0\}. \tag{B.4.3}$$

PROOF. The projection expression (B.4.2) can be viewed as a corollary of Proposition (B.1.6) (why?). For a more direct argument, let $x_L = \sigma^{x_L} B$, where $\sigma^{x_L} = \langle x \mid B' \rangle \langle B \mid B' \rangle^{-1}$. Then $x - x_L \in L^\perp$, since for all $y = \sigma^y B$,

$$\langle x_L \mid y \rangle = \sigma^{x_L} \langle B \mid B' \rangle \sigma^{y'} = \langle x \mid B' \rangle \langle B \mid B' \rangle^{-1} \langle B \mid B' \rangle \sigma^{y'} = \langle x \mid y \rangle .$$

By Corollary B.4.2, x_L is the projection of x on L. Since $L \subseteq L^{\perp\perp}$, $x - (x - x_L) = x_L \in L^{\perp\perp}$ and $x - x_L \in L^\perp$, and therefore $x - x_L$ is the projection of x on $L^\perp$. □

A useful application is to optimization problems of the form

$$\min \{\|x\| \mid x \in X, \langle B \mid x \rangle = b\} \quad (b \in \mathbb{R}^d).$$

The minimizing value of x is the projection of the zero vector onto the linear manifold of all x such that $\langle B \mid x \rangle = b$, which is characterized below.

COROLLARY B.4.7. *Suppose $B = (B_1, \ldots, B_d)'$ is a column matrix of linearly independent vectors and $M = \{x \in X \mid \langle B \mid x \rangle = b\}$ for some column vector $b \in \mathbb{R}^d$. Then $M^\perp = \mathit{span}(B_1, \ldots, B_d)$ and $b' \langle B \mid B' \rangle^{-1} B$ is the projection of the zero vector on M.*

PROOF. Fix any $m \in M$ and note that $M - m$ is the linear subspace $L^\perp$ of equation (B.4.3). Therefore, $M^\perp = L^{\perp\perp} = L = \operatorname{span}(B)$. The point 0_M is the projection of the zero vector on M if and only if $m - 0_M$ is the projection of m on $L^\perp$. We can therefore apply Proposition B.4.6 to conclude that the projection 0_M of 0 on M exists and $0_M = \langle m \mid B' \rangle \langle B \mid B' \rangle^{-1} B$. Since $\langle B \mid m \rangle = b$, the result follows. □

Recall (from Section B.1) that a basis B is said to be **orthonormal** if the corresponding Gram matrix is an identity matrix. Another useful application of Proposition B.4.6 is the construction of an orthonormal basis, which is the basis for the claim that every finite-dimensional inner product space is isomorphic to a Euclidean space. We call a basis B **orthogonal** if its elements are pairwise orthogonal: $\langle B_i \mid B_j \rangle = 0$ for $i \neq j$. Clearly, B is orthonormal if and only if it is orthogonal and normalized in the sense that $\|B_i\| = 1$ for all i. Every orthogonal basis can be normalized by scaling. If the linear subspace L has a finite orthogonal basis B, then $\langle B \mid B' \rangle$ is diagonal and formula (B.4.2) for the projection of x on L reduces to

$$x_L = \sum_{i=1}^{n} \frac{\langle x \mid B_i \rangle}{\langle B_i \mid B_i \rangle} B_i.$$

Assuming X is finite-dimensional, this equation can be used to recursively construct an orthogonal basis of X, a process known as **Gram-Schmidt orthogonalization**. Start with any nonzero vector B_1 in

X. Given n orthogonal vectors $B_1, \ldots, B_n$ such that

$$L = \operatorname{span}(B_1, \ldots, B_n) \neq X,$$

choose any $x \in X \setminus L$ and define $B_{n+1} = x - x_L$, where x_L is the projection of x on L. If X is d-dimensional, the recursive construction terminates for $n = d$. Normalizing the resulting orthogonal basis proves that *every finite-dimensional inner product space has an orthonormal basis.*[4]

B.5. Supporting hyperplanes and (super)gradients

A **hyperplane** H is a linear manifold whose orthogonal subspace is of dimension one. If $H^\perp$ is spanned by the vector y and $\alpha = \langle y \mid \bar{x} \rangle$, where $\bar{x}$ is any point in H, then $H = \{x \mid \langle y \mid x \rangle = \alpha\}$. Conversely, for every nonzero vector y and scalar α, the preceding expression defines a hyperplane. The set $\{x \mid \langle y \mid x \rangle \geq \alpha\}$ defines a **half-space** whose boundary is H in the direction of the orthogonal-to-H vector y. Recall that the vector y is said to support the set S at $\bar{s} \in X$ if

$$\langle y \mid \bar{s} \rangle = \inf \{\langle y \mid s \rangle \mid s \in S\}.$$

This condition can be visualized as the inclusion of the closure of S in the half-space $\{x \mid \langle y \mid x \rangle \geq \alpha\}$, where $\alpha \equiv \langle y \mid \bar{s} \rangle$, with $\bar{S}$ touching H at $\bar{s}$. It is intuitively compelling that one should be able to support a convex set at any point of its boundary. The supporting hyperplane theorem shows that this is indeed true for a finite-dimensional space.

THEOREM B.5.1. *(Supporting Hyperplane Theorem) Suppose S is a convex subset of the finite-dimensional space X. Then for every vector x that is not in the interior of S, there exists a nonzero vector y such that $\langle y \mid x \rangle \leq \langle y \mid s \rangle$ for all $s \in S$. If x is on the boundary of S, then y supports S at x.*

PROOF. Since x is not interior, we can construct a sequence $\{x_n\}$ of vectors such that $x_n \notin \bar{S}$ for all n and $x_n \to x$ as $n \to \infty$. For each n, let $\bar{s}_n$ be the projection of x_n on $\bar{S}$ and define $y_n = (\bar{s}_n - x_n) / \|\bar{s}_n - x_n\|$. The dual characterization of projections gives $\langle y_n \mid \bar{s}_n \rangle \leq \langle y_n \mid s \rangle$ for all $s \in S$. By the second part of Theorem B.4.1, the sequence $\{\bar{s}_n\}$ converges to the projection $\bar{s}$ of x on $\bar{S}$. The sequence $\{y_n\}$ lies in the closure of the unit ball, which is compact, and therefore we can extract a subsequence $\{y_{n_k}\}$ that converges to some vector y of unit norm. By the continuity of inner products, $\langle y \mid \bar{s} \rangle \leq \langle y \mid s \rangle$ for all $s \in S$. If x is on the boundary of S, then $x = \bar{s}$ is the limit of some sequence $\{s_n\}$ in S. Therefore, $\{\langle y \mid s_n \rangle\}$ converges to $\langle y \mid \bar{s} \rangle$, which shows (B.4.1). If x is not on the boundary of S, then $\{y_{n_k}\}$

[4]The Gram-Schmidt orthogonalization procedure described here is of theoretical value, but is not numerically stable when implemented as an algorithm—better alternatives exist. A complete discussion of this point is given in Meyer [2004].

converges to $y = (\bar{s} - x) / \|\bar{s} - x\|$. Since $\langle y \mid \bar{s} - x \rangle > 0$, it follows that $\langle y \mid x \rangle < \langle y \mid \bar{s} \rangle \leq \langle y \mid s \rangle$ for all $s \in S$. □

The following example shows that the preceding theorem is not valid in the infinite-dimensional case.[5]

EXAMPLE B.5.2. Suppose $X = l_2$, the space of square summable sequences with the inner product $\langle x \mid y \rangle = \sum_{n=1}^{\infty} x_n y_n$. Let S equal the positive cone $\{x \in X \mid x_n \geq 0 \,\text{for all}\, n\}$ and pick any $\bar{s} \in S$ such that $\bar{s}_n > 0$ for all n. There is no nonzero vector that supports S at $\bar{s}$ (why?). ♢

COROLLARY B.5.3. *(Separating Hyperplane Theorem) Suppose A and B are convex subsets of the finite-dimensional inner-product space X. If $A \cap B = \emptyset$, then there exists a nonzero $y \in X$ such that*

$$\inf_{a \in A} \langle y \mid a \rangle \geq \sup_{b \in B} \langle y \mid b \rangle \,. \tag{B.5.1}$$

PROOF. Since $A \cap B = \emptyset$, the zero vector is not in

$$S \equiv A - B = \{a - b \mid a \in A, b \in B\} \,.$$

By the supporting hyperplane theorem, there exists a nonzero $y \in X$ such that $\langle y \mid s \rangle \geq 0$ for all $s \in S$ and therefore $\langle y \mid a \rangle \geq \langle y \mid b \rangle$ for all $a \in A$ and $b \in B$. □

Another useful application of the supporting hyperplane theorem arises when the convex set being supported is defined as the space below the graph of a concave function, leading to the concept of a supergradient, which for a smooth function corresponds to a directional derivative. Consider any function $f : D \to \mathbb{R}$, where $D \subseteq X$. The vector y is a **supergradient** of f at $x \in D$ if it satisfies the **gradient inequality**:

$$f(x + h) \leq f(x) + \langle y \mid h \rangle \;\text{ for all } h \text{ such that } x + h \in D.$$

The **superdifferential** of f at x, denoted by $\partial f(x)$, is the set of all supergradients of f at x. The supergradient property can be visualized as a support condition in the space $X \times \mathbb{R}$ with the inner product

$$\langle (x_1, \alpha_1) \mid (x_2, \alpha_2) \rangle = \langle x_1 \mid x_2 \rangle + \alpha_1 \alpha_2, \quad x_i \in X,\ \alpha_i \in \mathbb{R}.$$

The **subgraph** of f is the set

$$\text{sub}(f) = \{(x, \alpha) \in D \times \mathbb{R} \mid \alpha \leq f(x)\} \,. \tag{B.5.2}$$

[5]The set S of the example has an empty interior relative to the topology defined by the induced norm (why?). If S is assumed to have a non-empty interior, then the validity of the Supporting Hyperplane Theorem is restored in any Hilbert space (as well as extensions), but this version of the result is of limited practical use, because in infinite-dimensional spaces, so many convex sets of interest have empty interiors.

The preceding definitions imply that

$$y \in \partial f(x) \iff (y,-1) \text{ supports } \operatorname{sub}(f) \text{ at } (x, f(x)). \tag{B.5.3}$$

THEOREM B.5.4. *In a finite-dimensional inner product space, the superdifferential of a concave function at an interior point of its domain is nonempty, convex and compact.*

PROOF. Suppose $f : D \to \mathbb{R}$ is concave and $x \in D^0$. By the supporting hyperplane theorem, $\operatorname{sub}(f)$ is supported by some nonzero $(y,-\beta) \in X \times \mathbb{R}$ at $(x, f(x))$:

$$\langle y \mid x\rangle - \beta f(x) = \min\{\langle y \mid z\rangle - \beta\alpha \mid \alpha \le f(z),\ z \in D,\ \alpha \in \mathbb{R}\}.$$

Since the left-hand side is finite, it follows that $\beta \ge 0$. If $\beta = 0$, then y supports D at x, which contradicts the assumption that x is an interior point of D. Therefore, $\beta > 0$ and $y/\beta \in \partial f(x)$. This proves that $\partial f(x)$ is nonempty. It follows easily from the definitions that $\partial f(x)$ is also convex and closed. Finally, we show that $\partial f(x)$ is bounded, utilizing the finite dimensionality assumption. We assume $x = 0$, which entails no loss of generality (why?). Theorem B.3.1 implies that f is bounded below over some ball centered at zero. Let $\varepsilon > 0$ and $K \in \mathbb{R}$ be such that for all $z \in X$, $\|z\| = \varepsilon$ implies $z \in D^0$ and $f(z) > K$. It follows that for every $y \in \partial f(0)$,

$$K < f\left(-\frac{\varepsilon}{\|y\|}y\right) \le f(0) + \left\langle y \mid -\frac{\varepsilon}{\|y\|}y\right\rangle = f(0) - \varepsilon\|y\|,$$

which results in the bound $\|y\| < (f(0) - K)/\varepsilon$. This proves that $\partial f(x)$ bounded. Since it is also closed, it is compact. □

Assuming it exists, the **directional derivative of f at x in the direction** h is denoted and defined by

$$f'(x;h) = \lim_{\alpha \downarrow 0} \frac{f(x+\alpha h) - f(x)}{\alpha}.$$

The gradient of f at x is said to exist if $f'(x;h)$ exists for every $h \in X$ and the functional $f'(x;\cdot)$ is linear. In this case, the Riesz representation of the linear functional $f'(x;\cdot)$ is the **gradient** of f at x, denoted by $\nabla f(x)$. Therefore, when it exists, the gradient $\nabla f(x)$ is characterized by the restriction $f'(x;h) = \langle \nabla f(x) \mid h\rangle$ for all $h \in X$.

THEOREM B.5.5. *Suppose D is a convex subset of the finite-dimensional inner product space X, the function $f : D \to \mathbb{R}$ is concave and x is any interior point of D. Then the directional derivative $f'(x;h)$ exists for all $h \in X$ and the following are true:*

(1) $\partial f(x) = \{y \in X \mid \langle y \mid h\rangle \ge f'(x;h) \text{ for all } h \in X\}$.
(2) $f'(x;h) = \min\{\langle y \mid h\rangle \mid y \in \partial f(x)\}$ for every $h \in X$.
(3) The gradient $\nabla f(x)$ exists if and only if $\partial f(x)$ is a singleton, in which case $\partial f(x) = \{\nabla f(x)\}$.

PROOF. Fix any $h \in X$, let $A = \{\alpha \in \mathbb{R} \mid x + \alpha h \in D^0\}$ (an open interval) and define the concave function $\phi : A \to \mathbb{R}$ by $\phi(\alpha) = f(x + \alpha h) - f(x)$. Note that $f'(x; h) = \phi'_+(0)$ (the right derivative ϕ at zero). For each $\alpha \in A$, let $\Delta(\alpha) = \phi(\alpha)/\alpha$, which is the slope of the line segment on the plane connecting the origin to the point $(\alpha, \phi(\alpha))$. Consider any decreasing sequence $\{\alpha_n\}$ of positive scalars in A. Concavity of ϕ implies that the corresponding sequence of slopes $\{\Delta(\alpha_n)\}$ is increasing. Moreover, fixing any $\varepsilon > 0$ such that $x - \varepsilon h \in D^0$, we have $\Delta(\alpha_n) \leq \Delta(-\varepsilon)$ for all n. Being increasing and bounded, the sequence $\{\Delta(\alpha_n)\}$ converges. This proves that the (monotone) limit $f'(x; h) = \lim_{\alpha \downarrow 0} \Delta(\alpha)$ exists and is finite.

(1) If $y \in \partial f(x)$, the gradient inequality implies that $\langle y \mid h \rangle \geq \Delta(\alpha)$ for all positive $\alpha \in A$. Letting $\alpha \downarrow 0$, it follows that $\langle y \mid h \rangle \geq f'(x; h)$. Conversely, suppose that $y \notin \partial f(x)$ and therefore $\Delta(1) = f(x + h) - f(x) > \langle y \mid h \rangle$ for some h. We saw earlier that $\Delta(\alpha)$ increases to $f'(x; h)$ as $\alpha \downarrow 0$ and therefore $f'(x; h) \geq \Delta(1) > \langle y \mid h \rangle$ for some h.

(2) In light of part 1, the claim of part 2 follows if given $h \in X$, we can produce some $y \in \partial f(x)$ such that $\langle y \mid h \rangle = f'(x; h)$. The gradient inequality $\phi(\alpha) \leq \phi'_+(0)\alpha$ for all $\alpha \in A$ (equivalently, $f(x + \alpha h) \leq f(x) + \alpha f'(x; h)$ for all $\alpha \in A$) can be restated as the condition that the line segment

$$L = \{(x + \alpha h, f(x) + \alpha f'(x; h)) : \alpha \in A\} \subseteq X \times \mathbb{R}$$

does not intersect the interior of the subgraph of f as defined in (B.5.2). By the separating hyperplane theorem (Corollary B.5.3), there exists nonzero $(p, \beta) \in X \times \mathbb{R}$ that separates the sets L and $\operatorname{sub}(f)$:

$$\inf_{\alpha \in A} \langle p \mid x + \alpha h \rangle + \beta (f(x) + \alpha f'(x; h)) \geq \sup_{(\tilde{x}, \alpha) \in \operatorname{sub}(f)} \langle p \mid \tilde{x} \rangle + \beta\alpha.$$

Since the right-hand side must be finite, $\beta > 0$. It follows that the right-hand side is at least as large as $\langle p \mid x \rangle + \beta f(x)$, which is also obtained as the expression on the left-hand side with $\alpha = 0$. Since $0 \in A^0$, the coefficient of α on the left-hand side must vanish: $\langle p \mid h \rangle + \beta f'(x; h) = 0$. Therefore, $f'(x; h) = \langle y \mid h \rangle$, where $y = -p/\beta$, and the separation condition reduces to the gradient inequality defining the condition $y \in \partial f(x)$.

(3) If $\partial f(x) = \{\delta\}$, then part 2 implies that $f'(x; h) = \langle \delta \mid h \rangle$ for all $h \in X$ and therefore $\delta = \nabla f(x)$. Conversely, if the gradient exists, then part1 implies that $\langle \nabla f(x) - y \mid h \rangle \leq 0$ for all $h \in X$ and $y \in \partial f(x)$. Letting $h = \nabla f(x) - y$, it follows that $y = \nabla f(x)$ if $y \in \partial f(x)$. □

B.6. Optimality conditions

The characterization of optima is a main application of convex analysis in economics. We already saw an example with the projection theorem. In this section, we focus on the optimization problem defining the function $F : \mathbb{R}^n \to [-\infty, +\infty]$, where

$$F(\delta) \equiv \sup\{f(x) \mid g(x) \leq \delta,\ x \in D\}, \quad \delta \in \mathbb{R}^n, \tag{B.6.1}$$

with the convention $\inf \emptyset = -\infty$, for given set $D \subseteq X$ and functions $f : D \to \mathbb{R}$ and $g : D \to \mathbb{R}^n$, for some positive integer n. Since δ can be absorbed in the specification of g, it is sufficient to consider the case $\delta = 0$. To do so, however, it is useful to consider the entire function F.

Assuming $F(0)$ is finite, the **superdifferential** of F at zero is

$$\partial F(0) \equiv \{\lambda \in \mathbb{R}^n \mid F(\delta) - F(0) \leq \lambda \cdot \delta \text{ for all } \delta \in \mathbb{R}^n\} \subseteq \mathbb{R}^n_+,$$

where the inclusion is due to the fact that $\delta \geq 0$ implies $F(\delta) \geq F(0)$. It is closely related to the **Lagrangian** function

$$\mathcal{L}(x, \lambda) \equiv f(x) - \lambda \cdot g(x), \quad x \in C, \quad \lambda \in \mathbb{R}^n.$$

LEMMA B.6.1. *Assume $F(0)$ is finite. For all $\lambda \in \mathbb{R}^n_+$, $\lambda \in \partial F(0)$ if and only if $F(0) = \sup_{x \in D} \mathcal{L}(x, \lambda)$.*

PROOF. In $\mathbb{R}^n \times \mathbb{R}$ with the Euclidean inner product, $\lambda \in \partial F(0)$ if and only if $(\lambda, -1)$ supports at $(0, F(0))$ the set A of all $(a, \alpha) \in \mathbb{R}^n \times \mathbb{R}$ such that $\alpha < F(a)$ (why?). Similarly, $F(0) = \sup_{x \in D} \mathcal{L}(x, \lambda)$ if and only if $(\lambda, -1)$ supports at $(0, F(0))$ the set B of all $(b, \beta) \in \mathbb{R}^n \times \mathbb{R}$ for which there exists $x \in D$ such that $g(x) \leq b$ and $\alpha < F(x)$ (why?). Finally, note that $A = B$. □

Assuming that the supremum defining $F(0)$ is a maximum and that $\partial F(0)$ is nonempty, the following main result provides a way of converting the constrained optimization problem defining $F(0)$ to a corresponding unconstrained problem. The intuitive idea is that through the condition $\lambda \in \partial F(0)$, the parameter λ, known as a **Lagrange multiplier**, provides an appropriate pricing of the constraint $g \leq 0$. If the constraint $g_i(x) \leq 0$ is slack, then the corresponding price of the constraint must be zero. This leads to the so-called **complementary slackness** condition $\lambda \cdot g(x) = 0$, which, given the inequalities $g(x) \leq 0$ and $\lambda \geq 0$, is equivalent to $g_i(x) < 0 \implies \lambda_i = 0$.

THEOREM B.6.2. *Suppose that $x \in D$, $g(x) \leq 0$ and $\lambda \in \mathbb{R}^n$. Then the following two conditions are equivalent:*

(1) $f(x) = F(0)$ and $\lambda \in \partial F(0)$.
(2) $\mathcal{L}(x, \lambda) = \max_{y \in D} \mathcal{L}(y, \lambda), \quad \lambda \cdot g(x) = 0, \quad \lambda \geq 0.$

PROOF. Suppose condition 1 holds. It has already been noted that $\lambda \in \partial F(0)$ implies $\lambda \geq 0$. Since $g(x) \leq 0$, $\mathcal{L}(x, \lambda) \geq f(x) = F(0)$. By the last lemma, $F(0) \geq \mathcal{L}(x, \lambda)$. Therefore, $\mathcal{L}(x, \lambda) = f(x)$ and

$\lambda \cdot g(x) = 0$. Conversely, condition 2 implies that $f(x) = \mathcal{L}(x, \lambda) \geq \mathcal{L}(y, \lambda) \geq f(y)$ for all y such that $g(y) \leq 0$. Therefore, $f(x) = F(0)$ and condition 1 follows by the last lemma. □

Assuming the existence of a maximum, the preceding optimality conditions are applicable if and only if $\partial F(0)$ is nonempty. The following lemma gives convexity-based sufficient conditions that are easy to check if they can be satisfied.

LEMMA B.6.3. *Suppose that D and g are convex, f is concave, $F(0)$ is finite and there exists $x \in D$ such that $g_i(x) < 0$ for every i. Then $\partial F(0) \neq \emptyset$.*

PROOF. By the supporting hyperplane theorem, the convex set $\text{sub}(F)$ is supported at $(0, F(0))$ by some nonzero $(\lambda, -\alpha)$, and therefore $\upsilon \leq F(\delta)$ implies $-\alpha F(0) \leq \lambda \cdot \delta - \alpha\upsilon$ for all $\delta \in \mathbb{R}^n$. If $\alpha < 0$, setting $\delta = 0$ leads to a contradiction. The Slater condition ensures that $F(\delta) > -\infty$ for all sufficiently small δ. Therefore $\alpha > 0$ and $\lambda/\alpha \in \partial F(0)$. □

The optimality conditions given so far are global in that they utilize the structure of the objective and constraint functions on their entire domain. Another type of optimality condition is local in that it provides implications of the fact that infinitesimal feasible perturbations from a reference optimum cannot improve the objective. The simplest instance of such an argument gives necessary optimality conditions for an unconstrained maximum. Suppose that x is interior to D and the gradient of f at x exists. If $f(x) = \max_{y \in D} f(y)$ then $\nabla f(x) = 0$. To see why, for every $y \in X$, consider the function on a small enough neighborhood of zero on the real line mapping α to $f(x + \alpha y)$. Since this function is maximized for $\alpha = 0$, its derivative at zero, $\nabla f(x) \cdot y$, must equal zero, and since y is arbitrary, $\nabla f(x) = 0$. In the converse direction, in order for the local condition $\nabla f(x) = 0$ to imply the global condition $f(x) = \max_{y \in D} f(y)$, we need a global regularity condition. For example, concavity of f works via the gradient inequality.

The preceding argument can be applied in particular to the function $\mathcal{L}(\cdot, \lambda)$ of the second condition of Theorem B.6.2. Assuming $x \in D^0$ and the existence of $\nabla f(x)$ and $\nabla g(x)$, $\mathcal{L}(x, \lambda) = \max_{y \in D} \mathcal{L}(y, \lambda)$ implies $\nabla f(x) = \lambda \cdot \nabla g(x)$, and the converse is true if D and g are convex and f is concave. We are therefore led to consider the **Kuhn-Tucker conditions:**

$$\nabla f(x) = \lambda \cdot \nabla g(x), \quad \lambda \cdot g(x) = 0, \quad \lambda \geq 0. \tag{B.6.2}$$

The sufficiency of the Kuhn-Tucker conditions for optimality under convexity assumptions is covered by Theorem B.6.2 and the gradient inequality. A necessity argument via Theorem B.6.2 and Lemma B.6.3 requires redundant global convexity assumptions. Instead, we can derive the necessity of the Kuhn-Tucker conditions for optimality with

the following local argument. Note that even though convexity makes no appearance in the theorem's statement, the separating hyperplane theorem is at the core of its proof.

THEOREM B.6.4. *Suppose that the gradients of f and g at $x \in D^0$ exist, and there exists some vector h such that $\langle \nabla g_i(x) \mid h \rangle < 0$ for all i. If $F(x) = f(x)$ and $g(x) \leq 0$, then the Kuhn-Tucker conditions (B.6.2) hold for some $\lambda \in \mathbb{R}^n$.*

PROOF. Define the convex set A of all $(a, \alpha) \in \mathbb{R}^n \times \mathbb{R}$ such that $a_i < 0$ for all i and $\alpha > 0$. Also define the convex set B of all $(b, \beta) \in \mathbb{R}^n \times \mathbb{R}$ for which there exists $h \in X$ such that

$$g(x) + \langle \nabla g(x) \mid h \rangle \leq b \quad \text{and} \quad \langle \nabla f(x) \mid h \rangle \geq \beta.$$

Optimality of x implies that for all $h \in X$, if $g_i(x) + \langle \nabla g_i(x) \mid h \rangle < 0$ for all i, then $\langle \nabla f(x) \mid h \rangle \leq 0$ (why?). Therefore $A \cap B = \emptyset$. By the separating hyperplane theorem (Corollary B.5.3), there exists nonzero $(-\lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$(a, \alpha) \in A \implies -\lambda \cdot a + \mu\alpha \geq 0, \tag{B.6.3}$$

$$(b, \beta) \in B \implies -\lambda \cdot b + \mu\beta \leq 0. \tag{B.6.4}$$

Condition (B.6.3) implies that $\lambda, \mu \geq 0$. The case $\mu = 0$ is ruled out by the regularity assumption on the gradient of g. After proper scaling, we can therefore assume that $\mu = 1$. Condition (B.6.4) with $\mu = 1$ implies that

$$-\lambda \cdot (g(x) + \langle \nabla g(x) \mid h \rangle) + \langle \nabla f(x) \mid h \rangle \leq 0 \quad \text{for all } h \in X,$$

which together with the inequalities $\lambda \geq 0$ and $g(x) \leq 0$ results in the Kuhn-Tucker conditions (B.6.2). □

This preceding result excludes equality constraints where $g_i(x) \leq 0$ and $g_i(x) \geq 0$ are both binding, but the ideas extend easily to this case. For example, consider the simple case of maximization of f over the linear manifold $M = \{x \in X \mid \langle B \mid x \rangle = b\}$, where $B = (B_1, \ldots, B_d)'$ is a column matrix of vectors and $b \in \mathbb{R}^d$. Assume that x is interior to D and the gradient of f at x exists. Suppose that x is optimal in the sense that $f(x) = \max_{y \in M \cap D} f(y)$. Then for every y such that $\langle B \mid y \rangle = 0$, $x + \alpha y \in M$ for all $\alpha \in \mathbb{R}$. Setting the derivative of $\alpha \mapsto f(x + \alpha y)$ at zero equal to zero, we conclude that $\nabla f(x)$ is orthogonal to all $y \perp \operatorname{span}(B)$, and therefore $\nabla f(x) \in \operatorname{span}(B)$. The latter is a necessary local optimality condition. By virtue of the gradient inequality, it is also a global sufficient optimality condition under the assumption that $x \in M \cap D$, D is convex and f is concave. The argument extends to equality constraints of the form $g(x) = 0$, with $\nabla g(x)$ playing the role of B in the preceding argument, but we have no need for this extension in this text.

Bibliography

J. Aczél. *Lectures on Functional Equations and Their Applications*. Dover Publications, Mineola, NY, 2006. 169

F. J. Anscombe and R. J. Aumann. A definition of subjective probability. *Annals of Mathematical Statistics*, 34:199–205, 1963. 172

D. Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge University Press, Cambridge, U.K., 2004. 76

K. J. Arrow. Le rôle des valeurs boursiéres pour la repartition la meillure des risques. *Econométrie, Colloques Internationaux du Centre National de la Recherche Scientifique*, 40:41–47, 1953. 21, 106

K. J. Arrow. The role of securities in the optimal allocation of risk bearing. *Review of Economic Studies*, 31:91–96, 1963. 21, 106

K. J. Arrow. *Aspects of the Theory of Risk Bearing*. Yrjo Jahnssonin Saatio, Helsinki, 1965. 144

K. J. Arrow. *Essays in the Theory of Risk Bearing*. North Holland, London, 1971. 59, 144

R. J. Aumann. Values of markets with a continuum of traders. *Econometrica*, 43(4):611–646, 1975. 111

D. P. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, Belmont, MA, 2003. 177

F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 3:637–654, 1973. 90

J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer-Verlag, New York, 2000. 177

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, United Kingdom, 2004. 177

E. Cinlar. *Probability and Stochastics*. Springer-Verlag, New York, 2010. 76

J. Cox and S. Ross. The valuation of options for alternative stochastic processes. *Journal of Financial Economics*, 3:145–166, 1976. 59

R. Dalang, A. Morton, and W. Willinger. Equivalent martingale measures and no-arbitrage in stochastic securities market models. *Stochastics and Stochastic Reports*, 29:185–201, 1990. 22

G. Debreu. *Theory of Value*. Cowles Foundation Monograph, Yale University Press, New Haven, 1959. 21, 106

G. Debreu. *Mathematical Economics: Twenty Papers of Gerard Debreu*. Cambridge University Press, New York, 1983. 163, 165

F. Delbaen and W. Schachermayer. *The Mathematics of Arbitrage*. Springer-Verlag, New York, 2006. 22

P. DeMarzo and C. Skiadas. Aggregation, determinacy, and informational efficiency for a class of economies with asymmetric information. *Journal of Economic Theory*, 80:123–152, 1998. 158

P. DeMarzo and C. Skiadas. On the uniqueness of fully informative rational expectations equilibria. *Economic Theory*, 13:1–24, 1999. 158

S. Dreyfus. Richard Bellman on the birth of dynamic programming. *Operations Research*, 50(1), 2002. 161

J. Drèze. Market allocation under uncertainty. *European Economic Review*, 15:133–165, 1971. 59

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, New York, 2002. 149, 185

D. Duffie and L. G. Epstein. Stochastic differential utility. *Econometrica*, 60:353–394, 1992a. 145, 150

D. Duffie and L. G. Epstein. Asset pricing with stochastic differential utility. *Review of Financial Studies*, 5:411–436, 1992b. 150

D. Duffie and C. Skiadas. Continuous-time security pricing: A utility gradient approach. *Journal of Mathematical Economics*, 23:107–131, 1994. 150

N. Dunford and J. T. Schwartz. *Linear Operators, Part I, General Theory*. Wiley, 1988. 177

I. Ekeland and R. Témam. *Convex Analysis and Variational Problems*. SIAM, Philadelphia, PA, 1999. 177

L. Epstein and S. Zin. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework. *Econometrica*, 57:937–969, 1989. 127

L. P. Hansen and R. Jagannathan. Implications of security market data for models of dynamic economies. *Journal of Political Economy*, 99: 225–262, 1991. 58

M. J. Harrison and D. M. Kreps. Martingale and arbitrage in multiperiod securities markets. *Journal of Economic Theory*, 20:381–408, 1979. 59

J. Jacod and P. Protter. *Discretization of Processes*. Springer Verlag Berlin Heidelberg, 2012. 76

J. Jacod and A. N. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer-Verlag, Berlin Heidelberg, second edition, 2003. 33, 76, 79

Y. M. Kabanov and D. O. Kramkov. No-arbitrage and equivalent martingale measure: An elementary proof of the Harrison-Pliska theorem. *Theory of Probability and its Applications*, 39:523–527, 1994. 22

D. H. Krantz, R. D. Luce, P. Suppes, and A. Tversky. *Foundations of Measurement*, volume I. Academic Press, Inc., San Diego, 1971. 165

D. Kreps and E. Porteus. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica*, 46:185–200, 1978. 145

D. M. Kreps. Arbitrage and equilibrium in economies with infinitely many commodities. *Journal of Mathematical Economics*, 8:15–35, 1981. 22

D. G. Luenberger. *Optimization by Vector Space Methods*. Wiley, New York, 1969. 177

H. Markowitz. Portfolio selection. *Journal of Finance*, 7:77–91, 1952. 54

R. C. Merton. Lifetime portfolio selection under uncertainty: The continuous time case. *Review of Economics and Statistics*, 51:247–257, 1969. 153

R. C. Merton. Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, 3:373–413, 1971. Erratum 6 (1973): 213–214. 153

C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2004. 192

P. Mörters and Y. Peres. *Brownian Motion*. Cambridge University Press, New York, 2010. 78

L. Narens. *Abstract Measurement Theory*. MIT Press, Cambridge, MA, 1985. 165

E. Pardoux and S. Peng. Adapted solution of a backward stochastic differential equation. *Systems and Control Letters*, 14:55–61, 1990. 145

J. W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32:122–136, 1964. 144

F. P. Ramsey. Truth and probability. In H. E. Kyburg, Jr. and H. E. Smokler, editors, *Studies in Subjective Probability (1980)*. Robert E. Krieger Publishing Company, New York, 1926. 172

P. J. Reny. A simple proof of the nonconcavifiability of functions with linear not-all-parallel contour sets. *Journal of Mathematical Economics*, 49:506–508, 2013. 111

D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*. Springer, New York, third edition, 1999. 78

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, New Jersey, 1970. 177

S. A. Ross. A simple approach to the valuation of risky streams. *Journal of Business*, 51:453–475, 1978. 22

L. J. Savage. *The Foundations of Statistics*. Dover Publications (1972), New York, 1954. 172

W. Schachermayer. A Hilbert-space proof of the fundamental theorem of asset pricing. *Insurance Mathematics and Economics*, 11:249–257, 1992. 22

M. Schroder and C. Skiadas. Optimal consumption and portfolio selection with stochastic differential utility. *Journal of Economic Theory*,

89:68–126, 1999. 145

M. Schroder and C. Skiadas. An isomorphism between asset pricing models with and without linear habit formation. *Review of Financial Studies*, 15:1189–1221, 2002. 119

M. Schroder and C. Skiadas. Optimal lifetime consumption-portfolio strategies under trading constraints and generalized recursive preferences. *Stochastic Processes and Their Applications*, 108:155–202, 2003. 147

M. Schroder and C. Skiadas. Lifetime consumption-portfolio choice under trading constraints and nontradeable income. *Stochastic Processes and their Applications*, 115:1–30, 2005. 147

C. Skiadas. *Advances in the Theory of Choice and Asset Pricing*. PhD thesis, Stanford University, 1992. 150

C. Skiadas. Subjective probability under additive aggregation of conditional preferences. *Journal of Economic Theory*, 76:242–271, 1997. 172

C. Skiadas. Recursive utility and preferences for information. *Economic Theory*, 12:293–312, 1998. 145

C. Skiadas. *Asset Pricing Theory*. Princeton Univ. Press, Princeton, NJ, 2009. 5, 53, 137, 169, 172

C. Skiadas. Scale-invariant asset pricing and consumption/portfolio choice with general attitudes toward risk and uncertainty. *Mathematics and Financial Economics*, 7:431–456, 2013a. 140

C. Skiadas. Scale-invariant uncertainty-averse preferences and source-dependent constant relative risk aversion. *Theoretical Economics*, 8: 59–93, 2013b. 169, 174

C. Skiadas. Dynamic choice with constant source-dependent relative risk aversion. *Economic Theory*, 3:393–422, 2015. 174

J. Stoer and C. Witzgall. *Convexity and Optimization in Finite Dimensions*. Springer-Verlag, Berlin, Germany, 1970. 22

R. H. Strotz. Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, 23:165–180, 1957. 104

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944. 172

P. P. Wakker. Cardinal coordinate independence for expected utility. *Journal of Mathematical Psychology*, 28:110–117, 1984. 172

P. P. Wakker. The algebraic versus the topological approach to additive representations. *Journal of Mathematical Psychology*, 32:421–435, 1988. 165, 172

P. P. Wakker. *Additive Representations of Preferences*. Kluwer, Dordrecht, The Netherlands, 1989. 165, 172

L. Walras. *Eléments d'Economie Pure*. Corbaze, Lausanne. English translation: *Elements of Pure Economics*, R. D. Irwin, Homewood, IL (1954), 1874. 106

P. Weil. The equity premium puzzle and the risk-free rate puzzle. *Journal of Monetary Economics*, 24:401–421, 1989. 127

P. Weil. Non-expected utility in macroeconomics. *The Quarterly Journal of Economics*, 105:29–42, 1990. 127

H. Xing. Consumption investment optimization with epstein-zine utility in incomplete markets. *Finance and Stochastics*, 21:227–262, 2017. 145

M. E. Yaari. A note on separability and quasiconcavity. *Econometrica*, 45:1183–1186, 1977. 167

J. A. Yan. Caracterisation d'une class d'ensembles convexes de l^1 ou h^1. *Lecture Notes in Mathematics*, 784:220–222, 1980. 22