(ii) follows from (i) by taking complements. We leave the details as an exercise.   □

Definition 1.11. A **probability space** is a triple $(\Omega, \mathcal{F}, P)$, where Ω is a set, called the **sample space**, $\mathcal{F}$ is a σ-field of subsets of Ω, and P is a probability measure on $\mathcal{F}$.

Definition 1.12. Elements of $\mathcal{F}$ are called **events.**

Remark 1.13. (i) In many cases, $\mathcal{F}$ is the class of all subsets of Ω. But sometimes we do not know the probabilities of all subsets, and sometimes—for technical reasons we need not go into here—it is simply not possible to define the probability of all sets. For those who have seen it, the primary example of this is Lebesgue measure. It cannot be defined on all subsets of the line, but only on a class of sets called the Lebesgue-measurable sets.

(ii) In any case, the class of sets on which we can define a probability is a σ-field. This is implicit in the very definition of a probability measure, since, for example, (iii) implies that the probability of a countable union of events can be defined, so the union must itself be an event.

(iii) There is another important reason for using σ-fields. It will not come up until Chapter 5.4, but it is worth mentioning here. It turns out that σ-fields are a very flexible way to represent information. One could say that they are information in the raw.

Problems 1.2

1.3. Let A, B, and C be sets. Prove or disprove:

    (a) $(A - B) - C = A \cap B^c \cap C^c$.

    (b) $(A \cap B) - C = (A - C) \cap (B - C)$.

    (c) $A \cup B = B \Longleftrightarrow A \subset B$.

1.4. Let n be an integer, let $\Omega = \mathbb{R}$, let $\mathcal{F}_n$ be the smallest σ-field containing all the intervals $(\frac{i}{n}, \frac{i+1}{n}]$, and let P be any probability measure on $\mathbb{R}$. Describe the random variables on $(\Omega, \mathcal{F}_n, P)$.

1.5. Show that the following two classes are fields, but not sigma-fields.

    (a) All finite subsets of $\mathbb{R}$ together with their complements.

    (b) All finite unions of intervals in $\mathbb{R}$ of the form $(a, b]$, $(-\infty, a]$, and (b, ∞).

1.6. Let $\mathcal{C}$ be the class of all subsets of $\mathbb{R}$ which are either countable or are the complement of a countable set. Show that $\mathcal{C}$ is a σ-field.

1.7. Finish the proof of Proposition 1.10 by showing that if B and B_n are events, then $B_n \downarrow B \Longrightarrow P(B_n) \downarrow P(B)$.

1.3. The Intuition

Think of a probability space as the mathematical description of an experiment. The experiment could be, for example, tossing a coin, rolling dice,

taking a number in a lottery, or getting up in the morning, opening a window, and checking the weather. In each case, there is a certain amount of randomness, or unpredictability in the experiment. The end result is called the *outcome* of the experiment.

To describe this mathematically, start with what we observe: the outcome. Think of Ω as the set of all possible outcomes of the experiment: each point of Ω represents an outcome. An event is a set of outcomes. The probability measure gives the probability of events. For instance, the probability that the outcome falls in a given event A is $P\{A\}$. The collection of all events is the σ-field $\mathcal{F}$.

In short, the probability space $(\Omega, \mathcal{F}, P)$ can be thought of as a mathematical description of an experiment: the elements of Ω are the possible outcomes, the elements of $\mathcal{F}$ are the events, and P gives us the probabilities of all the events—and, in particular, the probabilities of the outcomes.

We should emphasize that there is no need to tie a probability space to a particular experiment. That is just for the intuition. Indeed, Ω can be any set, $\mathcal{F}$ can be any σ-field of subsets of Ω, and P can be any probability measure on $\mathcal{F}$. Together, they form a probability space.

Example 1.13.1. Suppose the experiment is a roll of a single die. A die is a cube whose six faces each have a different number of spots, one through six. Take the the outcome to be a number from one to six. Thus $\Omega = \{1, 2, 3, 4, 5, 6\}$, and the σ-field $\mathcal{F}$ is all subsets of Ω. To get the probability measure, we note that if the die is well-made, the six sides are identical except for the number of spots on them. No one side can be more probable than another. Thus they must all have the same probability: $P\{1\} = P\{2\} = \cdots = P\{6\}$. The probabilities add up to one, so that $P\{1\} = P\{2\} = \cdots = P\{6\} = 1/6$.

Typical events would be $\{1, 3, 5\} = $ "odd"; $\{3, 4, 5, 6\} = $ "greater than two". Notice that each individual outcome is itself an event[3]. The probabilities are easily computed by simply counting the number of outcomes; for example $P\{2, 4, 6\} = P\{2\} + P\{4\} + P\{6\} = 1/2$.

Example 1.13.2. If the experiment is tossing a coin, the outcomes are heads and tails, and we can take $\Omega = \{H, T\}$. If the coin is well-balanced, there is no effective difference between the two sides, and the probabilities of heads and tails should be the same—$P\{H\} = P\{T\} = 1/2$. The σ-field $\mathcal{F}$ is the set of all subsets, which, in this case, is so simple that we can list it: $\emptyset$, $\{H\}$, $\{T\}$, and Ω, having probabilities zero, one-half, one-half, and one, respectively.

---

[3]Technically, the event is the set whose only element is the outcome.

Example 1.13.3. Equally Likely Outcomes. Many games of chance are designed to provide equally likely outcomes. In that case, the probability of each outcome is one over the total number of outcomes: if $|A|$ denotes the number of elements in A, the probability of each outcome is $1/|\Omega|$. This leads directly to the following formula for probabilities.

Proposition 1.14. *Suppose $|\Omega|$ is finite. If all outcomes are equally likely, then the probability of any event A is*

$$P\{A\} = \frac{|A|}{|\Omega|}.$$

Example 1.14.1. Roll two dice. In order to tell them apart, color one red and one green. Suppose they are well-made, essentially perfect cubes. Each die has a number of pips painted on each side, with the number of pips going from one to six. (Assume that the paint and the color of the dice make no difference to the outcome.) The outcome for each die is the number on the top face when the die comes to rest. We can represent the outcome by an ordered pair (m, n), where m is the number showing on the red die, and n the number showing on the green die.

Let the sample space Ω be the space of outcomes $\{(m, n), \ m, n = 1, \ldots, 6\}$. Then Ω has 36 points. To find the probability measure P, we note that since the dice are perfectly symmetric, no number can be more likely than any other. In fact, no pair of numbers can be more likely than any other pair. To see this, note that we can change any outcome into any other merely by repainting the faces: to replace $(2, 5)$ by $(1, 4)$, just interchange the numbers 1 and 2 on the red die, and 5 and 4 on the green. Repainting the dice doesn't change their probabilities—the paint is too light to make a difference. We conclude, purely from symmetry considerations, that all ordered pairs are equally likely. There are 36 outcomes, so that the probability of any event A is $|A|/36$. For instance, the event that the sum of the two dice is a four consists of ordered pairs $(1, 3)$, $(2, 2)$, $(3, 1)$ so that $P\{\text{ sum is 4 }\} = 3/36 = 1/12$.

Example 1.14.2. Choose a number at random from the interval $[0, 1]$. To choose at random implies, intuitively, that no number is more likely than any other. As there are uncountably many numbers in $[0, 1]$, the probability of each is zero, so we had better rephrase this: the probability that the number falls in any sub-interval $[a, a + h]$ is the same as the probability it falls in any other sub-interval of the same length. A moments reflection suggests that the probability must be proportional to the length of the interval, and, since the whole interval has length one, we conclude that the probability of falling in a sub-interval of $[0, 1]$ is actually equal to its length.

Therefore, we take the sample space Ω to be $[0,1]$ itself: the space of all outcomes. We will take the σ-field to be $\mathcal{B}(0,1)$, the σ-field of Borel[4] subsets of $[0,1]$. The probability measure is $P\{dx\} = dx$, $x \in [0,1]$, that is, Lebesgue measure. So for any event A, $P\{A\}$ is the length of A and—most importantly—$P\{A\}$ is the probability that our randomly-chosen number lies in A.

Thus $([0,1], \mathcal{B}(0,1), P)$ is a probability space, where $P\{dx\} = dx$ is Lebesgue measure.

Note. The previous examples concerned finite sample spaces, and the countable additivity of the probability measure was automatic. However, Example 1.14.2 is far from finite, and the fact that Lebesgue measure on $[0,1]$ is a probability measure on the Borel sets is far from trivial. While it certainly has total mass one, the fact that it is countably additive is deep. We will accept it here, and refer the reader to a text in measure theory, e.g. [**31**], §3.3, for its proof.

Remark 1.15. The probability space for an experiment is not unique. For example, the sample space for a single coin toss can be $\Omega = \{H, T\}$. But we could also use the phase-space representation mentioned in the Preface, which would make Ω a subset of twelve-dimensional Euclidean space! We indicated there that any probability measure sufficiently "smeared out" would work. Another possibility would be to take Ω to be the unit interval $(0,1)$, where any point $x \in (0, 1/2)$ corresponds to heads, and any $x \in [1/2, 1)$ corresponds to tails. In this case, $\mathcal{F}$ could be any σ-field of subsets of $(0,1)$ which contained $(0, 1/2)$ and $[1/2, 1)$, and P could be any probability for which $P\{(0, 1/2)\} = P\{[1/2, 1)\} = 1/2$.

We (perhaps too often) think mathematics is neat and tidy, and some might find this plurality of probability spaces off-putting. However, it is useful in practice: it allows us to choose the probability space which works best *in the particular circumstances,* or to not choose one at all; we do not always have to specify the probability space. It is usually enough to know that it is there if we need it.

1.3.1. Symmetry. The probabilities in the preceding examples were all derived from symmetry considerations. In each case, the possible outcomes were indistinguishable except by their labels. ("Label" just means "name" here. The two sides of a coin are labeled by the embossed figures we call heads and tails, dice are labeled by the number of pips on each side, cards are labeled by the pictures printed on their faces, and lottery tickets are labeled by the number printed on them.)

---

[4]We could equally well take the σ-field of Lebesgue sets, which is larger than $\mathcal{B}$, but it turns out that we cannot take the σ-field of *all* subsets.

This is no coincidence. In fact, it is about the only situation in which we can confidently assign probabilities by inspection. But luckily, while nature is not always obliging enough to divide itself into equally-likely pieces, one can, through art and science, start with the equally-likely case and then determine the probabilities in ever-more complex situations. Which is what the subject is about.

The idea of symmetry applies to events, not just outcomes. Consider a physical experiment with finitely or countably many outcomes, labeled in some convenient fashion.

Symmetry principle. *If two events are indistinguishable except for the way the outcomes are labeled, they are equally likely.*

"Indistinguishable" refers to the physical events in the experiment, not to the mathematical model. The principle says that if we can't tell the difference between one event and another except by asking for the names—the labels—of the individual outcomes, they must have the same probability. For example, roll a die and consider the events "even" and "odd", i.e., $\{2,4,6\}$ and $\{1,3,5\}$. If we physically renumber the faces of the die, so that we interchange n and $7-n$ on each face, so that $1 \leftrightarrow 6$, $2 \leftrightarrow 5$, and $3 \leftrightarrow 4$, then the events "even" and "odd" are interchanged. The symmetry principle says that the two events must have the same probability. (So does the very easy calculation which tells us that both have probability $1/2$.)

Let us illustrate this graphically. Recall the glasses used to view 3-D movies. They have one lens colored red, the other colored blue. One lens filters out blue colors, the other filters out red. Take a blank die. Put n red spots, and $7-n$ blue spots on each face, where n goes from one to six. Put on the 3-D glasses, and look at the die with one eye closed. If the left eye is closed, say, you look through the red lens—which filters out red—and you only see the blue spots. If the right eye is closed, you see the red. Roll the die and close one eye: suppose you see an even number of spots. Now open that eye and close the other: you now see an odd number. So the events "even" and "odd" must have the same probability, since, physically, they are the same event! They are just observed differently.

This is essentially J. M. Keynes "Principle of Indifference" [**18**]. To use it in a concrete case, number five identical marbles from one to five, put them in a sack, mix them up, close your eyes, and choose one. The numbers have no influence on the draw since we cannot see them. Then what is the probability that marble one is drawn first? What is the probability that marble five is drawn last? What is the probability that both happen: marble one is first and marble five is last?

Clearly, by symmetry all five marbles are equally likely to be the first choice, so the probability that number one is first is $1/5$. In the second

case, we could count the number of ways to choose all five marbles in order so that number five is last, but that is unnecessary, since it is again clear by symmetry that all five marbles are equally likely to be the last choice. Indeed, the problem is symmetric in the five marbles, so no one can be more likely than any other to be last, and the probability is 1/5 again.

The third probability might appear harder, but it is not. Simply look at the ordered pair (first marble, last marble). By the symmetry principle, all possible orders are equally likely. (To see that (1,5) has the same probability as, say, (2,3), just note that we can renumber the marbles so that marble 1 $\mapsto$ marble 2, and marble 5 $\mapsto$ marble 3, so that $(1,5) \mapsto (2,3)$. Thus $(1,5)$ must have the same probability as $(2,3)$.)

For a last example, let us take another look at the probabilities for a pair of dice. We argued above that we could distinguish the two dice by coloring them differently, and write the outcomes as 36 equally likely *ordered* pairs. So, for instance, the point 12 can only be made by $(6,6)$, so it has probability 1/36, while the point 11 can be made with $(5,6)$ or $(6,5)$, so it has probability 2/36.

But, in fact, a pair of dice are usually of the same color, and essentially indistinguishable. This led Gottfried Leibnitz, the co-inventor (with Newton) of calculus, to write in 1666 that the points 11 and 12 were equally likely. This was because each point can be made in only one way: the 12 by two sixes, the 11 by a six and a five[5].

This contradicts the previous calculation. Rather than laugh at Leibnitz, let us restate what he said. In effect, he thought that the 21 different *combinations* are equally likely, not the 36 different *ordered pairs*. (After some reflection, it may appear less obvious that Leibnitz was wrong. Indeed, we use equally-likely combinations to calculate the probabilities of card hands; why shouldn't we do it for dice too?)

Problem: Why are the 21 combinations not equally likely? (Hint: Is it possible to relabel the dice to change any given combination into any other? If it is, the symmetry principle would tell us Leibnitz was right, and we would have a genuine problem!)

Problems 1.3

1.8. (*i*) Throw 64 dice. What is the probability that their sum is odd? Even? [Hint: Try renumbering *one* of the dice.]

1.9. Construct probability spaces to describe the following experiments.

    (a) A fair coin is tossed three times.

    (b) A single coin is tossed and a single die is rolled.

---

[5]Do you think that Leibnitz gambled a lot?

1.10. Find the most ridiculous probability space you can think of to describe the roll of a single fair die. (But...you must be able to explain how your probability space actually describes the experiment!)

1.11. A ping pong ball is drawn at random from a sack containing 49 balls, numbered from one to 49.

(a) Find the probability that its number is divisible by 2 but not by 3.

(b) Find the probability that its number is divisible by both 2 and 3.

(c) Find the probability that its number is prime.

1.12. A pair of fair dice are rolled. Find the probability that the total number of spots showing is at least 11. Which number has the greatest probability?

1.13. What—if anything—would go wrong if we calculated the probabilities in the roll of a pair of fair dice by taking the *unordered* outcomes to be equally likely, instead of the ordered outcomes? That is, if we regarded, for example, $(3, 4)$ and $(4, 3)$ as the same outcome?

1.14. Roll six independent fair dice.

(a) Find the probability they all show different faces.

(b) Find the probability they form three pairs.

1.15. One card is lost from a standard bridge deck. Subsequently, a card is drawn from the remaining cards. What is the probability that the card drawn is an ace?

1.16. P.R. de Montmort, writing in 1708, posed a problem on the French game Jeu de Boules. The object of the game is to throw balls close to a target ball. The closest ball wins. Suppose two players, A and B have equal skill. Player A tosses one ball. Player B tosses two. What is the probability that player A wins?

1.17. Deal cards one-by-one from a well-shuffled bridge deck until only one suit is left. What is the probability that it is spades?

1.18. A number is chosen at random from the interval $[0, 1]$. It divides the interval into two parts. Find the probability that the longer part is at least twice as long as the shorter part.

1.19. Sicherman dice are like ordinary dice, except that they have different numbers of pips on their faces. One has 1,3,4,5,6,8 on its six faces, and the other has 1,2,2,3,3,4. A pair is thrown. Find the probabilities of rolling the numbers one thru twelve with these three dice. Compare with the usual probabilities.

1.20. 100 passengers wait to board an airplane which has 100 seats. Nassif is the first person to board. He had lost his boarding pass, but they allowed him to board anyway. He chose a seat at random. Each succeeding passenger enters and takes their assigned seat if it is not occupied. If it is occupied, they take an empty seat at random. When the final passenger enters, there is only one empty seat left. What is the probability that it is that passenger's correct seat?

1.21. Suppose A and B are disjoint events with probability $P\{A\} = .4$, $P\{B\} = .5$. Find

(a) $P\{A \cup B\}$

(b) $P\{A^c\}$.

(c) $P\{A^c \cap B\}$.

1.22. Let A and B be events. Show that $P\{A \cup B\} + P\{A \cap B\} = P\{A\} + P\{B\}$.

1.23. Let A, B and C be events. Show that $P\{A \cup B \cup C\}$ equals

$$P\{A\} + P\{B\} + P\{C\} - P\{A \cap B\} - P\{B \cap C\} - P\{C \cap A\} + P\{A \cap B \cap C\}.$$

1.4. Conditional Probability

Conditional probability is legendary for producing surprising results. Let us consider a particular—though unsurprising—example which will lead us to its definition.

Toss a dart at a circular dart board. Let the sample space Ω be the unit disk, which represents the target. The outcome, the point where the dart lands, is a point in the unit disk, and an event is just a subset of the unit disk. There is always a certain amount of randomness in where the dart lands, so the landing-point will be governed by a probability measure, say P: $P\{A\}$ is the probability that the dart lands in the set A.

Let C be the lower half of Ω. C is an event itself. Let A be another event. Suppose we know that C happens, i.e., that the dart lands in the lower half of the board. What is the probability of A given that this happens? In general, it is different from the probability of A. (For if A is in the upper half of Ω, we know the dart cannot have landed in A, so the new probability of A equals zero!)

Let us give this new probability a name and a notation: we call it the **conditional probability of A given C** and we write

$$P\{A \mid C\}.$$

Let us do a thought experiment to determine this. Throw a large number of darts at the target, independently, and use the law of averages[6]. Since this is a thought experiment, there is no limit to the number of darts we can throw. To find the probability of A, we could throw, say, N darts, count the number, say $n(A)$, which land in A, and say that $P\{A\}$ is the limit of $n(A)/N$ as $N \to \infty$.

Suppose we wish to find the probability the dart lands in A, *given that the dart lands in C*. Then we would simply count *only those darts that land in C*—there are $n(C)$ of them—and we would estimate this probability by the proportion of *those* darts which land in A. If we are only counting the darts that land in C, then the darts which land in A and also are counted are the darts which land in $A \cap C$. So we throw N darts, of which $n(C)$

---

[6]We all know the law of averages, and, wonder of wonders, it is actually true. We will prove it later.

Expectations II: The General Case

3.1. From Discrete to Continuous

We have defined the expectation of discrete random variables. We will extend it to the general case by approximating an arbitrary random variable above and below by discrete random variables[1].

Definition 3.1. Let X be a random variable. Its **upper and lower dyadic approximations** $\bar{X}_n$ and $\underline{X}_n$, $n = 0,\ 1, \ldots$ are

$$\bar{X}_n(\omega) = \frac{k+1}{2^n} \quad \text{if } k2^{-n} < X(\omega) \leq (k+1)2^{-n}\,,$$

$$\underline{X}_n(\omega) = \frac{k}{2^n} \quad \text{if } k2^{-n} < X(\omega) \leq (k+1)2^{-n}\,.$$

If ω is fixed, then $\big(\bar{X}_n(\omega)\big)$ is just a sequence of real numbers.

The following three statements are clear from the definition.

(1°) Both $\underline{X}_n$ and $\bar{X}_n$ are discrete random variables.

(2°) $\underline{X}_n(\omega) < X(\omega) \leq \bar{X}_n(\omega)$.

(3°) $\underline{X}_n(\omega) = \bar{X}_n(\omega) - 2^{-n}$.

The next four take a little more thought:

(4°) $\underline{X}_n(\omega) \leq \underline{X}_{n+1}(\omega) \leq \bar{X}_{n+1}(\omega) \leq \bar{X}_n(\omega)$.

---

[1]Our construction of the expectation follows that of the Lebesgue integral, albeit in different language. See Royden [**31**] Chapters 4 and 11 for an account of the classical and general Lebesgue integral respectively. The principle difficulty is to show that Lebesgue measure is countably additive, something we assumed. See [**31**] Chapter 3 for a proof.

($5°$) $\lim_{n\to\infty} \underline{X}_n(\omega) = \lim_{n\to\infty} \bar{X}_n(\omega) = X(\omega)$, and the limits are monotone.

($6°$) The $\bar{X}_n$ and $\underline{X}_n$ are either all integrable, or all non-integrable.

($7°$) If one (and therefore all) of the discrete approximations are integrable,

$$E\{\underline{X}_0\} \le E\{\underline{X}_1\} \le E\{\underline{X}_2\} \le \cdots \le E\{\bar{X}_2\} \le E\{\bar{X}_1\} \le E\{\bar{X}_0\}\,,$$
$$E\{\bar{X}_n\} - E\{\underline{X}_n\} = 2^{-n}\,.$$

To see $4°$, note that if, say, $X(\omega) \in \big(k2^{-n}, (k+1)2^{-n}\big]$ that $\bar{X}_n(\omega) = (k+1)2^{-n}$, and the only two possibilities for $\bar{X}_{n+1}(\omega)$ are $(k+1)2^{-n}$ and $(k+1/2)2^{-n}$, depending on whether $X(\omega) \in \big((k+1/2)2^{-n}, (k+1)2^{-n}\big]$ or $X(\omega) \in \big(k2^{-n}, (k+1/2)2^{-n}\big]$, respectively. In either case, $\bar{X}_{n+1} \le \bar{X}_n$. The same reasoning shows that $\underline{X}_n(\omega) \le \underline{X}_{n+1}(\omega)$.

For each ω the sequences $\bar{X}_n(\omega)$ and $\underline{X}_n(\omega)$ are monotone (by $4°$) and bounded, so that they have limits. The limits are equal by $3°$. Finally, $6°$ follows since all the $\bar{X}_n$ and $\underline{X}_n$ differ by at most one. The last conclusion is now immediate from $4°$ and $3°$.

We have defined integrability for discrete random variables. Let us extend it to general random variables.

Definition 3.2. X is **integrable** if $\bar{X}_0$ is integrable.

The following definition is now inevitable.

Definition 3.3. If X is integrable, then

$$E\{X\} = \lim_{n\to\infty} E\{\bar{X}_n\} = \lim_{n\to\infty} E\{\underline{X}_n\}\,.$$

If X is discrete, its expectation has already been defined by (2.5). To see that the new definition gives the same value, note that the $\bar{X}_n$ and $\underline{X}_n$ are all between $X - 1$ and $X + 1$, so they will be integrable if and only if X is. If X is integrable with expectation L, say, then $E\{\underline{X}_n\} \le L \le E\{\bar{X}_n\} = E\{\underline{X}_n\} + 2^{-n}$. It follows that $\lim_n E\{\bar{X}_n\} = L$, so the two definitions of expectation are indeed equal.

Proposition 2.28 gave the elementary properties of the expectation of discrete random variables. But the expectation of any random variable is the limit of expectations of discrete random variables, so we can essentially take limits in Proposition 2.28. However, the proof is not quite immediate, since we must deal directly with the dyadic approximations.

Before stating the theorem, notice that, as $|\bar{X}_n - X| \le 2^{-n}$,

$$(3.1) \qquad\qquad |\bar{X}_n| - 2^{-n} \le |X| \le |\bar{X}_n| + 2^{-n}\,.$$

Theorem 3.4. *Let X and Y be random variables and let $a \in \mathbb{R}$. Then*

(i) If $X = Y$ a.s., then Y is integrable iff X is, and, if so, $E\{X\} = E\{Y\}$.

(ii) If $|X| \leq |Y|$ a.s. and Y is integrable, then X is integrable. In particular, X is integrable if and only if $|X|$ is.

(iii) If X is integrable, so is aX, and $E\{aX\} = aE\{X\}$.

(iv) If X and Y are integrable, so is $X + Y$, and $E\{X + Y\} = E\{X\} + E\{Y\}$.

(v) If $X \geq 0$ a.s. and X is integrable, then $E\{X\} \geq 0$.

(vi) If X and Y are integrable and $X \leq Y$ a.s., then $E\{X\} \leq E\{Y\}$.

Proof. These conclusions are all known in the discrete case. We can *almost* just say that they all follow by taking limits. There is a slight technical difficulty: integrability is defined in terms of dyadic approximations of X, so we have to work with them. Notice that once we have shown (i), we can ignore the "a.s."s in the succeeding parts.

(i) If $X = Y$ a.s., then $\bar{X}_n = \bar{Y}_n$ a.s., and $E\{|\bar{X}_n|\} = \sum_j |j| 2^{-n} P\{\bar{X}_n = j 2^{-n}\} = \sum_j |j| 2^{-n} P\{\bar{Y}_n = j 2^{-n}\} = E\{|\bar{Y}_n|\}$. Thus $\bar{X}_n$, and hence X, is integrable iff $\bar{Y}_n$, and hence Y, is. If so, the same calculation without absolute value signs shows that $E\{\bar{X}_n\} = E\{\bar{Y}_n\}$. The result follows by letting $n \to \infty$.

(ii) Apply (3.1) with $n = 0$ to X and Y: $|\bar{X}_0| - 1 \leq |X| \leq |Y| \leq |\bar{Y}_0| + 1$. The right side is integrable by hypothesis. By Proposition 2.28, the discrete random variable $\bar{X}_0$ is also integrable. Then X is integrable by definition.

(iii) $|aX| = |a||X| \leq |a|(|\bar{X}_0| + 1)$. Since the right-hand side is integrable, so is $|aX|$. To see the integrals are equal, let $Z = aX$. Then

$$|\bar{Z}_n - a\bar{X}_n| \leq |\bar{Z}_n - Z| + |Z - aX| + |aX - a\bar{X}_n| \leq 2^{-n} + 0 + |a|2^{-n}.$$

Thus $|E\{\bar{Z}_n\} - aE\{\bar{X}_n\}| \leq (|a| + 1)2^{-n} \to 0$. It follows that $E\{aX\} = \lim_n E\{\bar{Z}_n\} = a \lim_n E\{\bar{X}_n\} = aE\{X\}$.

(iv) Let $Z = X + Y$, and apply (3.1) with $n = 0$:

$$|\bar{Z}_0| \leq |Z| + 1 \leq |X| + |Y| + 1 \leq |\bar{X}_0| + |\bar{Y}_0| + 3.$$

The right-hand side is integrable since X and Y are, so $|\bar{Z}_0|$, and hence Z, is integrable.

To check equality of the integrals, write the difference $|\bar{Z}_n - (\bar{X}_n + \bar{Y}_n)| \leq |Z_n - Z| + |Z - (X + Y)| + |X - \bar{X}_n| + |Y - \bar{Y}_n| \leq 3/2^n$. This goes to zero, so that the expectation of Z_n has the same limit as the sum of the expectations of $\bar{X}_n$ and $\bar{Y}_n$, which gives the desired equality.

(v) This is clear since $X \geq 0 \implies \bar{X}_n \geq 0$.

(vi) This follows from (iv), (iii), and the fact that $Y - X \geq 0$. □

If X and Y are integrable, it does not follow that their product XY is. (If, for instance, $Y \equiv X$, $XY = X^2$, and the integrability of X does not imply integrability of X^2.) However, if they are independent, it is true. Moreover, we have:

Theorem 3.5. *Let X and Y be independent random variables. If both X and Y are integrable, so is XY, and*

$$(3.2) \qquad\qquad E\{X\,Y\} = E\{X\}\,E\{Y\}\,.$$

Proof. Suppose first that X and Y are discrete with possible values (x_i) and (y_j), respectively. Then

$$
\begin{aligned}
E\{|X\,Y|\} &= \sum_{i,j} |x_i|\,|y_j| P\{X = x_i,\ Y = y_j\} \\
&= \sum_{i,j} |x_i|\,|y_j| P\{X = x_i\} P\{\,Y = y_j\} \\
&= \sum_{i} |x_i| P\{X = x_i\} \sum_{j} |y_j| P\{Y = y_j\} \\
&= E\{|X|\}\,E\{|Y|\}\,,
\end{aligned}
$$

where the change of order in the summation is justified since the summands are positive. Both X and Y are integrable, so this is finite. Thus XY is integrable. Now remove the absolute value signs. The series converges absolutely, so that the terms can be rearranged, and the same calculation (sans absolute value signs) gives (3.2).

Next, if X and Y are integrable, so are the dyadic approximations $\bar{X}_n$ and $\bar{Y}_n$, and, as $\bar{X}_n - X$ and $\bar{Y}_n - Y$ are both positive and less than 2^{-n},

$$
\begin{aligned}
E\{|\bar{X}_n\bar{Y}_n - XY|\} &\leq E\{|\bar{X}_n|\,|\bar{Y}_n - Y|\} + E\{|Y|\,|\bar{X}_n - X)|\} \\
&\leq 2^{-n}\big(E\{|\bar{X}_n|\} + E\{|Y|\}\big) \\
&\leq 2^{-n}\big(2 + E\{|\bar{X}_0|\} + E\{|\bar{Y}_0|\} \to 0
\end{aligned}
$$

where we have used the facts that $|\bar{X}_n| \leq 1 + |\bar{X}_0|$ and $|Y| \leq 1 + |\bar{Y}_0|$. In particular, since $\bar{X}_n\bar{Y}_n$ is integrable, so is XY. Therefore

$$
E\{X\,Y\} = \lim_n E\{\bar{X}_n\,\bar{Y}_n\} = \lim_n E\{\bar{X}_n\}\,E\{\bar{Y}_n\} = E\{X\}\,E\{Y\}\,.
$$

□

Corollary 3.6. *If X and Y are independent random variables with finite variances, then*

$$
\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)\,.
$$

Proof. $E\{(X+Y)^2\}-E\{X+Y\}^2 = E\{X^2\}+2E\{XY\}+E\{Y\}^2-(E\{X\}+E\{Y\})^2 = E\{X^2\} - E\{X\}^2 + E\{Y^2\} - E\{Y\}^2 = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y)\,.$ □

Example 3.6.1. Generating Functions. We have seen probability generating functions and moment generating functions, and we shall soon see a third type, characteristic functions. Let X be a random variable. The general form of these generating functions is this:

$$(3.3) \qquad\qquad G(t) = E\{e^{g(t)X}\}\,.$$

For the moment generating function, $g(t) = t$, for the probability generating function, $g(t) = \log t$, and for the characteristic function, $g(t) = it$. They have several things in common: they are power series in t whose coefficients have important connections with the distribution of X. Perhaps more important is that, leaving existence questions aside, they actually determine the distribution of X. The third common factor, which we will establish here, is that they work well with sums of independent random variables.

Proposition 3.7. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with generating functions $G_{X_1}(t)$, $G_{X_2}(t), \ldots, G_{X_n}(t)$. Suppose these exist at t. Then the generating function for the sum $X_1 + \cdots + X_n$ exists at t, and*

$$G_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} G_{X_i}(t)\,.$$

Proof. $E\{e^{g(t)(X_1+\cdots+X_n)}\} = E\{\prod_{i=1}^{n} e^{g(t)X_i}\}$. The random variables $e^{g(t)X_i}$ are independent, since the X_i are, and, by hypothesis, they are integrable. By Theorem 3.5 their product is also integrable, and its expectation is the product of the expectations, so $G_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} E\{e^{g(t)X_i}\} = \prod_{i=1}^{n} G_{X_i}(t)$, as claimed. □

Example 3.7.1. Application to Analysis. Consider the probability space

$$\big([0,1], \mathcal{B}, P\big)\,,$$

where $\mathcal{B}$ is the Borel sets and P is Lebesgue measure: $P(dx) = dx$.

Let g be a continuous function on $[0,1]$. Then g is a random variable! (In fact, as we saw in Exercise 2.2, any Borel function of a random variable is a random variable, and continuous functions are Borel.) It is bounded, so it is integrable and $E\{g\}$ exists. Let us compare the expectation with something familiar, the Riemann integral $\int_0^1 g(x)\,dx$. We claim that $\int_0^1 g(x)\,dx = E\{g\}$. To see why, write the integral as a limit of Riemann sums.

$$(3.4) \qquad\qquad \int_0^1 g(x)\,dx = \lim_{n\to\infty} \sum_{k=1}^{n} g(k/n)\frac{1}{n}\,.$$

Let $A_0 = [0, 1/n]$ and set $A_k = (k/n, (k+1)/n]$ for $k \geq 1$. Define $g_n(x) = g(k/n)$ if $x \in A_k$. In other words,

$$g_n(x) = \sum_{k=0}^{n-1} g(k/n) I_{A_k}(x).$$

Thus g_n is not only a step-function, it is a discrete random variable as well. Moreover, g_n converges uniformly to g, and its expectation is

$$E\{g_n\} = \sum_{n=0}^{n-1} g(k/n) P\{A_n\} = \sum_{n=0}^{n-1} g(k/n) \frac{1}{n}.$$

But this is the Riemann sum in (3.4)! The convergence is uniform so $E\{g\} = \lim_n E\{g_n\}$. Moreover, the Riemann integral is the limit of its Riemann sums, which proves the following proposition.

Proposition 3.8. *Let g be continuous on $[0, 1]$. Then $\int_0^1 g(x)\,dx = E\{g\}$.*

Note. This works for highly discontinuous g, too, such as the indicator function of the rationals, which, itself, is not even Riemann integrable. Still, we can compute its expectation. Indeed, let (r_n) be an enumeration of the rationals, and put $g(x) = I_{\mathbb{Q}} = \sum_n I_{\{r_n\}}(x)$. Then $E\{g\} = \sum_n P\{r_n\} = 0$, since, as P is Lebesgue measure, the probability of any singleton is zero. So this is a genuine extension of the Riemann integral. In fact, it **is** the Lebesgue integral.

To make this formal, if g is a Borel function on $[0, 1]$ (and therefore a random variable) we say it is *Lebesgue integrable* if $E\{|g|\} < \infty$, and define $\int_0^1 g(x)\,dx \stackrel{\text{def}}{=} E\{g\}$.

This is almost, but not quite the usual Lebesgue integral, which integrates Lebesgue measurable functions, a larger class than the Borel functions. But we can simply replace the Borel sets $\mathcal{B}$ by the Lebesgue sets $\mathcal{L}$ above. Then Lebesgue measurable functions are random variables on the probability space $([0, 1], \mathcal{L}, P)$, and the expectation gives the classical Lebesgue integral.

The restriction to $[0, 1]$ was purely for ease of exposition. The same idea works for any finite interval $[a, b]$. Indeed, normalized Lebesgue measure $P(dx) = dx/(b - a)$ is a probability measure on $([a, b], \mathcal{B})$, a Borel function g on $[a, b]$ is a random variable on $([a, b], \mathcal{B}, P)$ and, taking the normalization into account, $\int_a^b g(x)\,dx = (b - a)E\{g\}$.

Extended Real Values*. Let X be a positive random variable. We have defined $E\{X\}$ if X is integrable. If X is not integrable, we say $E\{X\} = \infty$.

So far, our random variables have had finite values. When we deal with limits, we will meet random variables with infinite values, and we have to be able to handle them.

Let X be a random variable with values in the extended interval $[0, \infty]$. If $P\{X = \infty\} > 0$, we say $E\{X\} = \infty$. If $P\{X = \infty\} = 0$, define $\hat{X}$ by

$$\hat{X}(\omega) = \begin{cases} X(\omega) & \text{if } X(\omega) < \infty, \\ 0 & \text{if } X(\omega) = \infty. \end{cases}$$

$\hat{X}$ equals X a.e., and it has finite values, so we define

(3.5) $$E\{X\} \stackrel{\text{def}}{=} E\{\hat{X}\}.$$

Intuitively, events of probability zero do not happen, and the values of a random variable on a set of probability zero are irrelevant. In the extreme case where $P\{\Lambda\} = 0$ and $X = \infty \cdot I_\Lambda$, the usual formula gives us $E\{X\} = \infty \cdot P\{\Lambda\} = \infty \cdot 0$. This is indeterminate, but our definition (3.5) says that it equals zero. In effect, we have decreed that infinity times zero is zero! It turns out that this is needed to make the theory work. Consider, for example, the sequence of random variables $X_n = nI_\Lambda$, $n = 1, 2, \ldots$, where $P\{\Lambda\} = 0$. These are finite-valued, so $E\{X_n\} = 0$, and $X_n \uparrow X$. If we wish—and we do[2]—to have $E\{X_n\} \uparrow E\{X\}$, we must have $E\{X\} = 0$.

For a general random variable X, let $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0)$ be the positive and negative parts of X. Both are positive, and $X = X^+ - X^-$. We then define

$$E\{X\} = E\{X^+\} - E\{X^-\},$$

whenever we can make sense of the right-hand side. That is, $\infty - \text{finite} = \infty$, $\text{finite} - \infty = -\infty$, and $\infty - \infty$ is indeterminate. In the last case, we simply say that the integral does not exist, even in the extended sense. In all other cases, we can give $E\{X\}$ a (possibly infinite) value.

3.2. The Expectation as an Integral

The use of the word "integrable" suggests that the expectation is directly connected with an integral in some way. Example 3.7.1 shows that this is true on the unit interval, where the expectation actually gives the Lebesgue integral. But the truth is even more general: the expectation *is* an integral. Indeed, we can write

$$E\{X\} = \int_\Omega X(\omega) \, P(d\omega), \quad \text{or simply} \quad E\{X\} = \int X \, dP,$$

---

[2]See the Monotone Convergence Theorem, Theorem 4.12.

where the right-hand side is a classical Lebesgue integral with respect to the measure P.

This is not the only connection between expectations and integrals. There is a second one, involving the distribution of X, which leads to the general form of the law of the unconscious statistician. Recall that (see Exercise 2.2) if X is a random variable and f is a Borel function, then $f(X)$ is also a random variable. The class of Borel functions includes most functions we can think of. For instance, a continuous function is Borel: if f is continuous, then $\{x : f(x) < \lambda\}$ is open, and therefore a Borel set.

Let X be a random variable on a probability space $(\Omega, \mathcal{F}, P)$, with distribution function F and distribution μ. As we saw in Proposition 2.10, μ is a probability measure on $\mathbb{R}$ and $(\mathbb{R}, \mathcal{B}, \mu)$ is also a probability space, where $\mathcal{B}$ is the σ-field of Borel sets. A random variable on this space is a real-valued function f on $\mathbb{R}$. Technically, to be a random variable, the function f must satisfy one mild condition: $\{x : f(x) < \lambda\} \in \mathcal{B}$ for all x. This means that f must be a Borel function. So, on this probability space, random variables are Borel functions.

This is of more than passing interest, since it means that we have a ready-made integration theory. The integral of f is just its expectation. Let us use E^{μ} for the expectation on $(\mathbb{R}, \mathcal{B}, \mu)$. Then we define an integral with respect to the distribution function:

$$(3.6) \qquad \int_{-\infty}^{\infty} f(x)\, dF(x) \overset{\text{def}}{=} E^{\mu}\{f\}.$$

This is a Lebesgue-Stieltjes integral. There are other ways of writing it[3]. We use the distribution function F instead of the distribution μ since it leads naturally to the probability density. The point is that we do not have to reconstruct the integral: it is already defined.

To see why the notation $\int dF$ makes sense, consider what happens when we integrate over an interval, say $(a, b]$. Certainly the integral of dF over this should be $F(b) - F(a)$, and it is: to see that, just choose $f = I_{(a,b]}$ and note that $E^{\mu}\{f\} = \mu\big((a, b]\big)$; but $\mu\big((a, b]\big) = F(b) - F(a)$. This in turn is equal to $\int_{a}^{b} dF(x) = \int f(x)\, dx$.

This brings us to the law of the unconscious statistician.

---

[3]T. J. Stieltjes defined this to generalize the Riemann integral in 1894. It was later extended to Borel (even Lebesgue) functions using the ideas of Lebesgue. Other notations for this are $\int_{\mathbb{R}} f(x)\, \mu(dx)$ or simply $\int f\, d\mu$. In this notation, it would be called a Lebesgue integral. While the ideas behind the two forms of integral are different, the integrals themselves are, in the end, the same.

Theorem 3.9. *Let X be a random variable with distribution function F, and let g be a positive Borel function. Then*

(3.7)
$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x)\, dF(x)\,.$$

In particular, $E\{X\} = \int_{-\infty}^{\infty} x\, dF(x)\,.$

Proof. Consider g as a random variable on $(\mathbb{R}, \mathcal{B}, \mu)$. We must show that the expectation of g—that is, $E^{\mu}\{g\}$—equals the expectation $E\{g(X)\}$ of $g(X)$ on $(\Omega, \mathcal{F}, P)$. Recall the dyadic approximation of g: $\bar{g}_n(x) = (k+1)2^{-n}$ if $k2^{-n} < g(x) \le (k+1)2^{-n}$. Let $\Lambda_k = \{x : \bar{g}_n(x) = k2^{-n}\} \subset \mathbb{R}$ and $A_k = \{X \in \Lambda_k\} \subset \Omega$. Thus $A_k = \{\bar{g}_n(X) = k2^{-n}\}$, and $P\{A_k\} = \mu(\Lambda_k)$. Then

$$
\begin{aligned}
E\{\bar{g}_n(X)\} &= E\Big\{\sum_{k=1}^{\infty} k2^{-n} I_{A_k}\Big\} \\
&= \sum_{k=1}^{\infty} k2^{-n} P\{A_k\} \\
&= \sum_{k=1}^{\infty} k2^{-n} \mu(\Lambda_k) \\
&= E^{\mu}\{\bar{g}_n\}\,.
\end{aligned}
$$

If we let $Z = g(X)$, then $\bar{Z}_n = \bar{g}_n(X)$. Let $n \to \infty$ on both sides; $E\{\bar{g}_n(X)\} \to E\{g(X)\}$ and $E^{\mu}\{\bar{g}_n\} \to E^{\mu}\{g\} \equiv \int g(x)\, dF(x)\,.$ $\square$

Remark 3.10. We have assumed g is positive, so that the integrals always have a value, though it might be infinite.

Corollary 3.11. *If X has a density f and if $g \ge 0$ is Borel measurable, then*

(3.8)
$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x)f(x)\, dx\,.$$

In particular, $E\{X\} = \int_{-\infty}^{\infty} x f(x)\, dx\,.$

Remark 3.12. If f and g are bounded and continuous, the distribution function F is differentiable, and integrating (3.7) by parts gives (3.8).

The general proof of Corollary 3.11 follows from Exercise 3.1 below.

Remark 3.13. If a random variable X has a density f, then by (3.8), its moment generating function is

$$M(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\, dx\,.$$

Modulo a change of sign of θ, this is the bi-lateral (i.e., 2-sided) Laplace transform of f. More generally, even if there is no density, we can write

$$M(\theta) = \int_{-\infty}^{\infty} e^{\theta x} \, dF(x) \,,$$

so that the moment generating function is the bilateral Laplace transform of the distribution of X.

The importance of these remarks is that there is an inversion formula for the Laplace transform, so that two functions having the same Laplace transform must be equal a.e. Thus, modulo questions of existence[4], the moment generating function determines the distribution.

Problems 3.2

3.1. Let μ be an absolutely continuous probability distribution with distribution function F and density f. For any Borel set Λ and $a < b$, let $\nu(\Lambda \cap [a, b]) = \int_a^b I_\Lambda(x) f(x) \, dx$. Show that this defines a probability measure ν on the Borel sets of $\mathbb{R}$, and that $\nu = \mu$. Conclude that for positive Borel functions g, $\int_a^b g(x) \, dF(x) = \int_a^b g(x) \, f(x) \, dx$.
[Hint: Let $\mathcal{G}$ be the class of sets on which $\nu = \mu$. Then intervals of the form $(a, b]$ are in $\mathcal{G}$.]

3.2. Let X be a random variable and let $\bar{X}_n$ and $\underline{X}_n$ be the upper and lower dyadic approximations to X. Show that $\bar{X}_n$ is the smallest random variable with values in $2^{-n}\mathbb{Z}$ which is greater than or equal to X, and that $\underline{X}_n$ is the greatest random variable with values in $2^{-n}\mathbb{Z}$ which is strictly smaller than X.

3.3. Let $X \geq 0$ be an integrable random variable. Show that if $E\{X\} = 0$, then $X = 0$ a.e.

3.4. Let X be a positive random variable with distribution function F. Show that

$$(3.9) \qquad\qquad E\{X\} = \int_0^{\infty} (1 - F(x)) \, dx \,.$$

Conclude that if $c > 0$, then $E\{X \wedge c\} = \int_0^c (1 - F(x)) \, dx$.

3.5. (Continuation.) Show that if X is an integrable random variable, then $\lim_{x \to \infty} x P\{|X| > x\} = 0$.
[Hint: Try integration by parts.]

3.6. Let $X \geq 0$ be a random variable. Show that for any $M > 0$ that

$$M \sum_{k=1}^{\infty} P\{X > kM\} \leq E\{X\} \leq M \sum_{k=0}^{\infty} P\{X > kM\} \,.$$

---

[4]These questions limit its utility, however. In practice, one replaces the Laplace transform by the Fourier transform, where existence is not a problem. This gives what it is called the characteristic function. We will meet it in Section 6.1.

3.3. Some Moment Inequalities

Chebyshev's Inequality.

Let us give several well-known inequalities involving moments. The first one bounds the probability that a random variable takes on large values. This is a very easy, but also very useful result, Chebyshev's inequality[5].

Theorem 3.14 (Chebyshev's Inequality). *Let $p > 0$, $\lambda > 0$, and let X be a random variable. Then*

$$(3.10) \qquad P\{|X| \geq \lambda\} \leq \frac{1}{\lambda^p} E\{|X|^p\}.$$

Proof. $P\{|X| \geq \lambda\} = P\{|X|^p \geq \lambda^p\} = \int_{\{|X|^p \geq \lambda^p\}} 1\, dP$. But $1 \leq |X|^p/\lambda^p$ on the set $\{|X|^p \geq \lambda^p\}$, so this is

$$\leq \int_{\{|X|^p \geq \lambda^p\}} \frac{|X|^p}{\lambda^p}\, dP \leq \int_\Omega \frac{|X|^p}{\lambda^p}\, dP = \frac{1}{\lambda^p} E\{|X|\}.$$

$\square$

If we replace X by $X - E\{X\}$, and take $p = 2$, we get

Corollary 3.15. $P\{|X - E\{X\}| \geq \lambda\} \leq \dfrac{1}{\lambda^2} \mathrm{Var}\,(X).$

Let X and Y be square-integrable random variables, which is another way of saying that X^2 and Y^2 are integrable. Here is a useful and venerable inequality, called—depending on the country—the Schwarz inequality, the Cauchy-Schwarz inequality, or the Cauchy-Schwarz-Buniakovski inequality[6].

Theorem 3.16 (Schwarz Inequality). *Suppose X and Y are square integrable. Then*

$$(3.11) \qquad E\{XY\} \leq \sqrt{E\{X^2\}}\sqrt{E\{Y^2\}}.$$

There is equality if and only if either X or Y vanishes a.s., or if $X = \lambda Y$ for some constant λ.

Proof. In spite of Doob's opinion, the proof is rather nice. If either of X or Y vanishes almost surely, (3.11) holds with equality. Suppose neither does and let $\lambda \geq 0$. Note that

$$0 \leq E\{(X - \lambda Y)^2\} = E\{X^2\} - 2\lambda E\{XY\} + E\{Y^2\},$$

---

[5]There are a number of similar inequalities which follow from (3.10) by simple substitutions for X. We call them all Chebyshev's inequality. However, many authors call (3.10) with $p = 1$ Markov's inequality, and reserve "Chebyshev's Inequality" for its corollary.

[6]When presenting this in his classes, J.L. Doob would remark that he didn't know why they were fighting about it, it was trivial anyway.

so

$$2\lambda E\{XY\} \leq E\{X^2\} + \lambda^2 E\{Y^2\}\,.$$

Divide by 2λ:

(3.12) $$E\{XY\} \leq \frac{1}{2\lambda}E\{X^2\} + \frac{\lambda}{2}E\{Y^2\}\,.$$

Set $\lambda = \dfrac{\sqrt{E\{X^2\}}}{\sqrt{E\{Y^2\}}}$ to get (3.11).

If neither X nor Y vanishes a.e., then there is equality iff there is equality in (3.12), which happens iff $E\{(X - \lambda Y)^2\} = 0$, i.e., iff $X = \lambda Y$ a.e.   □

3.4. Convex Functions and Jensen's Inequality

Definition 3.17. A function ϕ defined on an open interval (a, b) is **convex** if for each $a < x < y < b$ and each $\lambda \in [0, 1]$,

$$\phi\big(\lambda x + (1 - \lambda)y\big) \leq \lambda\phi(x) + (1 - \lambda)\phi(y)\,.$$

The geometrical interpretation of this is that the chord of the graph between x and y lies above the graph. It implies that the graph of ϕ is roughly cup-shaped. Convex functions are sometimes called "concave up", which is a good way to think about them. The negative of a convex function is sometimes called "concave down" or just concave.
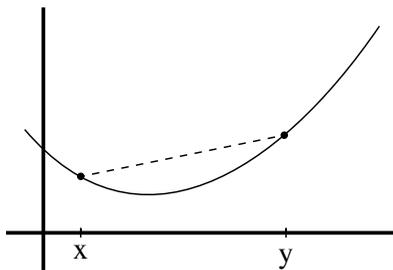

Figure 1. The graph of ϕ lies under the chord.

Functions like $|x|$, x^2, e^x, and e^{-x} are convex on the line, $1/x$ is convex on $(0, \infty)$ and $1/x^2$ is convex on both $(0, \infty)$ and $(-\infty, 0)$ (but not on $(-\infty, \infty)$!) Linear functions are both convex and concave.

Roughly speaking, convex functions generalize functions with positive second derivatives. However, they need not be differentiable: $|x|$ is convex, for instance. But they are *almost* differentiable, as the following shows.

(ii) is an exercise in number theory which we leave to the reader, and (iii) is a consequence of (ii).   □

7.4. Stopping Times

Consider a gambler playing roulette at a casino. He has decided that he will place one single bet on his lucky number seventeen, but he has not decided when to make it. We have no idea exactly how he will decide when to bet, but we can say a few general things. The first is that he cannot look into the future: certainly, he would like to place a bet on the $n + 1^{st}$ spin of the wheel only if seventeen will come up then, but he has to bet *before* the wheel is spun, so he cannot know whether or not his number will come up. All he knows when he makes the bet are the results of previous plays, and he has to base his decision on this. Of course, he may have other information, too, but let us ignore that for the sake of discussion.

Let $X_1, X_2, \ldots$ be the results of the plays. The gambler has to base his decision whether or not to bet on the $n + 1^{st}$ spin on the observed values of $X_1, \ldots, X_n$. That is all the information he has. So his strategy for betting must be this: there is a set $A \subset \mathbb{R}^n$ such that he will bet on the $n + 1^{st}$ spin if the observed values of the n-tuple $(X_1, \ldots, X_n)$ fall into the set A. Otherwise he will wait till later to bet.

The particular choice of A is the gambler's business, not ours. Our interest is this: it is the most general rule for deciding when to bet.

Let us focus on the time T that he places his bet. Then T is a random variable, and $T = n \Longleftrightarrow (X_1, \ldots, X_n) \in A$.

Note that strategies like "bet on seventeen immediately after the zero or double zero appears" and "bet on the fifth turn of the wheel" are of this form.

It is convenient to allow the possibility that the gambler never bets. If he never bets, we say $T = \infty$. Thus the betting time T is characterized by the following:

1° T is a random variable taking values in $\{1, 2, \ldots, \infty\}$;

2° for each n there is $A_n \subset \mathbb{R}^n$ such that $T = n \Longleftrightarrow (X_1, \ldots, X_n) \in A_n$.

We can streamline this criterion by introducing σ-fields. Let $\mathcal{F}_n$ be the class of all events of the form $\{\omega : (X_1(\omega), \ldots, X_n(\omega)) \in B\}$, where B is a Borel set in $\mathbb{R}^n$. (That is, $\mathcal{F}_n = \sigma\{X_1, \ldots, X_n\}$.)

Then T, the time of the bet, is random and satisfies:

- T takes values in $\{1, 2, \ldots, \infty\}$ and
- for each n, $\{T = n\} \in \mathcal{F}_n$.

Remark 7.20. $\mathcal{F}_n$ is a σ-field and for each n, $\mathcal{F}_n \subset \mathcal{F}_{n+1}$.

Indeed, we already know that $\mathcal{F}_n$ is a σ-field, and that if $\Lambda \in \mathcal{F}_n$, then there is a Borel set A for which $\Lambda = \{(X_1, \ldots, X_n) \in A\}$. We can also write Λ in terms of $(X_1, X_2, \ldots, X_{n+1})$, for $\Lambda = \{(X_1, \ldots, X_n, X_{n+1}) \in A \times \mathbb{R}\}$, which is in $\mathcal{F}_{n+1}$. Therefore, $\mathcal{F}_n \subset \mathcal{F}_{n+1}$.

Such times are called *stopping times*[5].

The gambler may also have extra information. Recalling that σ-fields are information in the raw, we can take this into account by adding the extra information to the σ-fields $\mathcal{F}_n$. Thus $\mathcal{F}_n$ may be larger than $\sigma\{X_1, \ldots, X_n\}$.

Definition 7.21. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ be a sequence of sub-sigma-fields of $\mathcal{F}$. A random variable T with values in $0, 1, 2, \ldots, \infty$ is a **stopping time** if for every $n = 0, 1, 2, \ldots$, the set $\{T = n\} \in \mathcal{F}_n$.

Here is an equivalent definition which is often used.

Proposition 7.22. *A random variable with values in $0, 1, 2, \ldots, \infty$ is a stopping time relative to the sigma-fields $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ if and only if $\{T \le n\} \in \mathcal{F}_n$ for $n = 0, 1, 2, \ldots$.*

Proof. If T is a stopping time, then $\{T \le n\} = \bigcup_{m=0}^{n}\{T = m\}$. If $m \le n$, then $\{T = m\} \in \mathcal{F}_m \subset \mathcal{F}_n$. Thus $\{T \le n\}$ is a finite union of elements of the σ-field $\mathcal{F}_n$, and is therefore in $\mathcal{F}_n$ itself. Conversely, if $\{T \le n\} \in \mathcal{F}_n$ for all n, then $\{T = n\} = \{T \le n\} - \{T \le n = 1\}$. By hypothesis, $\{T \le n\} \in \mathcal{F}_n$ and $\{T \le n - 1\} \in \mathcal{F}_{n-1} \subset \mathcal{F}_n$, so the difference, $\{T = n\} \in \mathcal{F}_n$, and T is a stopping time. □

Stopping times are important: they are times when interesting things can happen. And they will appear frequently in the sequel. Let us say more about the intuition behind them.

In the example above, the stopping time is a time at which the gambler can make a decision. The important thing is that the gambler can recognize the time when it arrives. That is exactly what the condition "$\{T = n\} \in \mathcal{F}_n$" means. For the gambler's total knowledge at time n is $\mathcal{F}_n$, and, since $\{T = n\}$ (and $\{T \ne n\}$ as well) are part of that knowledge, he knows whether or not T arrives then.

---

[5]These were introduced by A. A. Markov for Markov chains, and they are often called Markov times. They came up in more general settings, e.g., to show that gambling systems do not work and in particular, to answer the question of whether a gambler could improve the odds by opting to stop playing the game at a certain time, hence the name "stopping time". They are also called "optional times".

So, intuitively, a stopping time is simply a (possibly random) time that can be recognized when it arrives. It is not necessary to wait, or to look into the future to divine it. We can recognize it in real time.

In fact, stopping times are so pervasive that even times which seem to be fixed are actually random. For instance, the time of my 21st birthday—which seemed important at the time—depends on the date of my birth, which is definitely a random time. (A glance at the biology of reproduction makes it clear how frighteningly random that time was[6].)

Exercise 7.12. Show that if T is a stopping time, then for each n, the sets $\{T = n\}$, $\{T < n\}$, $\{T \geq n\}$ and $\{T > n\}$ are in $\mathcal{F}_n$.

Example 7.22.1. (i) The constant random variable $T \equiv n$ is a stopping time.

Indeed $\{T \leq m\}$ is either Ω or $\emptyset$, depending on whether $n \leq m$ or not. Both are in $\mathcal{F}_m$.

(ii) Let $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$, $n = 0, 1, 2, \ldots$, and let B be a Borel subset of the state space. Let $T_B = \inf\{n \geq 0 : X_n \in B\}$, where we make the convention that the infimum of the empty set is ∞. Then T_B is a stopping time, called the **first hitting time of B.**

For, $\{T_B = n\} = \{X_0 \in B^c, \ldots, X_{n-1} \in B^c, \ X_n \in B\} \in \mathcal{F}_n$.

(iii) If T is a stopping time and $m \geq 0$ is an integer, then $T + m$ is a stopping time.

Indeed, if $n \geq m$, $\{T + m = n\} = \{T = n - m\} \in \mathcal{F}_{n-m} \subset \mathcal{F}_n$, because the σ-fields increase. And if $n < m$, $\{T + m = n\} = \emptyset$, which is also in $\mathcal{F}_n$.

(iv) If S and T are stopping times, so are $S \wedge T = \min\{S, T\}$ and $S \vee T = \max\{S, T\}$. In particular, if m is a positive integer, $T \wedge m$ and $T \vee n$ are stopping times.

To see this, note that $\{S \wedge T \leq n\} = \{S \leq n\} \cup \{T \leq n\}$ and $S \vee T \leq n\} = \{S \leq n\} \cap \{T \leq n\}$. By Exercise 7.12, all four of these sets are in $\mathcal{F}_n$, hence so are $\{S \wedge T \leq n\}$ and $\{S \vee T = n\}$.

We will need to define the process at a stopping time, or, more generally, at any random time.

Definition 7.23. Let R be a random variable taking values in $0, 1, 2 \ldots$. Then X_R is the random variable defined by

$$X_R(\omega) = X_{R(\omega)}(\omega).$$

---

[6]Even the dates of the Gregorian calendar are random: biology aside, the official date of Christ's birth was not decided until 525 A.D., and was based on the best estimate they could make at the time. Given the incomplete and occasionally contradictory historical records available then, it could easily be off by several years. This raises an interesting point: 1 A.D. was not a stopping time. 526 A.D. was.

That is, X_R is defined by

$$X_R(\omega) = X_n(\omega) \quad \text{if } \omega \in \{R = n\} .$$

There is something to verify about this definition, namely that X_R actually is a random variable, i.e., that for any Borel set A, $\{X_R \in A\} \in \mathcal{F}$. But this is clear since $\{X_R \in A\} = \bigcup_{n=0}^{\infty} \{R = n\} \cap \{X_n \in A\}$, and both $\{R = n\}$ and $\{X_n \in A\}$ are in $\mathcal{F}$, hence so is their intersection.

Note: In particular, we will want to use X_T for a stopping time T, but there is a potential problem: we allow infinite values for stopping times, but X_T is not defined on the set $\{T = \infty\}$. However, we will be careful to confine ourselves to sets on which T is finite. (Notice that if $X_T = i$, then T must be finite[7].)

We have just argued that stopping times are really like fixed times, even though they are random. The σ-field $\mathcal{F}_n$ represents the information available—that is, everything we know—up to the present time n. We need to extend this idea to stopping times. That is, we need a σ-field to represent the history of the process up to a stopping time T. We are tempted to say that it is all the events determined by $X_0, \ldots, X_T$, but...T is random, so the number of elements in the sequence is not fixed. However, if $T = n$, the sequence is $X_0, \ldots, X_n$, and so the event is determined by $X_0, \ldots, X_n$. This leads us to define the past before T as follows.

Definition 7.24. Let T be a stopping time relative to the σ-fields $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$. We define $\boldsymbol{\mathcal{F}_T}$ by

$$\mathcal{F}_T \stackrel{\text{def}}{=} \left\{ \Lambda \in \mathcal{F} : \Lambda \cap \{T \leq n\} \in \mathcal{F}_n,\ n = 0, 1, 2, \ldots \right\} .$$

We call this "the past before T".

Remark 7.25. We can replace "$\{T \leq n\}$" by "$\{T = n\}$" in the definition; that is,

$$\mathcal{F}_T \stackrel{\text{def}}{=} \left\{ \Lambda \in \mathcal{F} : \Lambda \cap \{T = n\} \in \mathcal{F}_n,\ n = 0, 1, 2, \ldots \right\} .$$

The equivalence of these two characterizations of $\mathcal{F}_T$ follows from the equations

$$\begin{cases} \Lambda \cap \{T \leq n\} = \displaystyle\bigcup_{j=0}^{n} \Lambda \cap \{T = j\} , \\ \Lambda \cap \{T = n\} = \Lambda \cap \{T \leq n\} - \Lambda \cap \{T \leq n - 1\} . \end{cases}$$

---

[7]If we want to be formal, we can take a point ∂ which is *not* in the state space, and arbitrarily define $X_\infty = \partial$. With the obvious extension of the transition probabilities (e.g., set $P_{\partial\partial} = 1$), X becomes a Markov chain on $\mathbb{N} \cup \{\partial\}$, and X_T is well-defined even on $\{T = \infty\}$.

This gives us a choice of conditions to verify when we want to show[8] something is in $\mathcal{F}_T$.

The following proposition collects some elementary properties of $\mathcal{F}_T$.

Proposition 7.26. *Let S and T be stopping times. Then:*

(i) $\mathcal{F}_T$ *is a σ-field.*

(ii) *Both T and $X_T I_{\{T<\infty\}}$ are $\mathcal{F}_T$-measurable.*

(iii) *If $S \leq T$, then $\mathcal{F}_S \subset \mathcal{F}_T$.*

Proof. We leave (i) for the exercises.

(ii) Note that X_T is not defined on $\{T = \infty\}$, which is why we restrict ourselves to the set where T is finite. It is enough to prove that $\{X_T = j\} \in \mathcal{F}_T$. To see this, just test it with $\{T = n\}$: for $n = 0, 1, 2, , \ldots$

$$\{X_T = j\} \cap \{T = n\} = \underbrace{\{X_n = j\}}_{\in \mathcal{F}_n} \cap \underbrace{\{T = n\}}_{\in \mathcal{F}_n} \in \mathcal{F}_n \, .$$

To see T is $\mathcal{F}_T$-measurable, it is enough to show $\{T \leq m\} \in \mathcal{F}_T$ for all m. But for $m, n = 0, 1, 2, \ldots$

$$\{T \leq m\} \cap \{T \leq n\} = \{T \leq m \wedge n\} \in \mathcal{F}_{m \wedge n} \subset \mathcal{F}_n \, ,$$

so $\{T \leq m\}$ is indeed in $\mathcal{F}_T\}$.

(iii) Let $\Lambda \in \mathcal{F}_S$. To show $\Lambda \in \mathcal{F}_T$, test it with $\{T \leq n\}$. As $S \leq T$,

$$\Lambda \cap \{T \leq n\} = \underbrace{\Lambda \cap \{S \leq n\}}_{\in \mathcal{F}_n \ (\text{def of } \mathcal{F}_S)} \cap \underbrace{\{T \leq n\}}_{\in \mathcal{F}_n} \in \mathcal{F}_n \, ,$$

for $n = 0, 1, 2, \ldots$, hence $\Lambda \in \mathcal{F}_T$. $\square$

7.5. The Strong Markov Property

The Markov property of a Markov chain $X_0, X_1, \ldots$ can be expressed: "For each n, the process $X_n, X_{n+1}, \ldots$ is a Markov chain with the same transition probabilities as $X_0, X_1, \ldots$; moreover, given X_n, it is conditionally independent of $X_0, \ldots, X_n$."

It is remarkable that this remains true if we replace the fixed time n by any finite stopping time T whatsoever.

Let $\{X_n, \ n = 0, 1, 2, \ldots\}$ be a Markov chain with transition probability matrix $\mathbb{P} = (P_{ij})$, and let $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$. We can rephrase the Markov property in terms of $\mathcal{F}_n$: for $n \geq 1$, $\Lambda \in \mathcal{F}_n$, and states i, j, (7.1) becomes:

(7.9) $\Lambda \in \mathcal{F}_n \Longrightarrow P\{X_{n+1} = j \mid X_n = i, \Lambda\} = P_{ij} \, ,$

---

[8]This is a minor point in the present setting. It will be more important when we look at continuous parameter processes, where the situation is more delicate.

[Hint: First assume P is a probability measure and F is separable. Find increasing partitions $\mathcal{P}_n \subset \mathcal{P}_{n+1} \subset \ldots$ which generate $\mathcal{F}$, and put $X_n = Q(\Lambda)/P\{\Lambda\}$ if $\Lambda \in \mathcal{P}_n$. Show that (X_n) is a martingale. Use the previous problem to show it is uniformly integrable. Now remove the assumption that P is a probability measure, and use exercise 9.33 to remove the separability assumption.]

9.36. Let (X_n) be a sequence of random variables tending a.s. to a random variable X_∞. Let $\mathcal{F}_n$ be a sequence of σ-fields. Prove the following result, called Hunt's Lemma, which extends Theorem 9.38 to the case where both the random variables and the σ-fields vary.

> **Theorem** (Hunt's Lemma). Suppose that the X_n are dominated by an integrable random variable Z.
> (a) If $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$, then $\lim_{n \to \infty} E\{X_n \mid \mathcal{F}_n\} = E\{X_\infty \mid \bigvee_n \mathcal{F}_n\}$ a.s.
> (b) If $F_1 \supset \mathcal{F}_2 \supset \ldots$, then $\lim_{n \to \infty} E\{X_n \mid \mathcal{F}_n\} = E\{X_\infty \mid \bigcap_n \mathcal{F}_n\}$ a.s.

9.6. Financial Mathematics I: The Martingale Connection★

The aim of this section is to show the relation between financial markets and martingales. We will prove a very weak version of what is called the fundamental theorem of asset pricing, namely that in a complete market with no arbitrage, market prices are expectations relative to a certain "martingale measure", and that prices evolve as martingales[10].

Before we get to it, however, we need to learn something of the basic structure of a financial market, of buying, selling, and the idea of arbitrage. This will take time, but once done, the probability enters quickly, and from a surprising direction. The important probabilistic structure is not the one we first think it is.

Let us prepare the path with a quick look at a familiar financial market where the same thing happens, but in a less mysterious way: the race track[11].

A Day at the Races. Race tracks express probabilities in terms of odds: if the odds on a horse are five to two, that means that there are two chances the horse wins, five that it doesn't—seven chances in all—so the probability of winning is, in theory at least, $2/7$. More importantly, the odds give the payoff of bets: a bettor who bets \$2 at odds of five to two will win \$5 if the horse wins and lose the \$2 if it doesn't.

But the odds the bettor sees on the track's tote board are not the actual probabilities of winning. They are synthetic probabilities determined by the

---

[10]The stronger version of this theorem for discrete parameters is the Dalang-Morton-Willinger Theorem [**10**]. It was first discovered by Harrison, Kreps, and Pliska in 1979–1981, and extended by numerous authors. Delbaen and Schachermeyer proved the general continuous-parameter version in 1994. See [**11**] for a complete account.

[11]Look ahead to Remark 9.61 to see why this is not—quite—a joke.

amount of money bet on each horse. The race track sets the odds. Its aim
is to assure that, no matter which horse wins, the track will make a profit
of ρ percent (called the *take*) of the total money bet. To see how this works,
consider a race between three horses, call them[12] A, B, and C. Suppose the
there is a total of \$400 bet on A, \$600 bet on B, and \$1000 bet on C, with a
take of fifteen percent, a typical figure. The total betting pool is \$2000, and
the track wants fifteen percent of that for a \$300 profit. If A wins at odds
of x to one, the track keeps the \$1600 bet on B and C and pays $400x$ to
those who bet on A. This leads to $1600 - 400x = 300$, so $x = 13/4$. Thus A
starts at odds of $13/4$ to one, or thirteen to four. Similar calculations show
that the odds on B are eleven to six and the odds on C are seven to ten.
These odds imply that A, B, and C have probabilities of winning of $4/17$,
$6/17$, and $10/17$, respectively.

These probabilities are entirely synthetic, of course. They have little to
do with the true winning probabilities, for they are derived purely from the
betting... and they do not even sum to one.

Let us generalize this.

Exercise 9.37. Horses $H_1, H_2, \ldots, H_n$ are entered in a race. Race goers have bet
\$$M_i$ on horse H_i, $i = 1, \ldots, n$.

(a) Find odds on each of the n horses which assure that, after paying all winning
bets, the track has $\rho\%$ of the total betting pool left over for profit.

(b) What probabilities p_i are implied by these odds?

(c) Find $\sum_{i=1}^{n} p_i$. Is it less than, equal to, or larger than one?

Exercise 9.38. (Continuation.) Suppose that the take is $\rho = 0$.

(a) Show that $\sum_{i=1}^{n} p_i = 1$.

(b) Suppose that for each i, p_i is the true probability that horse H_i wins. Show
that all bets are fair.

Exercise 9.39. (Continuation.) Suppose the track runs m races in a day, and
that $\rho = 0$ for all of them. Suppose further that the races are independent and
the implied probabilities are actually the true probabilities of winning each race.
A gambler has an initial fortune F_0. As long as money remains, the gambler will
bet on none, one, or several horses to win in each race. The decision of whether
or not to bet on a given race, and how much, may depend on the results of the
previous races, the remaining fortune, and on the gambler's hunches. Let F_j be the
gambler's fortune after the jth race. Show that $F_0, F_1, \ldots, F_m$ is a martingale.

Of course, the true probabilities of winning are not those the bettors see
on the tote board. (Those don't even add up to one, since the track has a
non-zero take.) However, the important fact for what follows is that if the
race track's take is zero, there do exist probabilities—synthetic probabilities,

---

[12]Or Alydar, Bold Ruler, and Citation, if you prefer.

to be sure, but probabilities nonetheless—with respect to which all bets are fair and all gamblers' fortunes are martingales.

9.6.1. The Basic Financial Market. Financial markets in the real world contain multitudes of stocks, bonds, goods, services, contracts, and taxes, which can be bought and sold, delivered or refused, fulfilled, and paid or evaded under the rules, regulations, and customs which apply. They also involve people, which means that any mathematical hypotheses we might make about them will soon be violated.

We will not attempt to analyze anything so complicated. We will look instead at a simple idealized market containing a single stock and a single bond. In the real world, stocks represent shares of a company and have an intrinsic value which may change with the fortunes of the company. In our market, though, a stock is simply something to be bought and sold. Its price varies randomly with time. The bond acts like a bank: buying a bond is equivalent to putting money in a savings account—it earns interest—and selling a bond is equivalent to withdrawing money, or borrowing if there is not enough in the account. (We talk about bonds instead of bank accounts in order to use the metaphor of buying and selling for all items in the market.) The difference between a stock and a bond is that the stock is random, but the bond is not: if we know its price today, we know its price for all future times.

Simple as the market is, we shall have to explore its financial workings fairly deeply before we arrive at the probability. This is because there is another element implicit in the market: contracts or promises which depend on the stock and the bond. These are by no means simple. They are called **derivatives** because their value is derived[13] from the stock and bond values. A typical derivative would be a contract to deliver one share of stock one month hence. This is called a *future*. A *stock option* is another. A typical option would be the promise to sell to the buyer one share of stock one month hence at a fixed price K—*if* the buyer still wishes to buy it; but the buyer is not obligated to buy. If the stock price is less than K, the buyer would refuse the option: the stock would be cheaper at the market price.

The stock option cannot lose money, and may make a profit, so it must be worth something. But what? This is the kind of problem we will address.

Rules of the Market. The market consists of:

• A stock. The price of one share of the stock at time t is S_t: $\{S_t,\ t = 0, 1, 2, \dots\}$ is a stochastic process.

---

[13]"Derivative" in this sense has nothing to do with differentiation.

• A bond. Its value at time t is $B_0 e^{rt}$, where $r \geq 0$ is the interest rate[14]. The bond is deterministic and the interest rate is constant.

• Some or possibly all derivatives based on the stock and bond.

Furthermore:

• All items in the market can be traded according to the market trading rules.

The trading rules are simple:

• All items in the market may be freely bought and sold at the market price in any quantity desired.

That is, there is a price. If one wishes to buy at that price, there will be a seller. If one wishes to sell, there will be a buyer, and there is no limit, large or small, to the amount bought and sold.

• Short-selling is allowed.

This means that it is not necessary to own a stock or bond in order to sell it. "Short selling" is the selling of an item that one does not own. This is permitted in real-world markets, though it is normally regulated and subject to credit limits and margin requirements. It is akin to borrowing the stock in order to sell it. As with other borrowing, there is an obligation to return it later.

• There are no transaction costs or hidden expenses.

• Promises will be kept: there is no default, so contracts once made will be honored.

These are the basic rules of trading. There is one more important property, called the "no-arbitrage hypothesis" which has to do with the structure of the market, not the trading rules.

• There is no arbitrage in the market.

An **arbitrage** is a deal which makes a profit with strictly positive probability, but cannot lose money, and which requires zero net investment. When a possible arbitrage presents itself, we speak of an **arbitrage opportunity.** According to the hypothesis, this cannot happen. As they say, "There is no free lunch."

For example, suppose person A offers to sell something for a dollar and person B offers to buy it for two. Then person C could buy it from A and immediately sell it to B, making a sure profit of one dollar. The no-arbitrage hypothesis rules out this kind of transaction[15]. It is basically a hypothesis

---

[14]Note that interest is compounded continuously.

[15]It is also ruled out by the hypotheses, which say that quantities are bought and sold at a unique market price.

on the efficiency of the market. Market prices are supposed to adjust so quickly that the opportunity would disappear before C found out about it.

Remark 9.44. The no-arbitrage requirement actually forces prices to be unique. Note that if an item is sold, the buyer and seller must have agreed on the price. But if one person, say A, is willing to sell an item for the price P, and person B is willing to buy it for the same price, then person C cannot buy or sell it for a different price. Indeed, the reader can verify that if C offers to buy it for a strictly greater amount than P, or sell it for strictly less, this offers person D the possibility of an arbitrage. If C wishes to sell it for less than P, or buy it for more, there will be no takers, since D can deal with A or B at a better price. This means that there is a unique price—which we will call the *market price*—for both buying and selling, and everyone, buyer and seller alike, agrees on it. There are no transaction fees or hidden costs.

Future Values. The no-arbitrage hypothesis determines prices to a great extent, via what are called "hedging arguments", which compare alternate ways of achieving the same aims.

Let us see how it determines two things: the future value of the stock and the future value of a dollar.

Consider a question about stock value: What is the value today of the stock at a later time t? Let us rephrase that. Consider a contract which guarantees to provide one share of stock at time t. How much is that contract worth today, at $t = 0$?

The answer is that the contract is worth S_0, the value of the stock today. Any other value provides an arbitrage opportunity.

Indeed, let P be the contract price. There are two ways to have the stock at time t. One can either buy one share today for S_0 and keep it, or one can buy the contract today for P and collect the stock at time t.

Suppose first that $P < S_0$. We can construct an arbitrage as follows:

(a) Sell one share of stock today for S_0. (We are allowed to sell a stock without owning it—this is "short-selling".)

(b) With the money obtained from the sale, buy the contract for P. This leaves $S_0 - P > 0$. Invest that in a bond.

At this point, the total investment is zero. We own the contract, have a bond worth $S_0 - P > 0$, and we are short one share of stock which we must replace later.

(c) At time t, collect the share of stock from the contract and sell it for S_t, its price at time t. Use that money to buy back the share we sold short, also at S_t. This cancels out, but we still have the bond, which is

now worth $(P - S_0)e^{rt} > 0$. When we sell that, we have made a strictly positive profit from zero investment. This is an arbitrage, which contradicts the no-arbitrage hypothesis. Thus the price cannot be strictly less than S_0.

If $P > S_0$, reverse the strategy. Sell the contract for P—that is, agree to provide a share of stock at time t—and buy the stock itself for S_0. Invest the remaining $P - S_0$ in a bond. The total investment is zero. We have one share of stock, a bond, and the obligation to give one share of stock to the contract's buyer when t comes.

At time t, we provide the stock as agreed. Then we sell the bond for a strictly positive profit on zero investment. This is an arbitrage, which cannot happen. Thus P cannot be strictly greater than S_0. The only possible price is S_0.

Even without the no-arbitrage rule, there is a powerful reason to use this price. An investor who doesn't is open to an arbitrage. There is no limit on the amount of a transaction, so being on the wrong side of an arbitrage can cost a *lot* of money.

As the second example, consider the value of a dollar bill: how much is it worth some time in the future? (Cash money does not enter our model except as a way to keep score. We will see why after we answer this question.)

Let us rephrase the question. Consider a contract which guarantees to provide a one-dollar bill at a time t in the future. How much is that contract worth today?

We want to have a dollar at hand at time t. Of course we can do this by buying the dollar today, and keeping it until t. That will cost one dollar. But there is a better way.

We claim that the contract to supply one dollar at time t in the future is worth e^{-rt} dollars today.

Let P be the price of the contract. Suppose $P < e^{-rt}$. Then we do the following.

(a) Sell P worth of bonds. (This is short-selling.)

(b) Use that money to buy the contract for P.

The two transactions cancel, and there is no net investment. But now at time t,

(c) collect the dollar the contract promises and

(d) pay off the bonds at their new price, Pe^{rt}.

We now have cleared the transaction, and have $1 - Pe^{rt}$ left over. Since $P < e^{-rt}$, this is strictly positive. We have made a profit, with no risk whatsoever. This is arbitrage, and cannot happen. Thus, $P \geq e^{-rt}$.

If $P > e^{-rt}$, do the opposite: sell the contract for P at $t = 0$ and use the money to buy P worth of bonds. This is again a zero total investment. Then at time t, sell the bonds for Pe^{rt} and deliver the dollar that the contract promised. This leaves $Pe^{rt} - 1$, which is strictly positive because $P > e^{-rt}$. Another arbitrage! Thus P cannot be strictly greater than e^{-rt}. Since it is neither strictly greater nor strictly smaller than e^{-rt}, it must be equal.

Remark 9.45. This says that the value of a dollar bill some time t in the future is e^{-rt}. Its value decreases with time[16]. The prudent investor will never hold money, but will immediately exchange it for bonds[17]. That is why money is not explicitly in the model.

There is another consequence. We have to take inflation into account when comparing the value of items at different times. For example, salaries in the 1950s seem small by modern standards until we realize that today's dollar is worth about eleven cents in 1950s currency.

So the face value of money in the future must be **discounted** to measure it in terms of todays dollars. It is the discounted dollar, not the face-value dollar, that we must use to measure future value. Thus when we want to discuss future values, we should measure them in discounted dollars, not face-value dollars. Another way to express this is to say that future value should be measured in terms of the bond, not the dollar, since the bond has a constant value in discounted dollars.

Derivatives and Value. From now on, we will measure values in terms of todays dollar, or, as it is usually called, the **discounted dollar**. This allows us to compare values from one time to another. The value of the dollar bill decreases with time, as we saw. At time t, it is only worth e^{-rt} in today's dollars. That means that in terms of today's dollar, the stock and bond at time t are worth

$$\hat{S}_t \overset{\text{def}}{=} e^{-rt} S_t \,,$$

$$\hat{B}_t \overset{\text{def}}{=} e^{-rt} B_t \equiv B_0 \,.$$

This is purely a question of units: S_t and $\hat{S}_t$ are the same object expressed in different units. S_t is its value in nominal dollars, $\hat{S}_t$ in discounted dollars. For the rest of the discussion, we will express values in terms of discounted dollars. We will put hats over some discounted quantities such as the stock price, but this is simply as a reminder.

Let $\{\hat{S}_t, \ t \geq 0\}$ be defined on the probability space $(\Omega, \mathcal{F}, P)$. For $t \geq 0$, define

$$\mathcal{F}_t = \sigma\{\hat{S}_s, \ s \leq t\} \,.$$

[16]This is familiar: it is called inflation.
[17]There are no transaction costs in this model, so the exchange costs nothing.

Suppose for simplicity that $\hat{S}_0$ is constant, so that $\mathcal{F}_0$ is trivial. The bond is deterministic, so all the randomness in the market is supplied by the stock, and $\mathcal{F}_t$ represents the total state of knowledge at time t.

Derivatives get their name because their value is derived from values of the stock and the bond, which means that they are actually functions of the of the stock and bond. But the bond is deterministic, so we can regard a derivative as a function of $(\hat{S}_t)$ alone. In short, it is a random variable. Let us make that into a definition:

Definition 9.46. A **derivative X** on $[0, T]$ is a positive $\mathcal{F}_T$-measurable random variable.

If X is a derivative on $[0, T]$, it represents money, but the amount may depend on the behavior of $\hat{S}_t$ in the interim, i.e., on $[0, T]$. The derivative is positive (i.e., non-negative) by convention. Think of X as a contract that promises to provide X dollars at time T. The contract has the final value X, but this value may not be known before time T itself. So its price at an earlier time t is in some sense an estimate of its final value. We will speak of the price of this contract as "the value of X."

We can only speak of the price of derivatives that are freely traded, i.e., those that are part of the market. In real life, stocks are traded on the stock market, the bond is freely available, and some, but not all, derivatives are traded on exchanges. Other derivatives are sold by banks and other financial firms, but not traded, and still others are not traded at all. However, our market is ideal, not real, and, for the purpose of this section, we will assume that it is complete in the sense that *all* derivatives are freely traded. Therefore, they all have a value.

Definition 9.47. Let X be a derivative. Then $\boldsymbol{V_t(X)}$ denotes the **value of $\boldsymbol{X}$ at time $\boldsymbol{t}$**, measured in discounted dollars.

Notice that we have identified a derivative with its payoff, not with the process involved in arriving at the payoff. So if X is a derivative on $[0, T]$, any two schemes which give at the same payoff X at time T have the same value.

We already know something about V_t: if $t < T$, then

$$V_t(\hat{S}_T) = \hat{S}_t, \quad V_t(\hat{B}_T) = \hat{B}_t \equiv \hat{B}_0.$$

The final value of a derivative X on $[0, T]$ is X itself:

$$V_T(X) = X.$$

Moreover, the time-t value of X is known at time t, so $V_t(X)$ is $\mathcal{F}_t$-measurable. (In fact, it is itself a derivative on $[0, t]$: it is the time-t price of a contract to deliver X at time T.)

The basic problem is easy to state: "Find $V_t(X)$ for all derivatives X and all t."

The arbitrage opportunities discussed above are all elementary, in that they involve a finite number of securities at just two different times. To handle more complex dealings, let us introduce the idea of a portfolio. This is simply a list of the investor's holdings. It could include the stock, the bond, cash, and diverse derivatives. We assume the investor deals with a finite or at most countably infinite number of securities, say $X_1, X_2, X_3, \ldots$. Formally:

Definition 9.48. A **portfolio** is a function $\Phi(t) = (\phi_1(t), \phi_2(t), \ldots)$, $t = 0, 1, 2, \ldots$, where $\phi_k(t)$ is $\mathcal{F}_{t-1}$-measurable, $k = 1, 2, \ldots$. The **value** of the portfolio at time t is

$$M_t \overset{def}{=} \sum_{k=1}^{\infty} \phi_k(t) V_t(X_k) \,.$$

Here, $\phi_k(t)$ is the amount of the derivative X_k the investor holds at time t. It can be negative, which corresponds to short-selling. If a security is bought, it is added to the portfolio; if sold, it is subtracted. If the investor buys X_k, for example, then ϕ_k increases. In particular, short-selling something adds a negative amount to the portfolio. The reason that $\Phi(t)$ is $\mathcal{F}_{t-1}$-measurable rather than $\mathcal{F}_t$-measurable is that the investor cannot look into the future: the decision on the portfolio for day t is made after the stock-prices on day $t-1$ are known, but before those on day t are. (Otherwise, the investor could pick only stocks that are sure to go up, an arbitrage.)

If no money is injected into or taken from the portfolio, an increase in some ϕ_k must be balanced by a decrease in other ϕ_j used to pay for it. A portfolio from which no money is injected or extracted is called "self-financing": money used to buy a new security must come from the sale of others, so it does not require an influx of money to run.

The portfolio's value at time $t-1$ is $M_{t-1} = \sum_k \phi_k(t-1) V_{t-1}(X_k)$. Then securities are bought and sold at the old prices to make the new portfolio $\Phi(t)$. Its value—at the old prices—is $\sum_k \phi_k(t) V_{t-1}(X_k)$. No money has been added or removed, so this still equals M_{t-1}. Subtracting, we get

$$(9.13) \qquad \sum_k \big(\phi_k(t) - \phi_k(t-1)\big) V_{t-1}(X_k) = 0 \,, \ t = 1, 2, \ldots \,.$$

Definition 9.49. A portfolio satisfying (9.13) is called **self-financing**.

We can describe arbitrage opportunities in terms of portfolios.

Definition 9.50. A self-financing portfolio $(\Phi(t))$ is called an **arbitrage opportunity on** $[s, t]$ if $M_t - M_s \geq 0$ a.s. and $P\{M_t - M_s > 0\} > 0$.

Market trading rules allow one to convert the entire portfolio into cash and back at any time, so the owner of the portfolio Φ can have $M(s)$ in cash at time s and $M(t)$ in cash at time t, giving a genuinely riskless profit of $M(t) - M(s)$. This turns an arbitrage opportunity into a true arbitrage. So the no-arbitrage hypothesis is equivalent to saying there are no arbitrage opportunities. The self-financing requirement excludes the possibility that the profit comes from an outside investment.

Remark 9.51. Infinite values are financially impractical but mathematically inescapable. While the stock price is always finite, it may be possible to construct a derivative whose value is infinite. Since nobody could afford it, it should be irrelevant.

The St. Petersburg arbitrage[18] reveals a deeper problem. Suppose an investor invests repeatedly using the strategy "Double up after every loss." Suppose that the probability that the stock price goes up in any interval n to $n + 1$ is $p > 0$, and that the different intervals are independent. Let X_n be the derivative which pays \$2 if $S_{n+1} > S_n$, and and which pays zero otherwise. Suppose that X_n costs \$1. The investor buys X_1 for \$1 at time 1. If the stock goes up, she pockets the \$1 profit and stops playing. If not, she buys \$2 worth of X_2. If if the stock goes up, she pockets the \$4, again for a \$1 profit, and stops. Otherwise she doubles the investment again, and so on, until one investment finally pays off, as it must. If the stock falls during the first n intervals and rises on the $n + 1^{\text{st}}$, she has paid out $1 + 2 + \cdots + 2^n = 2^{n+1} - 1$, and, as her final investment was \$$2^n$, she gains \$$2^{n+1}$, for a profit of \$1. Since $p > 0$ and the different investments are independent, she is sure to eventually make a one-dollar profit. This is a genuine arbitrage.

Notice that in order to follow this strategy, the investor must short-sell bonds to buy the X_i; but there is no bound on the amount it might take. Indeed, this strategy requires the possibility of an unbounded amount of short-selling. Let us define:

Definition 9.52. A portfolio has **bounded risk** if there exists N such that for all t, $\sum_k V_t(X_k)^- \leq N$.

Here, $V_t(X_k)^-$ is the negative part of $V_t(X_k)$, i.e., the amount that the portfolio is short on X_k. This effectively limits the amount that can be borrowed. This is no restriction in practice, since N could be the total amount of money in circulation in the world, but it does rule out both the sale of infinite-value securities and the St. Petersburg arbitrage.

---

[18]This is a variant of the famous St. Petersburg game proposed by Nicolas Bernoullli in 1738. Its drawbacks in gambling and finance have been thoroughly explored.

Definition 9.53. There is **no-arbitrage with bounded risk** if no self-financing portfolio with bounded risk is an arbitrage opportunity.

Properties of the Value.

Lemma 9.54. *Let X and Y be derivatives on $[0, T]$ of finite value. If $X \leq Y$ a.s., then $V_0(X) \leq V_0(Y)$. If $X = Y$ a.s., then $V_0(X) = V_0(Y)$.*

Proof. Suppose that this is wrong, and that $V_0(X) > V_0(Y)$. Consider the portfolio that has one unit of Y and is short one unit of X. Its value at $t = 0$ is $M_0 \equiv V_0(Y) - V_0(X) < 0$. At time T, the same portfolio is worth $M_T = V_T(Y) - V_T(X) = Y - X > 0$. This is an arbitrage opportunity, which is a contradiction. This proves the first statement, and the second follows immediately. □

Theorem 9.55. *Let $a \geq 0$, let $0 \leq t < T$, and let X and Y be derivatives on $[0, T]$. Then*

> *(i) $V_0(0) = 0$, $V_0(1) = 1$, and $V_0(X) \geq 0$.*
>
> *(ii) $V_0(X) = 0 \iff X = 0$ a.e.*
>
> *(iii) V_0 is linear: $V_0(X + Y) = V_0(X) + V_0(Y)$ and $V_0(aX) = aV_0(X)$.*
>
> *(iv) If $t \geq 0$, then $V_0(X) = V_0(V_t(X))$.*
>
> *(v) If $A \geq 0$ is bounded and $\mathcal{F}_t$-measurable, then $V_t(AX) = AV_t(X)$.*

Proof. These equation are all true for the same reason: their two sides represent two ways to get the same payoff, and therefore have the same value by Lemma 9.54.

(i) $X \equiv 1$ gives the same payoff as a one-dollar bond, and therefore has the same initial value, namely $V_0(1) = 1$. Similarly, $X \equiv 0$ is equivalent to zero bonds, whose value is zero. Finally, derivatives are positive by convention, so by Lemma 9.54, $0 \leq X \implies 0 = V_0(0) \leq V_0(X)$.

(ii) If $X = 0$ a.e. $V_0(X) = 0$ by (i). Conversely, if $V_0(X) = 0$, the portfolio containing X alone has initial value $M_0 = V_0(X) = 0$ and final value $M_T = V_T(X) = X$. This is an arbitrage opportunity unless $X = 0$ a.s.

(iii) Consider two ways to get $X + Y$ at time T: either buy a contract to deliver $X + Y$ at time T, which has value $V_0(X + Y)$, or buy two separate contracts, one to deliver X, the other to deliver Y, which have values $V_0(X)$ and $V_0(Y)$. They have the same payoff at T, and therefore, by Lemma 9.54, the same value. Thus $V_0(X + Y) = V_0(X) + V_0(Y)$. A similar argument shows that $V_0(aX) = aV_0(X)$.

(iv) Consider two ways of getting X at time T: at time zero either (a) buy a contract to deliver X at time T, (value $V_0(X)$) or (b) buy a contract that will deliver $V_t(X)$ at the intermediate time t. (That is buy a contract

which will, at time t, deliver a second contract. This second contract will provide X at time T.) If it delivers $V_t(X)$ at time t its value must be $V_0\big(V_t(X)\big)$. The payoff is X in both cases, so the two values are equal.

(v) Note that A is known at time t, so from then on, it is possible to deal in A units of a security.

Let $\Lambda_1 \stackrel{def}{=} \{V_t(AX) < AV_t(X)\}$ and $\Lambda_2 \stackrel{def}{=} \{V_t(AX) > AV_t(X)\}$. Consider a portfolio Φ which is initially empty and therefore has initial value zero. It remains empty on $\Omega - \Lambda_1 \cup \Lambda_2$, and changes exactly once, at time t, on $\Lambda_1 \cup \Lambda_2$. On Λ_1, (where $V_t(AX) < AV_t(X)$), from time t on Φ contains $AV_t(X) - V_t(AX)$ in bonds, one unit of AX, and minus A units of X (i.e., A units of X were short-sold.) On Λ_2, after time t Φ contains $V_t(AX) - AV_t(X)$ in bonds, A units of X and minus one unit of AX. The value of Φ at day-t prices is $AV_t(X) - V_t(AX) + V_t(AX) - AV_t(X) = 0$ on Λ_1 and $V_t(AX) - AV_t(X) + AV_t(X) - V_t(AX) = 0$ on Λ_2. Therefore it is self-financing.

Both X and AX are derivatives on $[0, T]$ (A is $\mathcal{F}_t$ measurable and $\mathcal{F}_t \subset \mathcal{F}_T$) so at time T, $V_T(X) = X$ and $V_T(AX) = AX$. The bonds still have their original price. Thus on Λ_1, Φ has the value $M_T \equiv AV_t(X) - V_t(AX) + V_T(AX) - AV_T(X) = AV_t(X) - V_t(AX) > 0$. On Λ_2 its value is $M_T \equiv V_t(AX) - AV_t(X) + AV_T(X) - V_T(AX) = V_t(AX) - AV_t(X) > 0$. It is zero elsewhere, so $M_0 = 0$, $M_T \geq 0$, and $M_T > 0$ on $\Lambda_1 \cup \Lambda_2$. This is an arbitrage unless $P\{\Lambda_1 \cup \Lambda_2\} = 0$. Therefore $V_t(AX) = AV_t(X)$ a.s. $\square$

We need another result, one which doesn't come up in real-world finance: only a philosopher would ask[19] whether or not someone can simultaneously make an infinite number of different transactions. But this is mathematics, not the real world. Someone can. However, we impose a (loose) credit limit: we assume a bounded risk.

Proposition 9.56. *Let $X_1, X_2, \ldots$ be a sequence of derivatives on $[0, T]$ with $\sum_i V_0(X_i) < \infty$. Then*

$$V_0\Big(\sum_{i=1}^{\infty} X_i\Big) = \sum_{i=1}^{\infty} V_0(X_i)\,.$$

Proof. Construct $\sum_i X_i$ as follows. At time zero, short sell $V_0(X_i)$ worth of bonds and buy X_i with the proceeds. Do this for every i. The resulting portfolio contains one unit of each X_i and is short $\sum_i V_0(X_i)$ of bonds. This is finite by hypothesis, so that its value is $M_0 \equiv -\sum_i V_0(X_i) + \sum_i V_0(X_i) = 0$, and the portfolio has bounded risk. At time T, take delivery of the X_i and add them together to make $\sum_i X_i$. Let $P = V_0\big(\sum_{i=1}^{\infty} X_i\big)$. The

---

[19]Practical people wouldn't ask, for the answer is obviously "No." Mathematicians wouldn't ask either, for the answer is obviously "Yes."

portfolio now contains $-\sum_i V_0(X_i)$ of bonds along with $\sum_i X_i$ so its value is $M_T = P - \sum_i V_0(X_i)$, where $P \equiv V_0\big(\sum_i X_i\big)$.

Now $V_0(X)$ is increasing in X, so $P \geq V_0\big(\sum_{i=1}^{n} X_i\big) = \sum_{i=1}^{n} V_0(X_i)$ for each n. Let $n \to \infty$ to see that $P \geq \sum_{i=1}^{\infty} V_0(X_i)$. If the inequality is strict, $M_T > 0$, which is an arbitrage opportunity[20]. Therefore $P = \sum_i V_0(X_i)$. $\quad\square$

No-Arbitrage and Martingales. Now have seen the basic workings of the market, we can connect it to martingales. The connection is made by two theorems. The first one is this.

Theorem 9.57. *Suppose the market consists of the stock, bond and all derivatives. If the market has no arbitrage with bounded risk, then for each T there exists a probability measure Q on $(\Omega, \mathcal{F}_T)$ such that Q is equivalent[21] to P, and if X is a derivative on $[0, T]$, its value is*

$$(9.14) \qquad V_0(X) = \int X\, dQ\,.$$

Proof. Let $T > 0$. For $\Lambda \in \mathcal{F}_T$, define $Q(\Lambda) = V_0(I_\Lambda)$. This makes sense, since I_Λ is a positive $\mathcal{F}_T$-measurable random variable, and therefore a derivative. By Theorem 9.55 Q is a positive finitely additive set function on $\mathcal{F}_T$, and $Q(\Omega) = V_0(1) = 1$. It is even countably additive: if $\Lambda_1, \Lambda_2, \ldots$ are disjoint sets in $\mathcal{F}_T$, then $\sum_i I_{\Lambda_i} \leq 1$, so that $V_0\big(\sum_i I_{\Lambda_i}\big) \leq V_0(1) = 1$. It follows that $\sum_i V_0(\Lambda_i) = V_0\big(\sum_i I_{\Lambda_i}\big) \leq 1$. Proposition 9.56 implies

$$Q\Big(\bigcup_n \Lambda_n\Big) = V_0\Big(\sum_n I_{\Lambda_n}\Big) = \sum_n V_0(I_{\Lambda_n}) = \sum_n Q(\Lambda_n)\,.$$

Thus Q is a measure. To prove (9.14), just follow the construction of the integral, using the linearity of both the integral and the value function. By definition, $V_0(X) = \int X\, dQ$ if X is the indicator function of a measurable set. Therefore (9.14) holds if X is a positive simple function. In general, X, being positive, is the limit of an increasing sequence $X_0, X_1, X_2, \ldots$ of simple functions, with $X_0 \equiv 0$. Then $X = \sum_{n=1}^{\infty}(X_n - X_{n-1})$, and

$$\int X\, dQ = \sum_{n=0}^{\infty} \int (X_n - X_{n-1})\, dQ = \sum_{n=0}^{\infty} V_0(X_n - X_{n-1}) = V_0(X)\,,$$

where the next-to-last equality follows because $X_n - X_{n-1}$ is positive and simple, and the last follows by Proposition 9.56. This proves (9.14).

To see that Q and P are equivalent, apply Theorem 9.55 (ii) to indicator functions: $P\{\Lambda\} = 0 \iff V_0(I_\Lambda) = 0 \iff Q\{\Lambda\} = 0$. $\quad\square$

---

[20]The no-arbitrage argument fails if $P = \infty$: it is too expensive to sell! However, the rules of the market stipulate that buyers and sellers must agree on prices, and there cannot be agreement on an infinite price, since we would be willing to sell it for a finite amount!

[21]We say that Q and P are *equivalent* if for each $\Lambda \in \mathcal{F}_T$, $P\{\Lambda\} = 0 \iff Q\{\Lambda\} = 0$.

Corollary 9.58. *Under the hypotheses of Theorem 9.57, if X is a Q-integrable derivative on $[0, T]$, then for $t \geq 0$,*

(9.15) $$V_t(X) = E^Q\{X \mid \mathcal{F}_t\},$$

where E^Q is the expectation on $(\Omega, \mathcal{F}, Q)$. Thus $\{V_t(X), \mathcal{F}_t, \ t \geq 0\}$ is a martingale.

Proof. Let $\Lambda \in \mathcal{F}_t$. $V_t(X)$ is $\mathcal{F}_t$-measurable. It is also a derivative, so by Theorem 9.55 (iv) and (v),

$$\int_\Lambda V_t(X)\, dQ = \int I_\Lambda V_t(X)\, dQ = \int V_t(I_\Lambda X)\, dQ$$
$$= V_0(V_t(I_\Lambda X)) = V_0(I_\Lambda X)$$
$$= \int_\Lambda X\, dQ\,.$$

But this is true for all $\Lambda \in \mathcal{F}_t$, which identifies $V_t(X)$ as $E^Q\{X \mid \mathcal{F}_t\}$. $\square$

Definition 9.59. A measure Q on $(\Omega, \mathcal{F}_T)$ is an **equivalent martingale measure** if it is equivalent to P (i.e., $\forall\, \Lambda \in \mathcal{F}_T : P\{\Lambda\} = 0 \Longleftrightarrow Q\{\Lambda\} = 0$), and if the discounted stock price $\{\hat{S}_t, \mathcal{F}_t, 0 \leq t \leq T\}$ is a martingale on $(\Omega, \mathcal{F}_T, Q)$.

Let us restate this. If there is no arbitrage in a complete market, then the prices V_0 determine an equivalent martingale measure Q such that, under Q, not only the discounted stock prices $(\hat{S}_t)$, but the discounted prices of all derivatives are martingales. Moreover, once we know the measure Q, we know the prices of all derivatives at all times, for they are just conditional expectations.

Remark 9.60. One might have guessed that prices were expectations, even without the analysis. And it is true. They are. However, they are expectations with respect to **another** probability measure, the martingale measure Q. This is synthetic, derived from the money—the stock and derivative values—not from the actual probability distribution of the stock. Nevertheless, it is fundamental for evaluating worth.

Remark 9.61. It is instructive to revisit the horse-racing example. In one sense, the connection is clear: betting on a horse is equivalent to buying a derivative whose value is zero if the horse loses and is whatever the odds say if it wins. But there is a deeper connection. The tote board quotes odds that insure a profit for the track. In other words, it assures an arbitrage. Setting the track's take to zero removes that, and results in a betting system with no arbitrage. You proved in Exercises 9.37, 9.38 and 9.39 that the quoted odds defined a probability measure which made all gamblers' fortunes into martingales. You produced, in fact, a martingale measure for the race track.

The analogy doesn't end there. In addition to bets on win, place and show, race tracks offer combination bets like the *pick three*, which is a bet on the winners of three successive races. These are clearly derivatives: they depend on the results of the races. Once the martingale measure is established, Theorem 9.62 shows how to price them.

Conversely, a martingale measure generates a system of prices which have no arbitrage.

Theorem 9.62. *Let Q be an equivalent martingale measure and X a Q-integrable derivative. Define*

$$(9.16) \qquad V_t(X) \stackrel{\text{def}}{=} E^Q\{X \mid \mathcal{F}_t\}, \ t \geq 0.$$

This defines a market having no-arbitrage-with-bounded-risk.

Proof. If there is an arbitrage opportunity, there is a self-financing portfolio $\Phi(t)$ whose value M_t satisfies $M_0 = 0$, $M_T \geq 0$, and $P\{M_T > 0\} > 0$. There is a margin limit, so the value of the absolute portfolio must be finite at all times, i.e., $\infty > V_0\big(\sum_k |\phi_k(t)| V_t(X_k)\big) = \sum_k \int |\phi_k(t)| V_t(X_k)\, dQ$. In particular, M_t is Q-integrable for all t. Now

$$M_T = \sum_{j=1}^{T}(M_j - M_{j-1}) = \sum_{j=1}^{T}\sum_{k=1}^{\infty}\big(\phi_k(j)V_j(X_k) - \phi_k(j-1)V_{j-1}(X_k)\big)$$

$$= \sum_j \sum_k \phi_k(j)\big(V_j(X_k) - V_{j-1}(X_k)\big)$$

$$+ \sum_j \sum_k \big(\phi_k(j) - \phi_k(j-1)\big)V_{j-1}(X_k).$$

The portfolio is self-financing, so the last sum vanishes by (9.13). Take the expectation with respect to the martingale measure Q. As $\phi_k(j)$ is $\mathcal{F}_{j-1}$-measurable,

$$E^Q\{M_T\} = \sum_j \sum_k E^Q\big\{\phi_k(j)\big(V_j(X_k) - V_{j-1}(X_k)\big)\big\}$$

$$= \sum_j \sum_k E^Q\big\{\phi_k(j)\, E^Q\{V_j(X_k) - V_{j-1}(X_k) \mid \mathcal{F}_{j-1}\}\big\}.$$

But by (9.16), $j \mapsto V_j(X_k)$ is a Q-martingale, so the inner conditional expectations vanish and $E^Q\{M_T\} = 0$.

Now P and Q are equivalent, and $P\{M_T \geq 0\} = 1$, so $Q\{M_T \geq 0\} = 1$ as well. Then $E^Q\{M_T\} = 0 \Rightarrow Q\{M_T = 0\} = 1$, so $P\{M_T = 0\} = 1$ too. This contradicts the assumption that $P\{M_T > 0\} > 0$. $\qquad\square$

Remark 9.63. This settles a problem that has been lurking in the background: does there exist a consistent set of prices, that is, prices for which there is no arbitrage? This is not at all obvious, but we now see that if there exists an equivalent martingale measure, the answer is "Yes."

The problem of finding the value of derivatives comes down to the problem of finding an equivalent martingale measure, that is, a measure which makes the (discounted) stock price into a martingale. The corresponding (discounted) derivative values are given by (9.16). Theorem 9.62 guarantees that these are consistent market prices: there is no arbitrage. To summarize:

• Every consistent set of market prices corresponds to a martingale measure.

• Every martingale measure corresponds to a consistent set of market prices.

If the martingale measure is unique, then the market prices are uniquely determined and we can say that the no-arbitrage hypothesis determines all the prices. If there is more than one martingale measure, as there may be, then there is more than one set of market prices, and the no-arbitrage hypothesis does not uniquely determine the market.

Example 9.63.1. (A Binary Tree model.) The Cox-Ross-Rubinstein model is a simple model of stock prices which is suitable for computation. For simplicity, we will assume that the interest rate is zero.

Let S_n be the stock price for $n = 0, 1, 2, \ldots$, defined on a probability space $(\Omega, \mathcal{F}, P)$, where $\mathcal{F} = \sigma\{S_n, \ n = 0, 1, 2, \ldots\}$ is the natural σ-field. Suppose that at each time n, S_n can either increase by a factor of $b > 1$, or decrease by a factor of $1/b$, independent of the previous S_j. Let $p = P\{\hat{S}_{n+1} = b\hat{S}_n \mid \mathcal{F}_n\}$ be the probability of increasing.

Thus S_n is a Markov chain on $\{b^n, \ n = 0, \pm 1, \pm 2, \ldots\}$. It will be a martingale if $S_n = E\{S_{n+1} \mid S_n\} = pbS_n + (1-p)S_n/b$, or $p = 1/(b+1)$. Under the martingale measure, then, (S_n) is a Markov chain with transition probabilities $P_{b^n b^{n+1}} = 1/(b+1)$ and $P_{b^n b^{n-1}} = b/(b+1)$. If $S_0 = 1$, this uniquely determines the process modulo its initial distribution. We conclude that the martingale measure is unique, and the market prices are uniquely determined.

To get a feeling for how options are used, consider the plight of an investor who, thinking that the price of a stock is about to go down, has short-sold it, and who now hears a rumor that something is about to happen to make the price go way up. When it comes time to replace the short-sold stock—and that time will surely come—the investor will have to do it at the market price at that time, which may be considerably greater than it is now. (The investor has received today's price, but will have to repay the

stock at the future market price, so if the stock's price triples, the investor will lose his or her investment twice over[22].) Therefore, the investor hedges by purchasing an option to buy the stock at time $n = 3$, say, for its present price. If the price goes down, the investor will make a profit from the short-sold stock, and will not exercise the option. But if the price goes up, he will buy the stock at time $n = 3$ for today's price and use it to replace the short-sold stock. There will be no net loss. So the option is insurance: once it is bought, the investor is insured against loss. Question: How much should the option cost?

Suppose for example that $b = 2$, and that today's price is 1. The stock $\hat{S}$ can either double or halve each day, and there are three days. In the worst case, the stock goes up by a factor of eight, and the investor loses seven times the original investment.

The stock's possible values on the third day are 1/8, 1/2, 2, and 8. If the investor chooses to exercise the option, he must pay 1 for each share, so his profit will be $\hat{S}_3 - 1$. The investor will only exercise the option if it will profit him, which is if $\hat{S}_3 > 1$. So the payoff is $(S_3 - 1)^+$.

Neither the investor nor the option seller knows the actual probabilities but they can compute the option's price without them. They only have to know the transition probabilities under the martingale measure. With $b = 2$, the doubling probability is 1/3, the halving probability 2/3.

Thus the value of the option is $E^Q\{(S_3 - 1)^+\}$, where the expectation is with respect to the martingale measure Q. To calculate this, use binomial probabilities to find that the price is $E^Q\{(\hat{S}_3 - 1)^+\} = (8 - 1)^+(1/27) + (2 - 1)^+(6/27) = 13/27$. So the option costs 13/27 dollars per share, so if the investor owns N shares of stock, the option will cost $13N/27$.

It is useful to think of the stock price evolving on a tree: if its value at time n is s, it can take either of two branches, the upper branch, which leads to $2s$, and the lower branch, which leads to $s/2$. Under the martingale measure Q, the upper branch has probability 1/3 and the lower branch has probability 2/3. The value V_3 of the option at time 3 is known: $V_3(S_3) = (S_3 - 1)^+$. We can compute its initial value by using the transition probabilities to successively determine $V_2(S_2) = E^Q\{V_3 \mid S_2\}$, $V_1(S_1) = E^Q\{V_2 \mid S_1\}$, and finally $V_0 = E^Q\{V_1\}$, where Q is the martingale measure. This is straightforward. If $S_j = s$, then S_{j+1} is either $2s$ or $s/2$ with (conditional) probabilities 1/3 and 2/3, respectively, giving us

$$V_j(s) = \frac{1}{3}V_{j+1}(2s) + \frac{2}{3}V_{j+1}(s/2).$$

---

[22]Short-selling a stock is riskier than simply buying it. The worst that can happen to an investor who buys a stock is that the stock's value goes to zero, and the investment is lost. In short-selling, however, the investor can lose many times the investment.
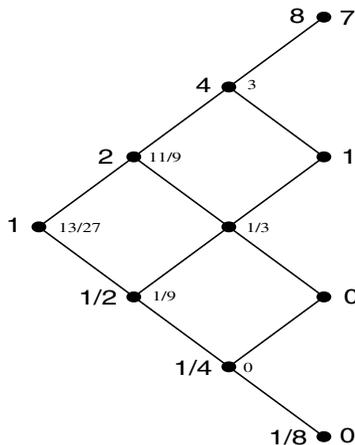
Figure 1. Values of $V_j = E^Q\{(S_3 - 1)^+ \mid \mathcal{F}_j\}$.

If we know the value of a derivative at time n, say, this equation gives us its value at times $n-1$, $n-2$, ..., 1, and 0 successively. See Figure 1.

Example 9.63.2. (Trinomial Tree.) Suppose that there are three possible choices for the stock: $\hat{S}_t$ can either increase by a factor of $b > 1$, decrease by a factor of $1/b$, or stay the same. It is easily seen that there is a whole family of distributions, one for each value of $P\{S_1 = S_0\}$, for which $E\{S_2 \mid \mathcal{F}_1\} = S_1$. Each generates a martingale measure, and each martingale measure leads to a different system of market values. A calculation shows that if $b = 2$, the option in the previous example has values between zero (when $P\{S_1 = S_0\} = 1$) and 13/27 (when $P\{S_1 = S_0\} = 0$).

Notice that the market values of the stock and bond are the same for all martingale measures. Only the values of the derivatives change.

Problems 9.6

9.40. Consider the Cox-Ross-Rubinstein model on the probablity space $(\Omega, \mathcal{F}, P)$ with step size $b > 1$: $P\{S_{n+1} = bS_n\} = p$, where $0 < p < 1$. Let S_n be the stock price, and assume the interest rate is zero. Let Q be the martingale measure.

(a) Show that $(\log S_n)$ is a simple random walk under either P or Q.

(b) Show that if $P\{S_1 = bS_0\} \neq Q\{S_1 = bS_0\}$, that P and Q are *not* equivalent on $\mathcal{F}_\infty = \sigma\{S_n, \ n \geq 0\}$.

[Hint: Find an event which has probability zero under P, probability 1 under Q.]

9.41. Consider the Cox-Ross-Rubinstein model with step-size $b > 1$. Find the value of the derivative which pays \$100 if $S_1 > S_0$ and zero otherwise.

9.42. In the Cox-Ross-Rubinstein model with step-size $b > 1$ and $S_0 = 1$, it is found that a call option which pays $(S_2 - 1)^+$ at time 2 has value 0.25. Find b.

9.43. Let P be the true measure on $(\Omega, \mathcal{F})$, and let Q be an equivalent martingale measure. By Exercise 9.35, for each t there exists a positive $\mathcal{F}_t$-measurable random variable, ζ_t, called the *Radon-Nikodym derivative of P with respect to Q* such that for each $\Lambda \in \mathcal{F}_t$, $P\{\Lambda\} = \int_\Lambda \zeta_t \, dQ$. Show that $\zeta_t > 0$ a.s., that $\{\zeta_t, \ t \geq 0\}$ is a Q-martingale of mean 1, and that, if $0 < s < t$, $E^Q\{\zeta_t \mid F_s\} = \zeta_s^{-1} E^P\{X_t \zeta_t \mid \mathcal{F}_s\}$.

The following three problems extend some martingale theorems from discrete to continuous parameter submartingales. The first one extends the upcrossing inequality to a countable dense parameter set $D \subset [0, \infty)$.

9.44. Let $D_1 \subset D_2$ be finite subsets of D. Let $\{x_t, t \in D\}$ be a family of reals and let $a < b$.

    (a) Define the number of upcrossings $\nu_{D_i}(a, b)$ of $[a, b]$ by $\{x_t, t \in D_i\}$.

    (b) Show that $\nu_{D_1}(a, b) \leq \nu_{D_2}(a, b)$.

[Hint: Show that $\alpha_j^2 \leq \alpha_j^1$ and $\beta_j^2 \leq \beta_j^1$.]

    (c) Define $\nu_D(a, b) = \sup\{\nu_{D_n}(a, b) : \ D_n \text{ finite}, \ D_n \subset D\}$. Let $\{X_t, \mathcal{F}_t, \ t \in D\}$ be a submartingale. Show that

$$(9.17) \qquad\qquad E\{\nu_D(a, b)\} \leq \sup_{t \in D} \frac{E\{(X_t - a)^+\}}{b - a}.$$

9.45. Let D be a countable dense subset of $[0, \infty)$. Let $\{X_t, \mathcal{F}_t \ t \in D\}$ be a submartingale. Show that with probability one, X_t has both left and right-hand limits along D at each $t > 0$, i.e., that for a.e. ω, $X_{t+} \overset{def}{=} \lim_{s \in D, s \downarrow t, s > t} X_s(\omega)$ and $X_{t-} \overset{def}{=} \lim_{s \in D, s \uparrow t, s < t} X_s(\omega)$ exist for all $t > 0$.

[Hint: If the right-hand limit fails to exist at some t, either X is unbounded there or $\limsup_{s \downarrow t, s > t} X_s(\omega) > \liminf_{s \downarrow t, s > t} X_s(\omega)$. Apply (9.17) to rule this out.]

9.46. Let D be a countable dense subset of $[0, \infty)$. Let $\{X_t, \mathcal{F}_t \ t \in D\}$ be a submartingale. Show that there exists a submartingale $\{\hat{X}_t, \hat{\mathcal{F}}_t, \ t \geq 0\}$ such that $t \mapsto X_t$ is almost surely right-continuous and has left limits at each $t > 0$, $\hat{X}_t = X_{t+}$ for each $t \geq 0$ a.s. What are the σ-fields $\hat{F}_t$? Show, moreover, that if the original submartingale is right-continuous in probability at each $t \in D$, that $\hat{X}_t = X_t$ a.s. for each $t \in D$.

9.47. Prove Doob's inequality: for $p > 1$, $E\{\max_{j \leq n} |X_j|^p\} \leq \left(\frac{p}{p-1}\right)^p E\{|X_n|^p\}$.

[Hint: First show that if $\xi, \zeta > 0$ satisfy $P\{\xi > a\} \leq \frac{1}{a} E\{\zeta; \xi > a\}$, then $\|\xi\|_p \leq \left(\frac{p}{p-1}\right)^p \|\zeta\|_p$.]

9.48. Let (X_n) be a submartingale with uniformly bounded increments. Show that $\lim_{n \to \infty} X_n$ exists a.s. on $\{\sup_n X_n < \infty\}$.

9.49. Let (ξ_n) be a sequence of i.i.d. Bernoulli $(1/2)$ random variables. Let $X_1 = 1$, and, by induction, $X_n = 2\xi_n X_{n-1}$. Show that X_n is L^1-bounded and converges a.e., but does not converge in L^1.

9.50. Let $(\mathcal{F}_n)$ be a filtration and let $\{M_n, \mathcal{F}_n \ n = 0, 1, 2, \ldots\}$ be an adapted process. Show that M is a martingale if and only if for all bounded stopping times T, $E\{M_T\} = E\{M_0\}$.