Important concepts from today: arbitrage; hedging; European option; risk-neutral measure and risk-neutral valuation.

1. Binomial market model

Consider a market that has one type of stock and one type of riskless bond. The stock price at time $t$ is $S_t$; the bond price at time $t$ is $B_t$. We will assume that time is discrete, e.g. measured in days.

Bonds are a completely known quantity: we have

$$B_{t+1} = (1 + r)B_t \implies B_t = (1 + r)^tB_0$$

Here, $r$ is the interest rate.

Stock prices, on the other hand, may fluctuate. Let us assume that

$$S_{t+1} = \begin{cases} US_t, & \text{(stock goes up)} \\ DS_t, & \text{(stock goes down)} \end{cases} \quad \text{where } D < (1 + r) < U.$$  

**Fundamental question:** how should we price financial derivatives in this market?

**Definition 1.** For the time being, a financial derivative is a product which will pay some amount of money at time $T$; the amount depends on stock prices up to and including $T$.

We will later encounter financial derivatives that do not have a fixed execution time, at which point we will need to be more careful about the payout.

**Example 1** (European options). The two most important financial derivatives for this course are the European options.

A European call option gives the holder an option, but not the obligation to buy one unit of the stock at time $T$ for a price $K$. Therefore, its payout at time $T$ is $(S_T - K)_+ = \max(S_T - K, 0)$.

A European put option gives the holder an option, but not the obligation to sell one unit of the stock at time $T$ for a price $K$. Therefore, its payout at time $T$ is $(K - S_T)_+ = \max(K - S_T, 0)$.

The price $K$ is called the strike.

We assume that the stocks, bonds, and derivatives are freely traded on the market. This means that at any time, you can make sure to hold $x$ units of stocks, or bonds, or derivatives, for any value of $x$ (positive or negative). Negative values of $x$ correspond to short-selling the corresponding commodity.

**Definition 2.** A portfolio is a combination of stocks, bonds, and financial derivatives. The value of the portfolio at time $t$ is the price of all the items in it at time $t$. 

*Date: 9/4/2014.*
No-arbitrage pricing: if your portfolio is worth 0 at time 0, then at no future time should it be worth a guaranteed positive amount. That is, you should not be able to make a sure profit without investing some money.

2. Hedging

Suppose that we can hedge or replicate a financial derivative using the underlying assets. That is, suppose there is a combination of stocks and bonds that is worth exactly the same amount as the derivative, regardless of what happens to the stock price in the future. Since we know how much stocks and bonds are worth, this should tell us how much the derivative is worth.

Example 2 (One-period price of a contingent claim). Let \( \Phi(S_T) \) be the amount paid out by our financial derivative, if the terminal stock price is \( S_T \). (For a European call, \( \Phi(S_T) = (S_T - K)_+ \).) This is the fair price of the derivative at time \( T \).

To determine the value of the contingent claim at time \( T - 1 \), suppose that at time \( T - 1 \) we have bonds worth $x$, and also \( y \) units of the stock. What will this combination be worth at time \( T \)?

\[
\begin{align*}
(1 + r)x + y(U S_{T-1}), & \quad \text{if stock goes up} \\
(1 + r)x + y(D S_{T-1}), & \quad \text{if stock goes down}
\end{align*}
\]

We can solve the system of equations

\[
\begin{align*}
(1 + r)x + y(U S_{T-1}) &= \Phi(U S_{T-1}) = (U S_{T-1} - K)_+ \\
(1 + r)x + y(D S_{T-1}) &= \Phi(D S_{T-1}) = (D S_{T-1} - K)_+
\end{align*}
\]

The solution is

\[
x = \frac{1}{1 + r} \frac{U \Phi(D S_{T-1}) - D \Phi(U S_{T-1})}{U - D}, \quad y = \frac{\Phi(U S_{T-1}) - \Phi(D S_{T-1})}{(U - D) S_{T-1}}
\]

The price of the European call at time \( T - 1 \) is then given by

\[ V_{T-1} = x + y S_{T-1}, \]

because that’s what the replicating portfolio is worth at time \( T - 1 \).

Note: We did not need to use probability to derive this answer. For the purposes of pricing derivatives, it does not matter how likely the stock is to go up or down. What does matter is the set of possibilities for the future values of the stock.

Exercise 1 (In-class exercise). Suppose the initial stock price is \( S_0 = 100 \), \( S_1 \) is either 120 (with probability 0.6) or 80 (with probability 0.4), the strike is \( K = 110 \), and for simplicity take \( r = 0 \) (no interest). Compute the value of the European call \( C(K,1) \) at time 0.

Solution 1. We have \( U = 1.2, \ D = 0.8; \) if the stock goes up, then \( \Phi(U S_0) = (120 - 110)_+ = 10 \), and if the stock goes down, then \( \Phi(D S_0) = (80 - 110)_+ = 0 \). Plugging into the formula, we get

\[
x = \frac{1.2 \cdot 0 - 0.8 \cdot 10}{1.2 - 0.8} = -20, \quad y = \frac{10 - 0}{120 - 80} = \frac{1}{4}
\]

and the price is

\[
-20 + \frac{1}{4} \cdot 100 = 5.
\]
**Question:** What happens if someone in the market is offering to buy this call option for $6? We know that it’s worth $5, so we should clearly sell him many options, but what do we do next?

**Answer:** Sell one call option for $6. Set 1 aside. Replicate the option according to the recipe above: borrow $20 units from the bank ("short-sell $20 worth of the bond") and use them and the extra $5 from the option to buy 1/4 of a stock.

If the stock goes up, your counterparty will call on you to deliver one unit of stock for $K = 110$. You already have 1/4 units of stock, so you need to buy another 3/4 units. This will cost you another $90. Thus, $K = 110$ is exactly enough to repay your bank loan, and you have just made $1.

If the stock goes down, your counterparty won’t exercise the option. Sell your 1/4 of a stock, and repay the bank loan. You again have made $1.

Thus, you earn $1 for each call option you sell, regardless of what happens to the stock price. This lets you make an arbitrarily large profit!

**Question:** What to do over the course of many periods?

**Answer:** Repeat the above procedure! If you know what the derivative must be worth at time $T - 1$, you can work backwards to solve for what it must be worth at time $T - 2$, and so on. For example, in the binary tree below, to compute the value of the call option at time 1 when $S_1 = 120$, we solve the system of equations

\[
\begin{align*}
x + (U \cdot 120)y &= 100 \\
x + (D \cdot 120)y &= 5
\end{align*}
\]

and the value is $x + y \cdot 120$.

**Example 3** (Many-period European call). Let $S_0 = 80$, $U = 1.5$, $D = 0.5$, $r = 0$. Consider the European call $C(80, 3)$ with strike price $K = 80$ and deadline $T = 3$. Then we have a binary tree of possibilities, and as a group exercise in class we solved for the replicating portfolios at all times, and determined the value of the call at all times.

The numbers inside the ovals are possible stock prices. The numbers inside rectangles are the values of the call at the corresponding time. $x$ and $y$ give the replicating strategy: to replicate the call for one time period, you should hold $x$ worth of bond, and $y$ units of the stock.
For example, to find the price when $S_1 = 120$, we looked at the values at time 2, and solved
\[
\begin{cases}
  x + y \cdot 180 = 100 \\
  x + y \cdot 60 = 5
\end{cases}
\]
and then computed $x + y \cdot 120$.

**Exercise 2** (In-class exercise). I am offering to buy or sell the option from the binary tree for $25. Show how to make sure money from my offer. Explain in particular what your actions will be at each of $t = 0, 1, 2, 3$ if the stock price trajectory is $S_t = 80, 120, 60, 30$.

**Solution 2.** Because 25 is less than the true value (27.5), you should buy the option. You should then replicate $(-1)$ times the option. That is:

1. At time 0, you have borrow $25 from the bank to buy the option. You want your portfolio to be $22.5 - 5/8S_0$, so you set $2.50 aside, and short-sell (“borrow”) $5/8$ of a share.
2. At time 1, the replicating portfolio says you should have $42.5 - 95/120S_1$. Since you've already short-sold $5/8 = 75/120$ of a share, you need to short-sell another $20/120 = 1/6$ of a share (for $20$).
3. At time 2, your position should be $5 - 1/6S_2$. So you use $37.50 to buy $5/8 of a share.
4. At time 3, since $S_t = 30$, there is no point in exercising the option you bought. Instead, you should settle your accounts by using the $5 you have left to buy the remaining $1/6 of a share. You have made a profit of $2.50.

3. **Risk-neutral valuation: an easier way to compute value**

If you look at the binary tree, you might notice a pattern: each value is the average of the two values to the right of it. This would have given us a much easier way to compute the value of the call, had we known it in advance!

The idea of the risk-neutral valuation is as follows. Because, as we saw, we don’t care about the true probability of the stock going up and down, we may pick our favorite probability distribution. Suppose we pick the probabilities in such a way that $E[S_{t+1}|S_t] = (1 + r)S_t$, i.e. on average the stock and the bond grow by the same amount. Then for any combination of the stock and bond,

$$E[xB_{t+1} + yS_{t+1}|S_t] = (1 + r)(xB_t + yS_t).$$

In particular, a combination of stocks and bonds that was worth $0 at some point will always be worth $0 in expectation. This means that it can’t always be positive. Since an option is really just some combination of stocks and bonds, this means no portfolio that is worth $0 at time 0 will be making a sure profit $\implies$ no arbitrage!

Let’s solve for the probabilities of stock going up and down that would make $E[S_{t+1}|S_t] = (1 + r)S_t$. Suppose stock goes up with probability $p$ and down with probability $1 - p$, then we must have

$$p \cdot U + (1 - p) \cdot D = 1 + r \implies p = \frac{(1 + r) - D}{U - D}.$$

**Risk-neutral valuation of a financial derivative:** Let $p$ be as above. Then the value of a product that pays $\Phi(S_T)$ at time $T$ is given by

$$V_0 = (1 + r)^{-T}E[V_T].$$
The value at time $t$ is given by

$$V_t = (1 + r)^{-(T-t)}E[V_T|S_t].$$

For example, for the European call, $p = 1/2$, so $V_{t-1} = \frac{1}{2}V_t(\text{up}) + \frac{1}{2}V_t(\text{down})$. We can also compute

$$V_0 = p^3\Phi(270) + 3p^2(1-p)\Phi(90) + 3p(1-p)^2\Phi(30) + (1-p)^3\Phi(10)
= \frac{1}{8} \cdot 190 + \frac{3}{8} \cdot 10 + \frac{3}{8} \cdot 0 + \frac{1}{8} \cdot 0 = $27.50.

Make sure you understand why it’s $3p^2(1-p)$ and not just $p^2(1-p)$!

Exercise 3 (In-class exercise). Suppose $T = 5$, so that there are 6 possible outcomes in the binary tree. Let the pay-off for them be given by 108, 81, 54, 27, 18, 9. Suppose also that $U = 1.6$, $D = 0.7$, and $r = 0$. What’s the value of this contract at time 0?


Next,

$$V_0 = p^5 \cdot 108 + 5p^4(1-p) \cdot 81 + \binom{5}{2}p^3(1-p^2) \cdot 54 + \binom{5}{3}p^2(1-p)^3 \cdot 27
+ 5p(1-p)^4 \cdot 18 + (1-p)^5 \cdot 9.
$$

We don’t actually care very much about the number here.

This method gives us an easy way to compute the value of a financial derivative without solving lots of linear equations. We do still need to solve the equations if we want to find the replicating portfolio, but we can now solve them left-to-right as we go along, instead of having to do it for the entire tree at the very beginning.

Remark 1 (Risk-neutral measure is not the truth!). The risk-neutral measure lets us compute values of financial derivatives as if all financial products had the same expected profit. It is just a computational device! We do not actually believe that all products have the same expected profit.

One of our goals in this course is to understand how to do this analysis in continuous time.
Important concepts from today: measure; probability measure; measurable function; random variable; expectation; change of measure.

A lot of probability discusses random variables. A random variable is a map $X : \Omega \to \mathbb{R}$ from the space of outcomes to $\mathbb{R}$ (or possibly to another set). In this lecture, we will clarify what sort of a map it is, how to make sense of probabilities of events, what information is contained in a random variable, and how to use the probability measure to compute weighted averages (aka expectations).

1. Measures and random variables

**Definition 1 (Measure).** A measure on a set $S$ is a map $m$ assigning to some subsets $A \subseteq S$ nonnegative real numbers, or infinity: $m(A) \in \mathbb{R}_+ \cup \{\infty\}$. The collection $\mathcal{F}$ of subsets of $S$ on which $m$ is defined is called the $\sigma$-algebra of measurable sets (or $m$-measurable sets).

The map $m$ and the set $\mathcal{F}$ satisfy certain properties:

1. Whenever a set $A$ is measurable, its complement is also measurable. That is, if $A \in \mathcal{F}$, then also $A^c \in \mathcal{F}$, where $A^c = S \setminus A = \{s \in S : s \notin A\}$.
2. The empty set $\emptyset$ and the entire space $S$ are always measurable; the empty set has measure 0: $m(\emptyset) = 0$.
3. Countable unions and intersections of measurable events are measurable. That is, if $A_i \in \mathcal{F}$ for all $i = 1, 2, \ldots$, then also $(\bigcup_i A_i) \in \mathcal{F}$ and $(\bigcap_i A_i) \in \mathcal{F}$.
4. If $A_i \in \mathcal{F}$ are nested, that is, $A_{i+1} \subseteq A_i$ always, then $m(\bigcap_i A_i) = \lim_{i \to \infty} m(A_i)$.
5. If $A_i \in \mathcal{F}$ are disjoint, that is, $A_i \cap A_j = \emptyset$ whenever $i \neq j$, then $m(\bigcup_i A_i) = \sum_i A_i$.

This property is called $\sigma$-additivity.

A $\sigma$-algebra is any collection of sets that satisfies the properties above (excluding the properties that say that $m$ is equal to something). In particular, the $\sigma$-algebra generated by some sets includes those sets, their (countable) unions, their (countable) intersections, their complements, the (countable) unions of everything we’ve generated so far, and so on.

If $m(S) = 1$, we say that $m$ is a probability measure, and we say that $(S, \mathcal{F}, m)$ is a probability space. In this case, we often use $\Omega$ to denote the set, and $\mathbb{P}$ to denote the measure.

We will primarily be interested in locally finite measures, which means that we require every bounded measurable set to have finite measure.

Usually, measures are denoted either $m$ or $\mu, \nu$. 

Date: 9/11/2014.
Example 1.  (1) \(S = \{\text{heads, tails}\},\ m(\{\text{heads}\}) = p,\ m(\{\text{tails}\}) = 1 - p.\) Here, \(\mathcal{F} = 2^S\) is the set of all subsets of \(S.\)

(2) \(S = \{1, 2, \ldots, 6\},\) with \(\mathcal{F} = 2^S.\) Because \(S\) is finite, to define \(m,\) it is enough to define \(m(\{i\})\) for each \(i \in S.\) We can put different measures on \((S, \mathcal{F}):\)

(a) We may take \(m(\{i\}) = 1/6\) for all \(i\) (fair die).

(b) or \(m(\{6\}) = 0.5\) and \(m(\{i\}) = 0.1\) for all other \(i\) (heavily loaded die that often comes up 6).

(c) We can also take \(m(\{i\}) = 1;\) then for any set \(A \subseteq S,\ m(A)\) is the number of elements in \(A.\)

(3) \(S = \{1, 2, \ldots, 6\},\) but only the sets \(\emptyset, \{2, 3, 5\}, \{1, 4, 6\}, S\) are measurable \((\mathcal{F} \neq 2^S).\) If \(m\) is a probability measure, it will be enough to define \(m(\{2, 3, 5\}).\) Note that \(m(\{2\})\) is not defined.

(4) \(S = \mathbb{N},\) with \(\mathcal{F} = 2^S.\) Because \(S\) is countable, it’s enough to define the measure of every point, e.g. \(m(\{n\}) = 1/n.\)

(5) The Lebesgue measure on \(\mathbb{R}:\) \(S = \mathbb{R},\ m([a, b]) = b - a.\) Here, \(\mathcal{F} \neq 2^S:\) there exist non-measurable sets (although it’s hard to give an explicit example). The measure of a set is its “length”. The \(\sigma\)-algebra of measurable sets is called the Borel \(\sigma\)-algebra, and sometimes denoted \(\mathcal{B}.\) Here, \(m(\{x\}) = m([x, x]) = 0\) for any \(x \in S,\) so it’s not enough to specify the measure on one-element sets.

Because \(m(\{x\}) = 0\) for any one point, we will also have \(m(A) = 0\) for any countable set \(A.\) Any nonempty open interval \((a, b) \subset \mathbb{R}\) contains uncountably many points, so there’s not a contradiction with giving it a measure that’s greater than 0.

(6) If \((S, \mathcal{F}, \mu)\) and \((T, \mathcal{G}, \nu)\) are measure spaces, then \(S \times T\) is a measure space. The measurable sets are everything in \(\mathcal{F} \times \mathcal{G}\) and everything generated by them; the measure is given by \(m(A \times B) = \mu(A)\nu(B)\) for one of the basic sets, and then whatever it has to be for the remaining measurable sets.

(a) \(S = \{\text{sequences of four } U\text{'s or } D\text{'s}\} = \{U, D\}^4.\) If \(\mu(\{U\}) = p\) and \(\mu(\{D\}) = 1 - p,\) then \(m(\{UUUU\}) = p^4(1 - p),\) and

\[m(\{UUDD, UDDU, UDUU, DUDU, DDUU\}) = 6p^2(1 - p)^2.\]

(b) \(S = \{\text{infinite sequences of } U\text{'s and } D\text{'s}\} = \{U, D\}^{\mathbb{N}}.\) In the case of an infinite product, the measurable sets are generated by sets of the form \(A_1 \times \ldots \times A_k \times \) (entire space in all the remaining components), so in this case, by sets of the form (fixed finite sequence followed by anything). There will be non-measurable subsets of this \(S.\)

(c) The Lebesgue measure on \(\mathbb{R}^n.\) The measure of a set is its \(n\)-dimensional volume (“area” if \(n = 2\)); there are non-measurable sets in \(\mathbb{R}^n,\) just like there are in \(\mathbb{R}.\) (If you are familiar with the Banach-Tarski paradox, it demonstrates that there are non-measurable sets in \(\mathbb{R}^3.\))

Exercise 1. Show that when \(m\) is a probability measure and \(A\) is measurable, \(m(A^c) = 1 - m(A).\)

Solution: if \(A\) is measurable, so is \(A^c.\) Since \(A\) and \(A^c\) are disjoint measurable sets,

\[m(A \cup A^c) = m(A) + m(A^c).\]

The left-hand side is \(m(S) = 1,\) so \(m(A^c) = 1 - m(A).\)

Exercise 2. Let \(S = \mathbb{N},\) and let \(m\) be given by \(m(\{n\}) = 1/n\) for \(n \geq 0.\) What is the measure of the set of even integers \(A = 2\mathbb{N}?\)
Exercise 3. Let \( m \) be the Lebesgue measure on \( \mathbb{R} \). Show that for any interval \([a, b]\),
\[
m([a, b]) = m((a, b)) = m((a, b]) = m([a, b]).
\]

**Solution:** Notice that we may write all of these sets as disjoint unions of \((a, b)\) and one or both of \([a], \{b\}\). It is therefore enough to show that \( m(\{a\}) = m(\{b\}) = 0 \), but we already saw that the Lebesgue measure assigns mass 0 to any point.

Exercise 4. Let \( m \) be the Lebesgue measure. Compute \( m(\mathbb{Z}) \).

**Solution:** The set \( \mathbb{Z} \) has a countable numbers of points. We know that any one point has Lebesgue measure 0, so by \( \sigma \)-additivity of measures, \( \mathbb{Z} \) has Lebesgue measure 0. By a similar logic, the set of rational numbers \( \mathbb{Q} \) has Lebesgue measure 0.

Exercise 5. Let \( S = \mathbb{R} \), and let \( m \) be given by \( m([a, b]) = \int_a^b 1_{[0, 1]}(x)\,dx \). What is the measure of the set \( A = \{ x : x^2 < \frac{1}{2} \} \)?

**Solution:** The set \( A \) takes the form of an open interval, \( A = (-1/\sqrt{2}, 1/\sqrt{2}) \). Just as for the Lebesgue measure, it doesn’t matter whether or not we include the endpoints of \( A \), because \( m(\{a\}) = 0 \) for any single point \( a \). We compute
\[
m(A) = \int_{-1/\sqrt{2}}^{1/\sqrt{2}} 1_{[0, 1]}(x)\,dx = \int_0^{1/\sqrt{2}} 1\,dx = \frac{1}{\sqrt{2}}.
\]

Definition 2 (Measurable function). Let \((S, \mathcal{F}, \mu)\) and \((T, \mathcal{G}, \nu)\) be two measure spaces. (We don’t really care about \( \nu \), only about the \( \sigma \)-algebra of \( \nu \)-measurable sets.) A function \( f : S \to T \) is called **measurable** if the preimages of \( \nu \)-measurable sets are \( \mu \)-measurable. That is, if \( B \in \mathcal{G} \), then \( f^{-1}(B) \in \mathcal{F} \).

If \( T = \mathbb{R} \), we almost always consider the Borel \( \sigma \)-algebra, so we require \( f^{-1}(B) \) to be measurable for every Borel set \( B \). (It’s enough to require that \( f^{-1}(B) \) is measurable for every open set \( B \).) In particular, if \( f : S \to \mathbb{R} \) is measurable, then for every \( x \in \mathbb{R} \), we must have \( f^{-1}(x) \in \mathcal{F} \).

Example 2. (1) When all subsets of \( S \) are measurable \( (\mathcal{F} = 2^S) \), then all functions \( f : S \to T \) are measurable. (The preimage of anything is a subset of \( S \), and therefore measurable.) (2) Suppose \( S = \{1, 2, \ldots, 6\} \), and \( \mathcal{F} = \{\emptyset, \{2, 3, 5\}, \{1, 4, 6\}, S\} \). Then any measurable function \( f : S \to T \) must be constant on \( \{2, 3, 5\} \) and on \( \{1, 4, 6\} \). For example, let \( x = f(2) \), and let \( B \) be any measurable set containing \( x \). Then \( 2 \in f^{-1}(B) \), and therefore \( \{2, 3, 5\} \subseteq f^{-1}(B) \). (3) Note that I can have different measures and \( \sigma \)-algebras on the same underlying set \( S \), so the same function \( f : S \to T \) may or may not be measurable depending on \( \mathcal{F} \). When we talk about conditional expectation, we will think about approximating a given function by \( \mu \)-measurable functions. (4) If \( S = \mathbb{R} \) and \( m \) is the Lebesgue measure, then any continuous function is measurable; but many measurable functions are not continuous. For example, \( f(x) = \lfloor x \rfloor \) is measurable: the preimage of any set is a union of intervals. It is quite hard to write down a function \( f : \mathbb{R} \to \mathbb{R} \) that is not measurable.
Definition 3 (Random variable). Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space. A real-valued random variable is a measurable map \(X : \Omega \rightarrow \mathbb{R}\). For any set \(S\) with \(\sigma\)-algebra \(\mathcal{G}\), an \(S\)-valued random variable is a measurable map \(X : \Omega \rightarrow S\).

For \(A \subset \mathbb{R}\) (or \(A \subset S\)), when we write \(\mathbb{P}(X \in A)\), we mean \(\mathbb{P}\{\omega \in \Omega : X(\omega) \in A\}\).

Notice that \(\mathbb{P}\) is a measure on \(\Omega\), so we have to feed a subset of \(\Omega\) to it!

In real life, we can think of \(\omega \in \Omega\) as the “state of the universe”, and \(X(\omega)\) is some observable quantity. For example, in rolling a die, \(\omega\) is the set of conditions (how hard I threw the die, how I held it, what is the die made of, what are the air currents in the room, ...) and if I knew \(\omega\) exactly, I should be able to determine how the die is going to land. The probability measure on \(\Omega\) quantifies our beliefs about which states of the world are more likely than others.

Note that we can also think of \(\Omega\) for this example as the set of numbers \(\{1, \ldots, 6\}\), which are about equally likely.

Definition 4 (\(\sigma(X)\) and induced probability). Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space, and let \((T, \mathcal{G}, \nu)\) be an arbitrary measure space. Whenever we have a random variable \(X : \Omega \rightarrow T\), we automatically get two induced objects: a \(\sigma\)-algebra on \(\Omega\) (which is often smaller than \(\mathcal{F}\)), and a measure on \(T\) (which is often different from \(\nu\)).

The \(\sigma\)-algebra generated by \(X\), denoted \(\sigma(X)\), is the preimage of \(\mathcal{G}\): \(\sigma(X) = X^{-1}(\mathcal{G})\). In particular, if \(X(\omega) = X(\omega')\) then every set in \(\sigma(X)\) contains either both \(\omega\) and \(\omega'\), or neither of them; this may not have been the case for \(\mathcal{F}\). In particular, if \(Y: \Omega \rightarrow S\) is measurable with respect to \(\sigma(X)\), then we must be able to express \(Y(\omega)\) as a function of \(X(\omega)\), i.e. \(Y\) is a function of \(X\).

The induced probability measure, also called the law of \(X\), is given by \(m(A) = \mu(\{\omega : X(\omega) \in A\})\). A lot of the time we care only about the law of \(X\), and not about the underlying probability space that \(X\) is defined on. Note that if \(T = \mathbb{R}\), the law of \(X\) is not the Lebesgue measure, since \(m(\mathbb{R}) = 1\); but it may be possible to express it in terms of the Lebesgue measure.

Example 3. (1) Let \(S = \{U, D\}^4\), and let \(X : S \rightarrow \mathbb{R}\) count the number of \(U\)'s in the sequence. (The range of \(X\) is thus the discrete set \(\{0, \ldots, 4\}\).) Suppose the measure on \(S\) is the product measure, with \(\mu(\{U\}) = p\) and \(\mu(\{D\}) = 1 - p\).

The induced measure – now on the set of numbers \(\{0, \ldots, 4\}\) and not on the set of sequences of \(U\)’s and \(D\)’s – is

\[
\begin{align*}
m(\{0\}) &= (1 - p)^4, & m(\{1\}) &= 4p(1 - p)^3, & m(\{2\}) &= 6p^2(1 - p)^2, \\
m(\{3\}) &= 4p^3(1 - p), & m(\{4\}) &= p^4.
\end{align*}
\]

If \(p = 0.5\), the CDF of \(m\) looks like this:

\[
\begin{array}{c|cccc}
\text{CDF} & 0 & 1 & 2 & 3 & 4 \\
\text{Value} & 0 & 1/2 & 3/4 & 1 & 1
\end{array}
\]

If \(Y : S \rightarrow \mathbb{R}\) has \(Y(UUDU) \neq Y(DUUU)\) then \(Y\) is not \(\sigma(X)\)-measurable. Any \(\sigma(X)\)-measurable function must be a function only of the number of \(U\)'s in a sequence.

(2) If \(S = \{U, D\}^\infty\) and \(X : S \rightarrow \{U, D\}^2\) takes the first two elements of the sequence, then any \(\sigma(X)\)-measurable function must depend only on the first two elements.
of the sequence. So if $Y(UUDUUDD\ldots) \neq Y(UUDUDDD\ldots)$ then $Y$ is not $\sigma(X)$-measurable.

We will use capital letters $(X, Y)$ for random variables, and lowercase letters for their values $(x, y, a, s)$.

**Definition 5 (Cumulative distribution function).** For a probability measure $m$ on $\mathbb{R}$, its cumulative distribution function (CDF) is the increasing function

$$F_m(x) = m((-\infty, x]).$$

In particular, for a random variable $X : \Omega \to \mathbb{R}$, the cumulative distribution function of (the law of) $X$ is

$$F_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in (-\infty, x]\})$$

$$= \mathbb{P}(X \leq x).$$

We often call this the CDF of $X$ itself, omitting the “law of”.

Notice that $F_m$ is always increasing (or at least nondecreasing), since it is looking at the measure of larger and larger sets.
Important concepts from today: expectation is an average, absolutely continuous measures, Radon-Nikodym derivative.

1. Review of last time

Last time, we defined measures, $\sigma$-algebras, random variables, and what it means for one random variable $X$ to be measurable with respect to $\sigma(Y)$. Finally, we defined the cumulative distribution function of a measure and of a random variable.

Here are the most important concepts from last time:

$X$ is $\sigma(Y)$-measurable if knowing the value of $Y$ lets you compute the value of $X$. (Formally, there is a requirement that $X$ must be a nice function of $Y$, but you’re unlikely to come across non-nice functions.)

The CDF of a measure is $F_m(x) = m((-\infty, x])$. To determine the CDF, you need to compute $m(A)$, where $A = (-\infty, x]$. The CDF of a random variable is $F_X(s) = \mathbb{P}(X \in (-\infty, s])$. If the CDF is continuous at $x$, then $\{x\}$ has measure 0, or $\mathbb{P}(X = x) = 0$; if the CDF has a jump at $x$, then $\{x\}$ has a positive measure equal to the size of the jump, or $\mathbb{P}(X = x) > 0$.

When the $\sigma$-algebra on $\mathbb{R}$ is the Borel sets, the CDF contains exactly the same information as the entire measure.

**Example 1.**

1. Write $(-\infty, b] = (-\infty, a] \cup (a, b]$. This means that $m((a, b]) = F_m(b) - F_m(a)$.

2. Write $\{a\} = \cap_n (a - \frac{1}{n}, a]$. This means that $m(\{a\}) = \lim_{n \to \infty} m((a - \frac{1}{n}, a]) = \lim_{n \to \infty} F_m(a) - F_m(a - \frac{1}{n})$.

   If $F_m$ is continuous at $a$, then $m(\{a\}) = 0$. Otherwise, if $F_m$ has a jump at $a$, then $m(\{a\})$ is the size of this jump. In this case, $a$ is called an atom for the measure.

3. We can now compute the measure of any open or closed interval, and therefore the measure of any Borel set. In particular:

   $$m([a, b]) = F_m(b) - \lim_{n \to \infty} F_m(a - \frac{1}{n});$$

   $$m((a, b]) = \lim_{b \to \infty} F_m(b - \frac{1}{n}) - F_m(a).$$

*Date: 9/11/2014.*
2. Expectations

The point of a (probability) measure is to be able to take weighted averages of the values of $X$. For example, if

$$S = \{1, 2, 3, 4, 5, 6\}, \quad m(\{6\}) = \frac{1}{2}, \quad m(\{i\}) = 0.1 \text{ otherwise}, \quad X(s) = s$$

then the expected value of $X$ is the weighted average of the outcomes:

$$E[X] = 0.1 \times 1 + 0.1 \times 2 + 0.1 \times 3 + 0.1 \times 4 + 0.1 \times 5 + 0.5 \times 6 = 4.5$$

In general, when $X$ takes values in a discrete set $S = \{s_1, s_2, \ldots \}$,

$$E[X] = \sum_i s_i P(X = s_i).$$

**Question:** how do we compute $E[X^2]$, or $E[\sin(X)]$, or more generally $E[g(X)]$ for some function $g$?

**Answer:** if $X = s_i$ then $g(X) = g(s_i)$, so:

$$E[g(X)] = \sum_i g(s_i) P(X = s_i).$$

This is (part of) the “Law of the Unconscious Statistician”.

We can also write

$$E[g(X)] = \sum_{u \in g(S)} u P(g(X) = u) = \sum_{u \in g(S)} u \left( \sum_{i : g(s_i) = u} P(X = s_i) \right)$$

where we take the sum over the possible values of $g(X)$.

**Remark** 1. The two sums are really equal if $S$ has a finite number of elements, if $g(X) > 0$ always, or if $E[|g(X)|] < \infty$. Otherwise, it is possible for only one of these sums to converge.

**Exercise** 1. Let $X$ take values in $\mathbb{Z}_+ = \{0, 1, 2, \ldots \}$ (we don’t care about the underlying probability space here), with $P(X = n) = e^{-\lambda} \lambda^n / n!$. This $\mathbb{Z}$-valued random variable is a Poisson random variable with rate $\lambda$; its distribution is the Poisson distribution with rate $\lambda$.

Compute $E[e^X]$. 

2
Solution:

\[
\mathbb{E}[e^X] = \sum_{n} e^{-\lambda} \frac{\lambda^n}{n!} e^n \mathbb{P}(X=n) g(n)
\]

\[
= \sum_{n} e^{-\lambda} \frac{\lambda^n}{n!} e^{\lambda}
\]

\[
= \sum_{n} e^{-\lambda} \frac{\lambda^n}{n!} e^{\lambda(n-1)} \frac{\lambda^n}{n!}
\]

\[
= e^{\lambda(n-1)} \sum_{n} e^{-\lambda} \frac{\lambda^n}{n!} = 1, \text{ Poisson distribution with rate } \lambda e
\]

In words, what we did was combine the two exponential terms and then rearrange the constants to get the probability mass function of a new Poisson distribution with a different parameter. We then used the fact that the probability mass function adds up to 1.

**Exercise 2.** Let \( X \) take values in \( S = \{U, D\}^{10} \) (so \( X \) is an \( S \)-valued random variable); let the probability measure induced by \( X \) be the uniform probability measure, i.e. \( \mathbb{P}(X = s) \) is the same for all \( s \in S \). Let the function \( g \) be given by

\[
g(s) = \begin{cases} 
1.5, & \text{first letter of } s \text{ is } U \\
0.5, & \text{first letter of } s \text{ is } D.
\end{cases}
\]

Compute \( \mathbb{E}[g(X)] \).

**Solution:** In this case, writing out \( \mathbb{E}[g(X)] = \sum_{s} \mathbb{P}(X = s) g(s) \) would be hideously boring: we’d have to enumerate all the \( 2^{10} = 1024 \) sequences \( s \in S \). Instead, we write

\[
\mathbb{E}[g(X)] = \sum_{u \in g(S)} u \mathbb{P}(g(X) = u)
\]

\[
= 1.5 \cdot \mathbb{P}(g(X) = 1.5) + 0.5 \cdot \mathbb{P}(g(X) = 0.5)
\]

\[
= 1.5 \cdot \mathbb{P}(X \in \{\text{sequences beginning with } U\}) + 0.5 \cdot \mathbb{P}(X \in \{\text{sequences beginning with } D\})
\]

\[
= 1.5 \cdot \frac{1}{2} + 0.5 \cdot \frac{1}{2} = 1.
\]

**Question:** what happens when \( X \) does not take values in a discrete set? For example, what about \( X \in \mathbb{R} \)?

**Answer:** let’s try to approximate \( X \) by a discrete random variable. Let \( g_n \) be the function approximating \( x \) to \( n \) decimal places, then we know how to compute \( \mathbb{E}[g_n(X)] \), and we expect it to be close to \( \mathbb{E}[X] \):

\[
\mathbb{E}[X] = \lim_{n \to \infty} \mathbb{E}[g_n(X)] = \lim_{n \to \infty} \sum_{x \text{ has } n \text{ decimal places}} x \mathbb{P}(g_n(X) = x).
\]

By definition, this is the Riemann-Stiltjes integral with respect to \( F_X \), or the Lebesgue integral with respect to the measure defined by the CDF \( F_X \):

\[
\mathbb{E}[X] = \int_{-\infty}^{\infty} x dF_X(x).
\]
But that’s not a very useful statement, because we still don’t know how to evaluate the integral.

Let’s think some more about $P(g_n(X) = x)$:

$$P(g_n(X) = x) = P(X \in [x - \frac{1}{2} \times 10^{-n}, x + \frac{1}{2} \times 10^{-n}]).$$

(We assume that 0.5 rounds up.) Here’s what it looks like when we plot the CDF of $X$:

![Plot of CDF](image)

The values of $x$ are shown in red; the height of the gray boxes gives $P(X_n = x)$.

Suppose the CDF of $X$ is differentiable at $x$, then

$$P(X \in [x - \frac{1}{2} \epsilon, x + \frac{1}{2} \epsilon]) \approx F_X'(x) \epsilon.$$

When $F_X$ is differentiable (almost) everywhere, we define

$$f_X(x) = F_X'(x), \quad \text{when } F_X \text{ is differentiable almost everywhere.}$$

This is the **probability density function** of $X$. When $F_X$ is differentiable,

$$E[X] = \lim_{n \to \infty} \sum_{x \text{ has } n \text{ decimal places}} x f_X(x) 10^{-n} = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Note that while the CDF is always increasing, the pdf must be positive, but doesn’t have to be increasing. Also, while the CDF is always defined, the pdf may not be defined ($F_X$ may not be differentiable.)

Update to the law of the unconscious statistician: if $X \approx x$, and $h$ is a (nice) function, then $h(X) \approx h(x)$. So:

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx, \quad F_X \text{ differentiable almost everywhere, } h \text{ nice.}$$

For all practical purposes, a function $h$ is nice if it’s always positive, or if $\int |h(x)| f_X(x) dx < \infty$.

**Exercise 3.** Let $X$ have the density $\frac{1}{3\sqrt{2\pi}} \exp(-\frac{1}{2\cdot 3^2} (x - 5)^2)$ (a normal random variable with mean 5 and variance 9). Compute $E[e^X]$. 


Solution:

$$
\mathbb{E}[e^X] = \int_{-\infty}^{\infty} e^x \frac{1}{3\sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{1}{9} (x-5)^2 + 2x \right)} dx 
$$

$$
= \int_{-\infty}^{\infty} \frac{1}{3\sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{1}{9} (x^2 - 10x + 25 + 18x) \right) \right) dx 
$$

$$
= \int_{-\infty}^{\infty} \frac{1}{3\sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{1}{9} (x^2 + 8x + 25) \right) \right) dx 
$$

$$
= \int_{-\infty}^{\infty} \frac{1}{3\sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{1}{9} ((x+4)^2 + 9) \right) \right) dx 
$$

$$
= \int_{-\infty}^{\infty} \frac{1}{3\sqrt{2\pi}} \exp \left( -\frac{11}{29} (x+4)^2 \right) e^{-1/2} dx 
$$

$$
= e^{-1/2} \int_{-\infty}^{\infty} \frac{1}{3\sqrt{2\pi}} \exp \left( -\frac{11}{29} (x+4)^2 \right) dx 
$$

$$
= e^{-1/2} \cdot \text{density of N}(-4,9) \text{ integrates to 1} 
$$

In words, what we did was combine the two exponential terms, complete the square, pull out the extra constant, and use the fact that what we got at the end was the integral of the density of a new normal random variable, so it must be equal to 1.

**Some properties of expectation:**

1. Expectation is linear: if $a, b \in \mathbb{R}$ and $X, Y$ are random variables, then $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. (Assuming $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ exist.)

2. Expectation of a constant is equal to that constant: if $X = c$ with probability 1, then $\mathbb{E}[X] = c$.

3. In general, $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$; for example, $\mathbb{E}[X^2] \neq (\mathbb{E}[X])^2$. If $g$ is a linear function, $g(x) = ax + b$, then the two will be equal.

**3. Density and change of measure**

We saw the probability density function above. This is the density of a real-valued random variable with respect to the Lebesgue measure on $\mathbb{R}$. We want to generalize this notion; in particular, we will want to write down densities of one random variable with respect to (the law of) another random variable.

**Definition 1 (Absolute continuity).** Let $\mathcal{F}$ be a $\sigma$-algebra of measurable sets on a set $S$, and let $\mu$ and $\nu$ be two measures on $\mathcal{F}$. We say that $\mu$ is absolutely continuous with respect to $\nu$, and write $\mu \ll \nu$, if $\nu(A) = 0 \implies \mu(A) = 0$.

If $\mu \ll \nu$ and also $\nu \ll \mu$ (i.e. the two measures have the same set of events of measure 0), we say $\mu$ is equivalent to $\nu$ and write $\mu \sim \nu$.

Suppose $\mu \ll \nu$ are measures on $\mathbb{R}$, and let’s think about their CDF’s:

1. If $F_\nu$ is flat on an interval, than that interval has $\nu$-measure 0, so $F_\mu$ must also be flat on that interval. If $\mu \sim \nu$, then $F_\nu$ and $F_\mu$ are flat on the same sets of intervals.
(2) If $F_\nu$ is continuous on an interval, then $F_\mu$ can’t have jumps on that interval: a jump point would have positive $\mu$-measure, but $\nu$-measure 0. If $\mu \sim \nu$, then the sets of jump points of $F_\nu$ and $F_\mu$ coincide.

(3) These two conditions are almost, but not quite enough to describe the CDFs of absolutely continuous measures; it’s possible for a continuous CDF to assign positive measure to sets of measure 0 (look up Cantor’s function, also called “Devil’s staircase”, if you’re interested).

**Theorem 3.1** (Radon-Nikodym derivative). If $\mu \ll \nu$, then it is possible to write the CDF $F_\mu$ as

$$F_\mu(x) = \int_{-\infty}^{\infty} h(x) dF_\nu(x),$$

for some measurable function $h : \mathbb{R} \to \mathbb{R}$. The function $h(x)$ is called the Radon-Nikodym derivative, or density, of $\mu$ with respect to $\nu$, denoted $h = d\mu / d\nu$. (We may also write $h = dF_\mu / dF_\nu$.)

When $F_\mu$ and $F_\nu$ both have densities, $h(x) = f_\mu(x) / f_\nu(x)$. The ratio $h$ will not be defined on any set where $f_\nu = 0$, but those sets by definition have $\nu$-measure 0, and hence $\mu$-measure 0 as well.

**Exercise 4.** Let $X \sim N(0, 1)$ be a normal random variable with mean 0 and variance 1. Let $Z = \exp(-3X + \frac{9}{2})$. Check that $\mathbb{E}[Z] = 1$.

Let $Y$ be a random variable such that $Z = dF_Y / dF_X$. What is the law of $Y$?

**Solution:** First, we compute

$$\mathbb{E}[Z] = \mathbb{E}[e^{-3X+9/2}] = \int_{-\infty}^{\infty} e^{-3x+9/2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x^2 - 6x + 9)\right) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - 3)^2\right) dx$$

$$= 1.$$

Because $X$ has a density, we will compute the density of $Y$:

$$\frac{f_Y(a)}{f_X(a)} = Z_{X=a} = \exp(-3a + \frac{9}{2}).$$

We’re looking at $f_X(a)$, so $X = a$.

Consequently,

$$f_Y(a) = f_X(a) \cdot \exp(-3a + \frac{9}{2}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a^2} e^{-3a+9/2} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(a-3)^2},$$

and $Y$ is a normal random variable with mean 3 and variance 1.

We’ll come back to this exercise towards the end of the semester, when we talk about Girsanov’s theorem.
Important concepts from today: independent random variables, what does it mean for many random variables to be independent, uncorrelated doesn’t mean independent, normal random variables.

1. Review of last time

Last time, we talked about what it means for one measure to be absolutely continuous with respect to the other:

\[ \mu \ll \nu \text{ means that if } \nu(A) = 0, \text{ then also necessarily } \mu(A) = 0. \]

We saw that if the only set \( A \) such that \( \nu(A) = 0 \) is \( A = \emptyset \), then \( \mu \ll \nu \) for any other measure \( \mu \) defined on the same space.

When \( \mu \) and \( \nu \) are measures on \( \mathbb{R} \), we also saw how to tell whether \( \mu \ll \nu \) by looking at their CDFs. (This was not a completely precise test, but we won’t get bogged down in the details.)

For two measures \( \mu \) and \( \nu \) on \( \mathbb{R} \), we will say that \( \mu \ll \nu \) if two conditions hold:

1. whenever the CDF of \( \nu \) is flat, the CDF of \( \mu \) is also flat;
2. whenever the CDF of \( \mu \) has a jump, the CDF of \( \nu \) also has a jump (but it may jump by a different amount).

We stated that if \( \mu \ll \nu \), then we can write \( \mu(A) = \int_A h(x) d\nu(A) \), where the function \( h \) was measurable, and also called \( h(x) = d\mu/d\nu(x) \), the Radon-Nikodym derivative. An important example of this was when \( \nu \) is the Lebesgue measure. We said that a measure (or CDF) is absolutely continuous with respect to the Lebesgue measure if the CDF is continuous, and differentiable almost everywhere. “Almost everywhere” meant “everywhere but a set of measure 0”, which in practice usually means “everywhere except for a countable set of points”.

If \( \nu \) is the Lebesgue measure, then the Radon-Nikodym derivative \( d\mu/d\nu(x) \) is \( F'_\mu(x) \), the ordinary calculus derivative of the CDF of \( \mu \) at \( x \). It may be undefined on a set of (Lebesgue) measure 0, which typically means “a set of countably many points”.

Finally, we explained how to find the Radon-Nikodym derivative of two random variables, both of which have a density:

If the CDFs of both \( \mu \) and \( \nu \) have a density (with respect to the Lebesgue measure), then \( d\mu/d\nu(x) = (F'_\mu(x))/(F'_\nu(x)) \), the ratio of the derivatives.

We didn’t have time to compute an example of the Radon-Nikodym derivative in the case of measures on discrete sets. The concept is actually much simpler there.

\[ \text{Date: 9/25/2014.} \]
If $\mu$ and $\nu$ are measures on a discrete set (but only then!), then we can take
\[
\frac{d\mu}{d\nu}(x) = \frac{\mu(A)}{\nu(A)}, \quad A \text{ is the smallest measurable set containing } x.
\]
In particular, if one-element sets are measurable,
\[
\frac{d\mu}{d\nu}(x) = \frac{\mu(\{x\})}{\nu(\{x\})}, \quad \text{on a discrete set, } \{x\} \text{ measurable.}
\]

We can’t use this definition on $\mathbb{R}$, because we’d be getting $0/0$ most of the time.

**Example 1.** Let $S = \{1, 2, 3, 4, 5, 6\}$. Let $\mathcal{F}$ be the $\sigma$-algebra containing the sets $\{1\}$, $\{2\}$, $\{3, 5\}$, and $\{4, 6\}$, as well as their arbitrary unions. (Thus, $\mathcal{F}$ has $2^4 = 16$ elements.)

Let $\mu$ be the measure given by $\mu(\{1\}) = 0$, $\mu(\{2\}) = \mu(\{3, 5\}) = \mu(\{4, 6\}) = 1/3$. Let $\nu$ be the measure given by $\nu(\{1\}) = 0$, $\nu(\{2\}) = 1/3$, $\nu(\{3, 5\}) = 0$, $\nu(\{4, 6\}) = 2/3$.

The only sets of $\mu$-measure $0$ are $\emptyset$ and $\{1\}$. Since these also have $\nu$-measure $0$, we see $\nu \ll \mu$. We also see $\mu \not\ll \nu$, since $\nu(\{3, 5\}) = 0$ but $\mu(\{3, 5\}) > 0$.

The Radon-Nikodym derivative $d\nu/d\mu$ must be a measurable function, that is, it will take one value on $3$ and $5$, and one value on $4$ and $6$. To compute it, we write
\[
\frac{d\nu}{d\mu}(1) = \frac{\nu(\{1\})}{\mu(\{1\})} = 0 = \text{not defined}
\]
\[
\frac{d\nu}{d\mu}(2) = \frac{\nu(\{2\})}{\mu(\{2\})} = \frac{1/3}{1/3} = 1
\]
\[
\frac{d\nu}{d\mu}(3) = \frac{\nu(\{3, 5\})}{\mu(\{3, 5\})} = \frac{0}{1/3} = 0
\]
\[
\frac{d\nu}{d\mu}(4) = \frac{\nu(\{4, 6\})}{\mu(\{4, 6\})} = \frac{2/3}{1/3} = 2
\]
\[
\frac{d\nu}{d\mu}(5) = \frac{\nu(\{3, 5\})}{\mu(\{3, 5\})} = 0
\]
\[
\frac{d\nu}{d\mu}(6) = \frac{\nu(\{4, 6\})}{\mu(\{4, 6\})} = 2.
\]

The derivative is not defined at $1$: this means that we could put any value there, and the theorem would still hold. This doesn’t matter, because the set of points where the derivative is not defined has measure $0$.

Take a look also at the posted solutions to the homework, which work out a few examples in this vein.

## 2. Independent Random Variables

Two random variables $X, Y$ defined on the same probability space are called independent if for any (measurable) events $A, B \subset \mathcal{T}$,
\[
\mathbb{P}(X \in A \text{ and } Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).
\]

Let’s do some examples, and see why this should mean that $X$ and $Y$ are independent.

**Example 2.** (1) The canonical example is tossing coins, or rolling dice. You toss two coins, and whether the second coin comes up heads doesn’t depend on what the first coin did. Therefore, the probability that the first coin comes up heads, and the second comes up tails, is $\mathbb{P}(H, T) = \mathbb{P}(H)\mathbb{P}(T)$. Or, you roll a die twice, and the probability that you roll a 3 and then a 5 is $\mathbb{P}(3, 5) = BP(3)\mathbb{P}(5)$. 

Let’s say this formally. When we roll a die twice (say), we are looking at a random variable defined on \( T = \{1, 2, 3, 4, 5, 6\}^2 \), the set of pairs of numbers 1 through 6. The two random variables we’re looking at are

\[ X((s_1, s_2)) = s_1, \quad Y((s_1, s_2)) = s_2. \]

The probability that \( X \in A \) and \( Y \in B \) (for example, \( X = 3 \) and \( Y = 5 \)) is

\[
\mathbb{P}(X \in A \text{ and } Y \in B) = \mathbb{P}(s_1 \in A \text{ and } s_2 \in B) = \mathbb{P}((s_1, s_2) \in A \times B). 
\]

If we’re looking at the uniform probability measure, i.e. all pairs of outcomes are equally likely, then this probability is simply

\[
\mathbb{P}((s_1, s_2) \in A \times B) = \frac{\#(\text{pairs } (s_1, s_2) \text{ in } A \times B)}{\#(\text{pairs } (s_1, s_2) \in S = \{1, \ldots, 6\}^2)} = \frac{\mathbb{P}(s_1 \in A) \cdot \mathbb{P}(s_2 \in B)}{\#(s \in S)^2}.
\]

But the right-hand side is also

\[
\frac{\mathbb{P}(s_1 \in A) \cdot \mathbb{P}(s_2 \in B)}{\#(s \in S)^2} = \mathbb{P}(s_1 \in A)\mathbb{P}(s_2 \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).
\]

In our example \( A = \{3\}, \ B = \{5\} \), this is of course \( 1/36 = 1/6 \cdot 1/6 \), but fundamentally this is just saying that the number of outcomes in the product set \( A \times B \) is the product of the number of outcomes in \( A \) and the number of outcomes in \( B \), and we’re assuming that all pairs of outcomes are equally likely.

(2) In general, when I have two random variables \( X, Y : S \to T \), I can define a pair of independent random variables

\[ \tilde{X} : S \times S \to T, \quad \tilde{X}((s_1, s_2)) = X(s_1), \quad \tilde{Y} : S \times S \to T, \quad \tilde{Y}((s_1, s_2)) = Y(s_2). \]

That is, I make \( \tilde{X} \) and \( \tilde{Y} \) be functions of independent components of \((s_1, s_2) \in S \times S\). Notice that \( \tilde{X} \) and \( \tilde{Y} \) are defined on the set \( S \times S \) – not the original set \( S \)! In particular, while \( X \) and \( \tilde{X} \) induce the same law on \( T \), i.e. \( \mathbb{P}(X \in A) = \mathbb{P}(\tilde{X} \in A) \) for any \( A \subset T \), they are not the same random variable!

(3) Usually, when \( X \) and \( Y \) are independent random variables defined on a space \( S \), it’s possible to think of \( S \) as a product set, but sometimes it’s easier not to do that. For example, let \( X, Y : [0, 1] \to \mathbb{R} \) be given by

\[
X(s) = \begin{cases} 
0, & s < \frac{1}{2} \\
1, & 1/2 \leq s \leq 1
\end{cases}, \quad Y(s) = \begin{cases} 
0, & s < \frac{1}{2} \\
1, & 1/4 \leq s < 1/2 \\
0, & 1/2 \leq s < 3/4 \\
1, & 3/4 \leq s \leq 1.
\end{cases}
\]

Let’s check that they are independent. Since \( X \) and \( Y \) take on two values each (0 and 1), and each with probability 1/2, it’s enough to check that \( \mathbb{P}(X = 0, Y = 0) = 1/4 \) and \( \mathbb{P}(X = 1, Y = 1) = 1/4 \). (Notice that \( \mathbb{P}(X = 0, Y = 1) = \mathbb{P}(X = 0) - \mathbb{P}(X = 0, Y = 0) \), so if the product rule works for the event \( (X = 0, Y = 0) \), it’ll work for the event \( (X = 0, Y = 1) \).)

Checking:

\[
\mathbb{P}(X = 0, Y = 0) = \mathbb{P}(s < 1/4) = 1/4; \quad \mathbb{P}(X = 1, Y = 1) = \mathbb{P}(3/4 < s) = 1/4.
\]

Here we can also think of \( X \) as reading off the first digit of \( s \) in binary notation, and \( Y \) is reading off the second digit of \( s \) in binary notation. In this sense,
they are defined on a product space, but we didn’t need to use that to show the independence.

(4) If $X$ and $Y$ are independent, and $g_1$ and $g_2$ are measurable functions, then $g_1(X)$ and $g_2(Y)$ are independent. Indeed:

$$P(g_1(X) \in A, g_2(Y) \in B) = P(X \in g^{-1}(A), Y \in g^{-1}(B)),$$

and it’s pretty clear from here.

(5) Reminder: independent random variables must be defined on the same probability space. It doesn’t matter if they take the same values.

(6) Reminder: you can also have variables defined on a product space that aren’t independent. For example, let $S = \{U, D\}^2$, let $X : S \to \mathbb{R}$ count the number of $D$’s, and let $Y : S \to \mathbb{R}$ be the indicator of whether the second letter is a $D$. These are not independent: if $X = 2$ then necessarily $Y = 1$, so

$$P(X = 2, Y = 1) \neq P(X = 2)P(Y = 1),$$
meaning $X$ and $Y$ are not independent.

We now have three concepts under our belt that we can apply to a pair of random variables: we can test whether they’re independent; we can test whether they are measurable with respect to each other; and we can test whether their laws are absolutely continuous with respect to each other. These are all different concepts. For absolute continuity, it’s enough to look at each of $X$ and $Y$ separately. On the other hand, measurability and independence both require us to look at the values $X(\omega)$ and $Y(\omega)$ simultaneously, on the same $\omega$. That is, measurability and independence are about the relationship between the measurable (identifiable) sets $\sigma(X)$ and $\sigma(Y)$, whereas absolute continuity is about the values that $X$ and $Y$ take. Notice, however, that independence also involves looking at the law of $X$ and $Y$: I can’t tell whether two random variables are independent without knowing their probability distribution.

Independence is a little like the opposite of measurability. If $X$ is $\sigma(Y)$-measurable, then knowing the value of $Y$ should let me compute the value of $X$. If $X$ and $Y$ are independent, then knowing the value of $Y$ should tell me absolutely nothing about the value of $X$.

**Example 3** (The constant random variable). Here’s an important example of a random variable that’s independent of any other random variable, and also measurable with respect to any other random variable: $X(s) = c$, a fixed constant, for all $s \in S$.

Let’s check that if we have another random variable $Y : S \to \mathbb{R}$, then $X$ is independent of $Y$. Indeed, we have

$$P(X \in A, Y \in B) = \begin{cases} P(Y \in B), & c \in A \\ 0, & c \notin A \end{cases}$$
as it should be. Let’s also check that $X$ is $\sigma(Y)$-measurable, by showing that to each value of $Y$ is associated only one value of $X$: of course, that value is $c$.

This is essentially the only overlap between measurable and independent random variables: if $X$ is both $\sigma(Y)$-measurable and independent of $Y$, then we must have $X(s) = c$ for some $c$ with probability 1, i.e. $X$ is a constant.

Independence can be extended to more than two random variables. When we say that $X_1, X_2, \ldots, X_n$ are independent, we mean

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_1 \in A_1)P(X_2 \in A_2)\ldots P(X_n \in A_n).$$
This implies that every pair among these random variables is independent, but it also says more than that.
Here’s an important concept that we’ll come back to later: when we say that an infinite collection of random variables is independent, we mean that every finite subcollection is independent. That is, we don’t look at infinite products above.

How do we use independence? Here are two of the most important examples.

**Theorem 2.1.** If \( X_1, X_2, \ldots \) all have densities, then their joint density is the product of the individual densities:

\[
f_{(x_1, x_2, \ldots, x_n)}(a_1, \ldots, a_n) = f_{X_1}(a_1) f_{X_2}(a_2) \ldots f_{X_n}(a_n).
\]

We haven’t actually worked with joint densities before, but they do the same thing as ordinary densities:

\[
\Pr(X_1 \in (a_1 \pm \frac{1}{2}\epsilon_1), X_2 \in (a_2 \pm \frac{1}{2}\epsilon_2), \ldots, X_n \in (a_n \pm \frac{1}{2}\epsilon_n)) \approx f_{(X_1, X_2, \ldots, X_n)}(a_1, \ldots, a_n) \times \epsilon_1\epsilon_2 \ldots \epsilon_n.
\]

Here, the top line is just shorthand for \( X_1 \in (a_1 - \frac{1}{2}\epsilon_1, a_1 + \frac{1}{2}\epsilon_1) \) and so on. That is, the probability that \( X_1, \ldots, X_n \) are in some small \( n \)-dimensional set around the point \((a_1, \ldots, a_n)\) scales with the volume of the set, and \( f_{(X_1, \ldots, X_n)} \) is the constant of proportionality (the derivative). It is also the Radon-Nikodym derivative with respect to the Lebesgue measure on \( \mathbb{R}^n \) (the \( n \)-dimensional volume).

**Theorem 2.2.** If \( X_1, X_2, \ldots, X_n \) are independent, then for any real-valued functions \( g_1, g_2, \ldots, g_n \)

\[
\mathbb{E}[g_1(X_1)g_2(X_2) \ldots g_n(X_n)] = \mathbb{E}[g_1(X_1)]\mathbb{E}[g_2(X_2)] \ldots \mathbb{E}[g_n(X_n)]
\]

provided all of these expectations are defined. In particular, if \( X_i \) are real-valued,

\[
\mathbb{E}[X_1X_2 \ldots X_n] = \mathbb{E}[X_1]\mathbb{E}[X_2] \ldots \mathbb{E}[X_n].
\]

We saw on the last homework that in general, expectation of the product is *not* the product of expectations; but for independent random variables it is. Notice that it’s *still* the case that \( \mathbb{E}[g(X)] \neq g(\mathbb{E}[X]) \).

We will prove this result later, after we have carefully defined conditional expectation and that approach to independence.

**Warning:** The converse is not true: \( \mathbb{E}[X \cdot Y] = \mathbb{E}[X]\mathbb{E}[Y] \) does *not* mean that \( X \) and \( Y \) are independent. For example, suppose \( X \) is uniform on the set \( \{1, 2, \ldots, 9, 10\} \), and suppose that \( Y = \pm X \) with probability 1/2 for each of the possibilities. Then \( X \) and \( Y \) are definitely not independent: knowing \( Y \) tells you the value of \( X \) exactly (so \( X \) is \( \sigma(Y) \)-measurable), and knowing \( X \) severely restricts the possibilities for \( Y \). On the other hand, \( XY \) can be either \( X^2 \) or \( -X^2 \) with probability 1/2 each, so when I compute \( \mathbb{E}[XY] \) I will get 0, which is also \( \mathbb{E}[X]\mathbb{E}[Y] \).

(Although it is actually true that if the equality holds for *all* functions \( g_i \), then the random variables are independent.)

For two random variables, we define their **covariance** as

\[
\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].
\]

When the covariance is 0, we say the random variables are uncorrelated. (Correlation is related to covariance, but we won’t be using it directly in this course.) Notice that the **variance** of a random variable is its covariance with itself.

**Proposition 2.3.** Here’s an alternative expression for the covariance:

\[
\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].
\]
Proof. Expand the product, and recall that expectation is linear:
\[
\]

The point of this is to show that covariance isn’t about the mean of the random variables, it’s about oscillations about that mean.

Here are some other properties of covariance:

1. If \( X \) is a constant random variable, and \( Y \) is any other variable, then \( \text{Cov}(X,Y) = 0 \). This follows directly from independence. (There’s a partial converse: if \( \text{Var}(X) = \text{Cov}(X,X) \) is zero, then \( X = E[X] \) is a constant with probability 1.)

2. Covariance is \textit{bilinear}: \( \text{Cov}(aX + bY, cZ + dW) = ac \text{Cov}(X,Z) + bc \text{Cov}(Y,Z) + ad \text{Cov}(X,W) + bd \text{Cov}(Y,W) \). (Expand the product and check!)

3. As a consequence of this, \( \text{Cov}(X + a, Y + b) = \text{Cov}(X,Y) \) for any constants \( a \) and \( b \). We just saw one example of this when \( a = -E[X], b = -E[Y] \): the covariance of \( X \) and \( Y \) is equal to the covariance of those random variables \textit{centered} to have mean zero.

Here’s an important statement about independence and covariance:

Independent random variables are uncorrelated (have covariance 0). But uncorrelated random variables may not be independent.

3. NORMAL RANDOM VARIABLES

The normal random variable \( N(\mu, \sigma^2) \) has mean \( \mu \), variance \( \sigma^2 \), and pdf
\[
f_X(x) = \frac{1}{\sqrt{2\pi \sigma}} \exp \left( -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right).
\]

The \textit{standard} normal random variable is \( N(0,1) \).

The normal random variable will be important to our course, essentially for the following reason.

Let \( X_1, X_2, \ldots \) be independent real-valued random variables (hence defined on the same probability space), and suppose they have the same law. Such variables are called \textit{iid}, meaning independent and identically distributed. Let’s also suppose that \( E[X_i] = \mu \) and \( \text{Var}(X_i) = \sigma \). Then the average of many of the \( X_i \),
\[
\frac{1}{n} \sum_{i=1}^{n} X_i \to \mu, \quad \text{as } n \to \infty.
\]
that is, the average of many iid random variables converges to their mean. This is the law of large numbers. (We should be a little more precise by what we mean about convergence here, because we’re working with random variables. The correct statement here is that we have almost sure convergence, i.e. for almost all \( \omega \in \Omega \), \( \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} X_i(\omega) = \mu \).

However, suppose we’re interested in the deviations from that mean, then we have the (much harder!) central limit theorem:

\textbf{Theorem 3.1} (Central limit theorem).
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \Rightarrow N(0,1), \quad \text{as } n \to \infty.
\]
Here, what we mean is that the random variables on the left converge to the random variable on the right in distribution: if we plot the CDF of the left-hand side as a function of $n$, it will converge to the CDF of the $N(0, 1)$ random variable. I’ve used two different arrows ($\to$ and $\Rightarrow$) to differentiate the two notions of convergence.

There’s actually a more general statement of the central limit theorem, which allows the $X_i$ to have different means and variances, and even allows them to be slightly dependent, but we won’t go there. In practice, the central limit theorem says that anything that’s the sum of many small independent contributions should be approximately normal. For example, people’s height is influenced by many factors and is approximately normal; size of apples in an orchard tends to be approximately normal; and so on. (Another reason the normal distribution appears often is that it turns out to be a very nice mathematical object for proving results.)

Let’s come back to normal variables. Here’s one fundamental result about normal random variables:

**Proposition 3.2.** If $X$ is a standard normal variable, then $Y = \sigma X + \mu$ is a $N(\mu, \sigma)$ random variable.

**Proof.** To check that the random variables are equal, we need to compare their CDFs.

$$F_Y(a) = \mathbb{P}(Y \leq a) = \mathbb{P}(X < \frac{a - \mu}{\sigma})$$

$$= F_X((a - \mu)/\sigma) = \int_{-\infty}^{(a-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx$$

On the other hand, for a $N(\mu, \sigma^2)$ random variable, we have

$$F_{N(\mu, \sigma^2)}(a) = \int_{-\infty}^{a/\sigma} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2\right) \, dx$$

Let’s change variables in this second integral to $u = (x - \mu)/\sigma$, then $du = dx/\sigma$ so $dx = \sigma du$. Thus,

$$F_{N(\mu, \sigma^2)}(a) = \int_{-\infty}^{(a-\mu)/\sigma} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} u^2\right) \sigma du,$$

which after cancelling the $\sigma$ is exactly what we wanted to get. \qed

We will refer to the CDF of a standard normal variable as $\Phi$:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \, du.$$  

The fact that we can express all normal random variables through the standard normal variable should mean that many of the computations involving normal random variables might depend on $\Phi$ as well as $\mu$ and $\sigma$. For an example, see the homework solutions.
Important concepts from today: jointly normal, uncorrelated = independent for normals, Brownian motion

1. Review of last time

Last time, we discussed independence of random variables:

$$X$$ and $$Y$$ are independent means
$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

for all measurable sets $$A, B$$.

On a discrete set, this is something that you can verify directly by looking at the probability mass function. On $$\mathbb{R}$$, you can verify for $$A, B$$ that are intervals, and then it’ll be true on the entire Borel $$\mathcal{F}$$-algebra, because it’s generated by intervals.

We saw that if independent random variables have a density, then their joint density is

$$f_{(X_1, X_2, \ldots, X_n)}(a_1, \ldots, a_n) = f_{X_1}(a_1)f_{X_1}(a_1) \cdots f_{X_n}(a_n).$$

This makes it easy to show that

$$E[g_1(X_1) \cdots g_n(X_n)] = E[g_1(X_1)] \cdots E[g_n(X_n)]$$

for independent random variables with density, but it’s actually true for all independent random variables.

A corollary of this is that independent random variables are uncorrelated:

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y] = 0$$

if $$X$$ and $$Y$$ are independent.

The converse isn’t true: there are many examples on the homework.

We saw that independence is the opposite of measurability: if $$X$$ and $$Y$$ are independent, then knowing $$X$$ tells you nothing about $$Y$$; if $$Y$$ is $$\sigma(X)$$-measurable, then knowing $$X$$ tells you everything about $$Y$$. There is a lot of room in between. We saw that if $$Y$$ is both independent of $$X$$ and $$\sigma(X)$$-measurable, then $$Y$$ is a constant.

Finally, we defined the random normal variable $$N(\mu, \sigma^2)$$ as a random variable with pdf

$$f_{N(\mu, \sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\frac{(x - \mu)^2}{\sigma^2}\right),$$

and showed that if $$X \sim N(0, 1)$$ is a standard random normal variable, then $$\mu X + \sigma \sim N(\mu, \sigma^2)$$.
2. Multivariate normals

We say that \((X_1, \ldots, X_n)\) are jointly normal or multivariate normal with mean \(\mu = (\mu_1, \ldots, \mu_n)\) and covariance \(\Sigma = (\Sigma_{ij}, 1 \leq i, j \leq n)\) if their joint density is

\[
f_{X_1,\ldots,X_n}(a_1, \ldots, a_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp \left( -\frac{1}{2} (a - \mu)^T \Sigma^{-1} (a - \mu) \right).
\]

Here, \(^T\) is the transpose of a vector. In longhand notation, if \((\Sigma^{-1})_{ij}\) is the \(ij\) entry of the matrix \(\Sigma^{-1}\), then inside the exponential is

\[
-\frac{1}{2} \sum_{i,j} (\Sigma^{-1})_{ij} (a_i - \mu_i)(a_j - \mu_j).
\]

The vector \(\mu\) is the mean of \((X_1, \ldots, X_n)\), that is, \(E[X_i] = \mu_i\). The covariance matrix \(\Sigma\) contains the covariances, i.e. \(\text{Cov}(X_i, X_j) = \Sigma_{ij}\).

The normal pdf has level sets (sets where \((X_1, \ldots, X_n)\) is equally likely to be) that look like ellipsoids (or, in \(\mathbb{R}^2\), just ellipses). With a bit of effort, you can read off the covariance matrix \(\Sigma\) from the length and orientation of the axes of the ellipsoids.

**Reminder: how to invert a matrix.** If \(M\) is a matrix, then to find \((M^{-1})_{ij}\) you compute \(\det(M)\) without row \(j\) or column \(i\), multiply by \((-1)^{i+j}\), and divide by \(\det M\).

In particular, for a \(2 \times 2\) matrix,

\[
M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \implies M^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.
\]

**Proposition 2.1.** If \(X_1, \ldots, X_n\) are jointly normal, then each \(X_i\) is normal with mean \(\mu_i\) and variance \(\Sigma_{ii} = \text{Cov}(X_i, X_i)\).

In fact, a much stronger statement is true:

**Theorem 2.2.** If \(X_1, \ldots, X_n\) are jointly normal, then any linear combination of the \(X_i\) is univariate normal:

\[
\sum_i c_i X_i \sim N \left( \sum_i c_i \mu_i; \sum_{i,j} c_i c_j \Sigma_{ij} \right).
\]

Conversely, if the collection of random variables \(X_1, \ldots, X_n\) has the property that any linear combination is univariate normal, then \(X_1, \ldots, X_n\) are jointly normal, i.e. have the multivariate normal pdf.

We’ll show this later, but for now let’s check that we got the right mean and variance in the theorem:

\[
E[\sum_i c_i X_i] = \sum_i c_i E[X_i] = \sum_i c_i \mu_i
\]

by linearity of expectation, and

\[
\text{Var}(\sum_i c_i X_i) = \text{Cov}(\sum_i c_i X_i, \sum_i c_i X_i) = \sum_{i,j} c_i c_j \text{Cov}(X_i, X_j) = \sum_{i,j} c_i c_j \Sigma_{ij}
\]

by bilinearity of covariance.

We can extend the statement even further: if \((X_1, \ldots, X_n) \sim N(\mu, \Sigma)\) and \(A\) is an \(m \times n\) matrix, then \(AX\) is normal, with mean \(A\mu\) and covariance matrix \(A\Sigma A^T\). Notice that we multiply \(\mu\) by \(A\) once, and we multiply covariance by \(A\) twice, just as in the single-variable case. Again, we’ll show this later.
Theorem 2.3. If $X_1, \ldots, X_n$ are independent random variables, then they are jointly normal with diagonal covariance matrix, i.e. $\Sigma_{ii} = \sigma_i^2$ and $\Sigma_{ij} = 0$ for $i \neq j$. Conversely, if $X_1, \ldots, X_n$ are jointly normal and uncorrelated, then they are independent.

Proof. Joint density of independent normals looks like this:

$$f(x_1, \ldots, x_n)(a_1, \ldots, a_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left( -\frac{1}{2} \left( \frac{a_i - \mu_i}{\sigma_i} \right)^2 \right).$$

To find the joint density of multivariate normal with diagonal covariance matrix, notice that $\det \Sigma = \prod_i \sigma_i^2$, and $\Sigma^{-1}$ has $\sigma_i^{-2}$ on the diagonal and 0 elsewhere. Thus, the density is

$$f(x_1, \ldots, x_n)(a_1, \ldots, a_n) = \frac{1}{(2\pi)^{n/2} \prod_i \sigma_i} \exp \left( -\frac{1}{2} \sum_i (a_i - \mu_i)^2 \sigma_i^{-2} \right),$$

because the off-diagonal terms inside the exponential vanish. These two are equal, proving the equivalence. □

Thus, we have shown:

For jointly normal random variables, uncorrelated $\implies$ independent.

Warning: We already know that this isn’t true for non-normal random variables. It can even fail to be true for normal random variables if they aren’t jointly normal. For example, if $X \sim N(0,1)$ and $Y = X$ or $-X$ with probability 1/2 each, then both $X$ and $Y$ are normal, but they’re clearly not independent: for example,

$$\mathbb{P}(X \in (-1, 1), Y \in (-1, 1)) = \mathbb{P}(X \in (-1, 1)) = \mathbb{P}(Y \in (-1, 1)) = p.$$ 

Since $0 < p < 1$, we see that $p \neq p^2$, which is what we’d need for independence.

Coming back to jointly normal variables, the standard multivariate normal is $(X_1, X_2, \ldots, X_n) \sim N(0, I_n)$ with the pdf

$$f(x_1, x_2, \ldots, x_n)(a_1, a_2, \ldots, a_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2} (a_1^2 + a_2^2 + \ldots + a_n^2)}.$$ 

The matrix $I_n$ is the $n \times n$ identity matrix. Similarly to the univariate case, we would like to show that all multivariate normals can be obtained from the standard multivariate normal. This is true.

Proposition 2.4. If $X = (X_1, \ldots, X_n)$ is standard multivariate normal, $\mu$ is a vector, and $S$ is an $n \times n$ matrix, then $Y = \mu + SX$ is normal with mean $\mu$ and covariance $SS^T$.

Let’s try to prove this in the same way we did for univariate normals:

$$f(y_1, \ldots, y_n)(a_1, \ldots, a_n) \approx \frac{1}{(2\epsilon_1)(2\epsilon_2)\ldots(2\epsilon_n)} \mathbb{P}(y_1, \ldots, y_n = (a_1 \pm \epsilon_1, \ldots, a_n \pm \epsilon_n))$$

write this as $Y \in a \pm \epsilon$

$$= \frac{1}{(2\epsilon_1)(2\epsilon_2)\ldots(2\epsilon_n)} \mathbb{P}(SX \in a - \mu \pm \epsilon)$$

$$= \frac{1}{(2\epsilon_1)(2\epsilon_2)\ldots(2\epsilon_n)} \mathbb{P}(X \in S^{-1}(a - \mu \pm \epsilon)).$$
We now have a slight hitch: what happens if $S$ isn’t invertible? Let’s pretend $S$ is invertible for now, and move on with life:

$$= \frac{1}{(2\pi)^{n/2}} \exp\left( -\frac{1}{2} \sum_i (S^{-1}(a - \mu)_i^2 \right) \times \text{Volume}(S^{-1}(\pm \epsilon)) \frac{(2\epsilon_1)(2\epsilon_2) \ldots (2\epsilon_n)}{(2\pi)^n}$$

How much does $S^{-1}$ change the volume of a box? By a factor of $|\det(S^{-1})|$.

$$= \frac{1}{(2\pi)^{n/2}} \exp\left( -\frac{1}{2} \sum_i (S^{-1}(a - \mu)_i^2 \right) \times |\det(S^{-1})|$$

$$= \frac{1}{(2\pi)^{n/2} |\det(S)|} \exp \left( -\frac{1}{2} \sum_i \sum_j \sum_k (S^{-1})_{ij}(S^{-1})_{ik}(a_j - \mu_j)(a_k - \mu_k) \right)$$

This is exactly what we should see, since $\det(S) = \sqrt{\det(SS^T)}$.

We now have two questions:

1. Can we get all covariance matrices $\Sigma$ as $SS^T$? In one dimension this was very easy: to go from variance $\sigma^2$ to $\sigma$ all we needed to do was take the square root.

   In matrices the answer is still yes, but matrix square roots are a little harder.

   We’ll not talk about this now, but I’ll post a separate document discussing the question.

2. What happens when $S$ (and then $\Sigma = SS^T$) is not invertible?

Let’s look at a toy example of a non-invertible matrix $\Sigma$:

$$\Sigma = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

If $(X, Y, Z)$ is normal with mean $(\mu_X, \mu_Y, \mu_Z)$ covariance $\Sigma$, then $X$, $Y$, and $Z$ should be independent (because all off-diagonal entries are 0), $X$ should be $N(\mu_X, 2)$, $Y$ should be $N(\mu_Y, 1)$, and $Z$ should be $N(\mu_Z, 0)$. That is, $Z$ is a constant, and the entire distribution lives in the plane $z = \mu_Z$. If we look at just the plane, i.e. just the points of the form $(x, y, \mu_Z)$, then we’ll see a normal distribution on the plane, with covariance $\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$.

This means that $(X, Y, Z)$ will not have a density in $\mathbb{R}^3$, because the probability of belonging to a small box around, say, $(\mu_X, \mu_Y, \mu_Z)$ doesn’t scale with the third dimension of the box – it scales as area rather than as volume. But we’ll still say that $(X, Y, Z)$ is jointly normal. We also see that we can write

$$(X, Y, Z) = (\mu_X, \mu_Y, \mu_Z) + \begin{pmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} W_1 \\ W_2 \\ 0 \end{pmatrix},$$
where \((W_1, W_2)\) is a standard normal in \(\mathbb{R}^2\). Since we don’t actually need a third random variable, we put a 0 in for it.

In general, if \(\Sigma\) is not invertible, this will happen in some basis. That is, there will be some linear combinations of \(X_1, \ldots, X_n\) that are constant, and then there will be an affine subspace of \(\mathbb{R}^n\) on which we get a multivariate normal distribution. We will be able to write \((X_1, \ldots, X_n)\) as a linear combination of \((W_1, \ldots, W_d)\) with \(d < n\), where \((W_1, \ldots, W_d)\) will be a standard normal in \(\mathbb{R}^d\). (To write this in matrix form, we’ll need to pad the \(W\) with extra zeros.)

In real life, singular (non-invertible) correlation matrices mean that there are fewer underlying sources of randomness than the number of random variables you’re observing. (Actually, in real life you’re likely to have not \(\det(\Sigma) = 0\) but \(\det(\Sigma)\) very small.)

3. **Stochastic processes and Brownian motion**

We’ve already talked about finite collections of random variables \(X_1, \ldots, X_n\), and about their joint properties. We’ll now graduate to stochastic processes, which is the same thing with more variables:

\[
(X_t : t \in T) = (X_t(\omega), t \in T, \omega \in \Omega).
\]

Here, we can talk about the random variable \(X_t\) defined on \(\Omega\); but we will also often talk about the trajectory: that is, the function of time \(f(t) = X_t(\omega)\). Here, we fix the element of \(\Omega\) and look at what happens to \(X\) on it at all times.

**Example 1.**

1. Here’s a finite example: suppose \(\Omega = \{U, D\}^{16}\), and each one of \(X_1, \ldots, X_{16}\) reads off the corresponding letter of the sequence. Then \((X_1, \ldots, X_{16})(\omega)\) is the entire random sequence of \(U\)’s and \(D\)’s, but we also can talk about \(X_4\), the distribution of just the 4th letter. (And by the way, while we’ve usually talked about the \(X_i\) as if they were independent, they don’t have to be.)

2. An infinite, but still discrete example is a time series: \(t = 0, \pm 1, \pm 2, \ldots\). Let \(\phi\) be some real number, and let \(Z_t\) be an iid sequence. We could have a process that satisfies

\[
X_t = \phi X_{t-1} + Z_t, \quad t \in \mathbb{Z}, \quad Z_t \text{ iid sequence};
\]

this is an autoregressive process of order 1. Again, we can think about the trajectory, which will satisfy \(x_t = \phi x_{t-1} + z_t\), or we can think about the distribution of \(X_t\) at time \(t\), which might, for example, be normal.

3. One of the most important processes for us will be the Gaussian process. Here, \(t = \mathbb{R}\). We give an indirect definition: a stochastic process is Gaussian if all the finite-dimensional distributions are Gaussian, i.e. if

\[
(X_{t_1}, \ldots, X_{t_n}) \text{ are jointly normal for all } t_1, \ldots, t_n.
\]

The distribution of a Gaussian process is determined by the means and covariance matrices of these finite-dimensional distributions; so it’s enough to specify

\[
\mu_X(t) = \mathbb{E}[X_t], \quad c_X(t, s) = \text{Cov}(X_t, X_s).
\]

At this point you should be worried: I’ve suggested that all the variables \(X_t\) are defined on the same probability space; can I actually do that? That is, will there be a probability space on which I can define a bunch of random variables (functions) which satisfy all the constraints for a Gaussian process, i.e. which are jointly normal? The answer is yes, but it’s not an easy theorem.

We will care about a particular Gaussian process called Brownian motion. Here are its properties:
(1) Mean and covariance functions of Brownian motion: \( \mu_X(t) = 0,\ c_X(t,s) = \min(t,s). \) In particular, \( B_0 = 0. \)

(2) Finite-dimensional distributions:
\[
(X_{t_1}, \ldots, X_{t_n}) = N(0, \Sigma)
\]
where \( \Sigma_{ij} = t_i \) if \( i \leq j \). In particular, \( \mathbb{E}[X_t^2] = \mu_t^2 + c_X(t,t) = t, \) so \( X_t \sim N(0,t) \).

(3) Increments are stationary: the distribution of \( X_{t+h} - X_t \) is the same as the distribution of \( X_h \). We can check this: \( X_t \) and \( X_{t+h} \) are jointly normal with mean \( (0,0) \) and covariance matrix \( \begin{pmatrix} t & t \\ t & t+h \end{pmatrix} \), so \( X_{t+h} - X_t \) is normal with mean \( 0 - 0 = 0 \) and variance \( \Sigma_{22} + \Sigma_{11} - 2\Sigma_{12} = h. \) This is also the distribution of \( X_h. \) Notice that \( X_{t+h} - X_t \) and \( X_h \) are different random variables, they may or may not be independent, and we’re just saying they have the same probability distribution.

(4) Increments are independent: for all \( t_1, \ldots, t_n, \) the \( n - 1 \) differences
\[
X_{t_2} - X_{t_1},\ X_{t_3} - X_{t_2},\ \ldots,\ X_{t_n} - X_{t_{n-1}}
\]
are independent.

(5) The sample paths are continuous. Here’s the heuristic argument for this. Because increments are stationary, it’s enough to check that \( X_h \to 0 \) as \( h \to 0, \) and then we’ll automatically get \( X_{t+h} \to X_t \) as \( h \to 0. \) The convergence \( X_h \to 0 \) follows because \( \mathbb{E}[X_h] = 0 \) and \( \text{Var}(X_h) = h \to 0, \) so \( \mathbb{E}[X_h^2] = h \to 0 \) and \( X \) can’t be large with any sizeable probability. What this argument shows is
\[\mathbb{P}(X_h > \epsilon) \to 0 \text{ as } h \to 0,\]
which is not at all the same thing as saying that \( X_h(\omega) \to X_0(\omega) = 0 \) for any particular \( \omega. \) You can’t actually prove that sample paths must be continuous, because all the other properties of Brownian motion allow them to not be continuous on a set of measure 0.

(6) In addition to independent increments, we have the much stronger Markov property: the “past” trajectory \( (X_s : s < t) \) is independent of the future trajectory \( (X_{t+s} - X_t : s > 0), \) except for the initial value \( X_t. \)

Here’s what’s enough to conclude that something is a Brownian motion:

**Theorem 3.1** (Lévy). Any process satisfying the following properties is a Brownian motion (and thus has all the properties listed above and below):

- \( B_0 = 0; \)
- The increments are independent;
- \( \mathbb{E}[B_t - B_s] = 0 \) for all \( t \) and \( s, \) and \( \text{Var}(B_t - B_s) = t - s \) for \( s \leq t; \)
- Sample paths are continuous.

Notice that you don’t even need to require the increments to be Gaussian – that will happen automatically! It’s a hard theorem.

Here are some more properties of Brownian motion:

(1) The paths of a Brownian motion are continuous, by definition. However, almost all sample paths are nowhere differentiable. A derivative would mean that \( (B_{t+h} - B_t) \approx dh \) for some constant \( d \) (the value of the derivative). However, heuristically, we see that \( B_{t+h} - B_t \) has mean \( 0 \) and variance \( h, \) which means that it has standard deviation \( \sqrt{h}, \) which means that it scales roughly as \( \sqrt{h} \) rather than \( h. \) Since \( \sqrt{h} \) is much bigger than \( h \) for small \( h, \) Brownian motion doesn’t have a derivative anywhere.
This argument isn’t a proof of the non-differentiability (it could be that some sample paths of Brownian motion are differentiable, while others are very badly not so). There’s a proof in Mikosch, Appendix A3. For us, the take-away message is that the size of $B_h$ is approximately $\sqrt{h}$, which leads to problems when $h \approx 0$. 
Important concepts from today: Brownian motion is invariant under translation, scaling, and time inversion; Brownian motion with drift, Brownian bridge, geometric Brownian motion.

1. REVIEW OF LAST TIME

Last time, we joint normality: we say \((X_1, \ldots, X_n) \sim N(\mu, \Sigma)\) if

\[
f_{(X_1, \ldots, X_n)}(a_1, \ldots, a_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp \left( -\frac{1}{2} \sum_{i,j} (a_i - \mu_i)(\Sigma^{-1})_{ij}(a_j - \mu_j) \right)
\]

Note the \(\sqrt{\det \Sigma}\) in the denominator – in last lecture we had the typo involving \(\det \Sigma\) there. We have \(E[X_i] = \mu_i\) and \(\text{Cov}(X_i, X_j) = \Sigma_{ij}\).

We said that any linear transformation of the vector \((X_1, \ldots, X_n)\) will then also be jointly normal; the mean and covariance matrix can be determined from linearity of mean and bilinearity of covariance. In particular, linear combinations of the \(X_i\) are univariate normal, and any subset of the \(X_i\) are jointly normal. We also established that we can write \(X\) as \(\mu + SY\) where \(Y \sim N(0, I)\) is a standard multivariate normal, and \(S\) is a matrix such that \(SS^T = \Sigma\). We discussed briefly what happens when \(\Sigma\) is not invertible: in that case, the “multivariate normal in \(\mathbb{R}^n\)” will actually only take values in some subspace (e.g. a line or a plane), and this corresponds to \(Y\) being standard normal in lower dimension (fewer sources of randomness than dimensions).

We established that for jointly normal random variables, independence is the same thing as being uncorrelated.

We then moved on to talk about stochastic processes, and particularly Brownian motion. We saw that a stochastic process is a collection of random variables indexed by \(t\), which can come from any set but for us will usually be time (either discrete or continuous). The random variables are defined on the same probability space, and usually we discuss the properties of trajectories (the collection of values \(X_t(\omega)\) for one fixed \(\omega\) and all \(t\)) or marginals (the random variable \(X_t\) for a fixed \(t\) and all \(\omega\)). For a Brownian motion, we had the following characterization:

1. \(B_0 = 0\).
2. Finite-dimensional distributions are jointly normal: \((B_{t_1}, \ldots, B_{t_n}) \sim N(0, \Sigma)\) where \(\Sigma_{ij} = \min(t_i, t_j)\).
3. Increments are stationary and independent: the distribution of the random variable \(B_{t+h} - B_t\) is the same as the distribution of \(B_h - B_0 = B_h\) (stationary), and the increments \(B_{t_i} - B_{s_i}\) corresponding to nonoverlapping intervals \((s_i, t_i)\) are independent.
Sample paths are continuous, but nowhere differentiable. (Strictly speaking, we should say that this holds on a set of $\omega \in \Omega$ that has probability 1.) This was essentially because $B_h \sim \sqrt{h}$, so $h^{-1}(B_h - B_0)$ diverges.

We had Lévy’s theorem, which told us minimalistic conditions for showing that something is a Brownian motion.

**Theorem 1.1** (Lévy). *Any process satisfying the following properties is a Brownian motion (and thus has all the properties listed above and below):*

- $B_0 = 0$;
- The increments are independent;
- $\mathbb{E}[B_t - B_s] = 0$ for all $t$ and $s$, and $\text{Var}(B_t - B_s) = t - s$ for $s \leq t$;
- Sample paths are continuous.

2. **Some more properties of Brownian motion**

1. If $B_t$ is a Brownian motion, then $W_t = B_{t+t_0} - B_{t_0}$ is also a Brownian motion. That is, Brownian motion is translation-invariant. Let’s check it based on Lévy’s theorem:

   $W_0 = B_{t_0} - B_{t_0} = 0$.

   Increments $W_{t_i} - W_{s_i} = B_{t_0+t_i} - B_{t_0+s_i}$ are independent: if $(s_i, t_i)$ are nonoverlapping, so are $(t_0 + s_i, t_0 + t_i)$.

   This also shows that the mean and variance of the intervals is correct, because $B$ had stationary increments, so shifting them by $t_0$ does nothing to the mean and variance.

   Sample paths are still continuous.

2. Brownian motion is self-similar: if $B_t$ is a Brownian motion, then so is $W_t = \frac{1}{\sqrt{c}}B_{ct}$, for any $c > 0$.

   Proof: we check each of the four properties.

   $W_0 = c^{-1/2}B_{c \cdot 0} = 0$, since $B_0 = 0$.

   The increments are still independent (that isn’t affected by rescaling time or multiplying by constants).

   The sample paths are still continuous (that isn’t affected by continuous changes of time scale, or by multiplication by constants).

   Finally, $\mathbb{E}[W_t - W_s] = c^{-1/2}\mathbb{E}[B_{ct} - B_{cs}] = 0$, and

   $\text{Var}(W_t - W_s) = c^{-1}\text{Var}(B_{ct} - B_{cs}) = c^{-1} \cdot c(t - s) = (t - s)$

   as required.

3. Finally, Brownian motion is invariant under time inversion in the following sense: if $B_t$ is a Brownian motion, then so is $W_t = tB_{1/t}$. (We define $W_0 = 0$.) Checking continuity at 0 here will be a little harder.
Important concepts from today: conditional expectation as the best measurable approximation; rules for computing conditional expectation; Brownian motion is a martingale.

1. Conditional expectation

Conditional expectation expresses the following intuitive idea: if I get some additional information about a random variable (such as stock price), then I should be able to incorporate this additional knowledge into my beliefs about the random variable.

A couple of examples where conditional expectation is a natural thing:

(1) I roll two dice, and tell you that one of the dice came up 3. I would like to know the distribution of the sum of the rolls conditioned on this information.

(2) I believe that stock prices are correlated across time (for example, stock prices behave approximately like a Brownian motion). I have observed the stock price up until today, and would like to predict the price tomorrow.

In elementary probability theory we learned to condition events on a single event of positive probability:

\[ P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0. \]

This would let us answer the first of the above questions, in that it lets us compute \( P(X = a) \) for all values of \( a \). But it doesn’t let us do anything with the second, because any particular trajectory of a Brownian motion has probability zero! So, what to do?

(Note: the answer to the second question should be that the price tomorrow is normal, with mean equal to the price today, and standard deviation equal to the daily oscillation of the stock price. This is because Brownian motion has independent increments, so the difference (price tomorrow – price today) is normal and independent of the price today.)

First, I want to rephrase the basic conditional expectation. Begin by aggregating the probability information into a CDF:

\[ F_{X|B}(a) = \frac{P(X \leq a \text{ and } B)}{P(B)}. \]

This is the CDF of a random variable \( X|B \), which looks exactly like \( X \) but is defined only on \( B \). (Note: Mikosch writes this as \( F_X(a|B) \).

Next, we compute the expected value of this new random variable:

\[ \mathbb{E}[X|B] = \frac{\mathbb{E}[X 1_B]}{P(B)} = \int_B X(\omega) d\mathbb{P}(\omega) \frac{1}{P(B)}. \]

For example, if \( \Omega \) is a discrete space, then \( \mathbb{E}[X|B] \) is computed by averaging only those values of \( X \) that correspond to the event \( B \) occurring (and, of course, dividing by \( P(B) \)).

Date: 10/23/2014.
Now, suppose that we have several events $A_1, A_2, \ldots, A_n$ which together partition the space. That is, $\bigcup_i A_i = \Omega$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$. Consider the collection of numbers $\mathbb{E}[X | A_i]$. These define a new random variable: namely, for each $\omega \in \Omega$, figure out which $B_i$ contains $\omega$, and then take $\mathbb{E}[X | A_i]$.

Example 1 (Conditional expectation of uniform random variable). Let $X : (0, 1] \to (0, 1]$ be given by $X(\omega) = \omega$, where the underlying probability measure is uniform. Then $X$ is uniformly distributed on $(0, 1]$. Consider the events $A_1, \ldots, A_5$ given by $A_i = ((i - 1)/5, i/5]$. Then

$$\mathbb{E}[X | A_i] = \frac{\mathbb{E}[X 1_{A_i}]}{\mathbb{P}(A_i)} = \frac{\int_{(i-1)/5}^{i/5} ada}{1/5} = \frac{12i - 1}{2} \cdot \frac{5}{5}.$$ 

These numbers define a new random variable, whose graph is given below:

![Figure 1. Uniform random variable conditioned on the five events $A_i$.](image)

What we’ve just done was compute the conditional expectation of $X$ based on a $\sigma$-algebra generated by the sets $A_i$. (Actually, it’s not clear so far that we needed a $\sigma$-algebra structure, we just had the collection $\{A_1, \ldots, A_5\}$, but the $\sigma$-algebra structure will be useful momentarily.) Another way of saying it is that if $Y$ is a random variable such that $\sigma(Y) = \sigma(A_i)$, then we’ve just computed $\mathbb{E}(X|\sigma(Y)) = \mathbb{E}(X|Y)$. (We usually write $\mathbb{E}(X|Y)$ rather than $\mathbb{E}(X|\sigma(Y))$.)

**Properties of conditional expectation:**

1. $\mathbb{E}(X|Y)$ is a $\sigma(Y)$-measurable random variable. That is, $\mathbb{E}(X|Y)$ will be a function of $Y$ (which function is controlled by $X$).
2. If $A \in \sigma(Y)$, then $\mathbb{E}[\mathbb{E}(X|Y)1_A] = \mathbb{E}[X 1_A]$. In particular, to compute the value of $\mathbb{E}(X|Y)$ on one of the smallest measurable sets, we simply average $A$ over that set. But also, $\mathbb{E}[\mathbb{E}(X|Y)1_{A_i \cup A_j}] = \mathbb{E}[X 1_{A_i \cup A_j}]$: the average of $\mathbb{E}(X|Y)$ over the union of several of the $A_i$’s is equal to the average of $X$ over them, because I can compute the average of $X$ over the union of several $A_i$’s by integrating over each of the $A_i$ first, and then taking the sum. (This is where the $\sigma$-algebra structure is coming into play.)
3. In particular, $\mathbb{E}[\mathbb{E}(X|Y)] = \mathbb{E}[X]$ (take $A = \Omega$). Notice that $\mathbb{E}(X|Y)$ is a random variable, and $\mathbb{E}[\mathbb{E}(X|Y)]$ is a number. I’ll try to stick to having round parentheses for conditional expectation (which returns a random variable), and square brackets for expectation (which returns a number).
4. Conditional expectation is linear: $\mathbb{E}[c_1 X_1 + c_2 X_2 | Y] = c_1 \mathbb{E}[X_1 | Y] + c_2 \mathbb{E}[X_2 | Y]$. Notice that there is no linearity in the second argument: this is because we don’t care about the values that $Y$ takes, we only care about the information those values convey about $\omega \in \Omega$ (i.e. we care about $\sigma(Y)$ in the domain, not the values of $Y$ in the range).
5. If $X$ is already $\sigma(Y)$-measurable, then $\mathbb{E}(X|Y) = X$. 

2
Theorem 1.1. Conditional expectation exists, provided measure zero is zero. So, where do we get information on $\mathbb{E}$ sets of positive measure, but here is the problem when replace $E$ value of $E$ obvious that there will be a way to define answer, of course, is from sets of positive measure that contain $E$ expectation: $X$ refers to the filtration with just one letter $F$.

Now let’s extend the idea of conditioning on a $\sigma$-algebra.

Definition 1. Let $X : \Omega \to T$ be a random variable, and let $\mathcal{F}$ be a $\sigma$-algebra on $\Omega$. Then $\mathbb{E}(X|\mathcal{F})$ is a $\mathcal{F}$-measurable random variable, such that for every $A \in \mathcal{F}$, $\mathbb{E}[\mathbb{E}(X|\mathcal{F})1_A] = \mathbb{E}[X1_A]$.

All of the properties we stated above hold for this more general conditioning (if we replace $Y$ and $\sigma(Y)$ by $\mathcal{F}$ throughout).

Theorem 1.1. Conditional expectation exists, provided $\mathbb{E}(|X|) < \infty$.

We know that it exists when the $\sigma$-algebra consists of finitely (or countably) many sets of positive measure, but here is the problem when $A \in \mathcal{F}$ has measure 0. Then $\mathbb{E}[X1_A] = 0$ and $\mathbb{E}E(X|\mathcal{F})1_A = 0$ automatically: the integral of anything over a set of measure zero is zero. So, where do we get information on $\mathbb{E}(X|\mathcal{F})(\omega)$ for $\omega \in A$? The answer, of course, is from sets of positive measure that contain $A$, but it is by no means obvious that there will be a way to define $\mathbb{E}(X|\mathcal{F})$ consistently on all of them.

We will not prove this theorem, but here is an alternative characterisation of conditional expectation:

Theorem 1.2. Let $X$ be a random variable, and let $\mathcal{F}$ be a $\sigma$-algebra on the underlying probability space. Then among all the $\mathcal{F}$-measurable random variables $Y$, the smallest value of $\mathbb{E}[(X - Y)^2]$ is achieved by $Y = \mathbb{E}(X|\mathcal{F})$.

This tells us that in some sense, the best approximation of $X$ given information contained in $Y$ is to take the conditional expectation (not the mode or median). It also suggests a way of proving that conditional expectation exists: we would just need to show that the function $\mathbb{E}[(X - Y)^2]$ on the set of all $\mathcal{F}$-measurable random variables attains its minimum.

2. Martingale

Definition 2. A filtration is a collection of increasing (finer) $\sigma$-algebras $\mathcal{F}_s$. We tend to refer to the filtration with just one letter $\mathcal{F}$, but we mean the entire collection.
The easiest way to get a filtration is to let $\mathcal{F}_s = \sigma(X_u, u \leq s)$ for some stochastic process $X$. That is, $\mathcal{F}_s$ will contain all information about the trajectories of $X$ up to time $s$.

**Definition 3.** Given a filtration $(\mathcal{F}_t)$, a stochastic process $(X_t)$ is a martingale if:

1. $X_t$ is $\mathcal{F}_t$-measurable. That is, $\sigma(X_t) \subseteq \mathcal{F}_t$. (If $\mathcal{F}_t$ is generated by the trajectories of some stochastic process up to time $t$, then given those trajectories we must be able to determine the trajectories of $X$ up to time $t$.) This is called *adapted.*

2. $$\mathbb{E}(X_s | \mathcal{F}_s) = X_s, \quad \forall s \leq t.$$ Equivalently, $\mathbb{E}(X_t - X_s | \mathcal{F}_s) = 0$ if $s \leq t$. (Here, we mean the constant random variable $0$.)

Often, but not always, the filtration is the filtration generated by $X$ itself, i.e. $\mathcal{F}_s = \sigma(X_u, u \leq s)$. It’s clear that $X$ will be adapted to its own filtration, but sometimes it’s easier to prove that the conditional expectation does not change with respect to a finer filtration (one that tells apart more events).

Notice that by the property of conditional expectation, if $X$ is a martingale, then $\mathbb{E}[X_t] = \mathbb{E}[X_0]$ at all times $t$. That is, the overall expectation of $X$ does not change from its starting value.

**Example 2.** (1) Brownian motion is a martingale with respect to its own filtration. Indeed:

$$\mathbb{E}(B_t | \sigma(B_u, u \leq s)) = \mathbb{E}(\underbrace{B_s}_{\text{F-measurable}} + \underbrace{(B_t - B_s)}_{\text{independent of } \mathcal{F}} | \sigma(B_u, u \leq s)) = B_s + \mathbb{E}[B_t - B_s] = B_s.$$ Here the answer was true because the increments of Brownian motion are independent, but this doesn’t always have to be like that.

(2) But if we change filtration, we might not get a martingale. For example, let $\mathcal{G}_s = \sigma(B_u, u \leq 2s)$, then $B$ is not a martingale with respect to $\mathcal{G}$ because $\mathbb{E}[B_t | \mathcal{G}_s] = B_t$ for $t \leq 2s$ (since $B_t$ is $\mathcal{G}_s$-measurable). That is, if we could see the future, then Brownian motion wouldn’t actually have random mean-zero fluctuations.

(3) $X_t = B_t^2$ (where $B$ is a Brownian motion) is not a martingale with respect to anything: we have $X_0 = 0$ but $\mathbb{E}[X_t] > 0$. Let $Y_t = B_t^2 - t$; then $(Y_t, t \geq 0)$ is a martingale with respect to the filtration given by $\mathcal{F}_s = \sigma(B_u, u \leq s)$. This is not the natural filtration of $Y$; it is finer, since it tells the sign of $B_s$.

We check:

$$\mathbb{E}(Y_t | \mathcal{F}_s) = \mathbb{E}((B_s + (B_t - B_s))^2 - t | \mathcal{F}_s)$$

$$= \mathbb{E}(\underbrace{B_s^2}_{\text{F-s-measurable}} + 2 \underbrace{B_s (B_t - B_s)}_{\text{independent of } \mathcal{F}_s} + \underbrace{(B_t - B_s)^2}_{\text{independent of } \mathcal{F}_s} - t | \mathcal{F}_s)_{\text{non-random}}$$

$$= B_s^2 + 2B_s \mathbb{E}[B_t - B_s] + \mathbb{E}[(B_t - B_s)^2] - t = B_s^2 - s.$$ This is exactly what we wanted to show.

Notice that $Y$ is also a martingale with respect to any coarser filtration to which it’s adapted, including its own natural filtration. This follows from the tower property of conditional expectation:

$$\mathbb{E}(Y_t - Y_s | \sigma(Y_u, u \leq s)) = \mathbb{E}(\mathbb{E}(Y_t - Y_s | \mathcal{F}_s) | \sigma(Y_u, u \leq s)) = 0.$$
That is, if you can show that the increments of a process have mean zero using extra information, then they have mean zero without the extra information as well.

(4) This is worth highlighting: if \( X_t \) is a martingale with respect to \( \mathcal{F} \), and \( \mathcal{G} \) is coarser than \( \mathcal{F} \) but \( X_t \) is still measurable with respect to \( \mathcal{G} \), then \( X_t \) is a martingale with respect to the coarser filtration as well.

(5) We had some friends of Brownian motion that we never got to meet, so we might as well meet them now. Consider the geometric Brownian motion \( X_t = \exp(\mu t + \sigma B_t) \). This is the Black–Scholes model for stock prices, which has the advantage of always being nonnegative. In particular, it’s not a Gaussian process, although it’s natural filtration is the same as the filtration of the Brownian motion (because the map \( B_t \mapsto X_t \) is invertible). Let’s begin by computing the expectation:

\[
\mathbb{E}[X_t] = \mathbb{E}[e^X], \quad \text{where } X \sim N(\mu t, \sigma^2 t).
\]

We’ve computed these sorts of expectations already on a homework; the answer turns out to be

\[
\mathbb{E}[X_t] = e^{\mu t + \frac{1}{2} \sigma^2 t}.
\]

For good measure, let’s see how to compute the covariance function:

\[
\begin{align*}
\mathbb{E}[X_t X_s] - \mathbb{E}[X_t] \mathbb{E}[X_s] &= \mathbb{E}[e^{\mu(t+s)}e^{\sigma(B_t+B_s)}] - \mathbb{E}[e^{\mu t}e^{\sigma B_t}]\mathbb{E}[e^{\mu s}e^{\sigma B_s}].
\end{align*}
\]

Notice that \( B_t + B_s \) is normal, with mean 0 and variance \( \text{Var}(B_t) + \text{Var}(B_s) + 2 \text{Cov}(B_t, B_s) = 3s + t \). Thus, we have three more expectations of \( \mathbb{E}[e^X] \) for a normal random variable \( X \), and when the dust settles, we get the answer

\[
\text{Cov}(X_t, X_s) = e^{(\mu + \frac{1}{2} \sigma^2)(t+s)} \left(e^{\sigma^2 s} - 1\right).
\]

If we subtract the mean from \( X_t \), we won’t get a martingale (check this!). We don’t really expect to: \( X_t \) is the exponential of something, and it makes more sense to divide exponentials than to subtract from them. So, we’ll show that \( Y_t = e^{-\mu t - \frac{1}{2} \sigma^2 t} X_t \) is a martingale (with respect to its own filtration, or equivalently, with respect to the filtration of the underlying Brownian motion). Notice that \( Y_t \) always has mean 1.

We compute:

\[
\mathbb{E}(Y_t | \mathcal{F}_s) = \mathbb{E}(e^{-\mu t - \frac{1}{2} \sigma^2 t} e^{\mu t + \sigma B_t} | \mathcal{F}_s) = \mathbb{E}(e^{-\frac{1}{2} \sigma^2 t} e^{\sigma B_s} e^{\sigma(B_t-B_s)} | \mathcal{F}_s)
\]

\[
= e^{-\frac{1}{2} \sigma^2 t} e^{\sigma B_s} \mathbb{E}[e^{\sigma(B_t-B_s)}] = e^{-\frac{1}{2} \sigma^2 t} e^{\sigma B_s} e^{\frac{1}{2} \sigma^2 (t-s)} = e^{-\mu s - \frac{1}{2} \sigma^2 s} e^{\mu s + \sigma B_s} = Y_s.
\]

(Remember that we are trying to get \( e^{\mu s + \sigma B_s} \) on the right-hand side, so that it resembles \( X_t \)!) A different example of a martingale: let \( Z \) be a random variable (for example, “the weather on January 1, 2015”) and let \( \mathcal{F}_s \) be a filtration (for example, “all the information in the weather patterns up to time \( s \)”). Then

\[
X_t = \mathbb{E}(Z | \mathcal{F}_t),
\]

the prediction of the future weather made on day \( t \), is a martingale. This follows directly from the tower law for conditional expectations:

\[
\mathbb{E}(X_t | \mathcal{F}_s) = \mathbb{E}(\mathbb{E}(Z | \mathcal{F}_t) | \mathcal{F}_s) = \mathbb{E}(Z | \mathcal{F}_s), \quad s \leq t.
\]
Another way of saying this is that the best prediction, on day $s$, of what the forecast will be on day $t$ (still for January 1 weather) coincides with the forecast made on day $s$. (This is a good example to keep in mind for the tower law itself!)

Notice that in this case we expect $\lim_{t \to \infty} X_t = Z$, provided we can define the limit appropriately (and provided also that $Z$ is measurable with respect to $\mathcal{F}_\infty = \bigcup_t \mathcal{F}_t$). It’s true for many martingales that $\lim_{t \to \infty} X_t$ exists as a random variable. For example, the limit exists for geometric Brownian motion (can you guess what it should be?). On the other hand, the limit doesn’t exist for Brownian motion: the variance of Brownian motion grows with time, and as $t \to \infty$ Brownian motion diffuses over all of space.

Here’s an easy way of getting more martingales: if $X_t$ and $Y_t$ are martingales with respect to the same filtration $\mathcal{F}$, and $a$ and $b$ are constants, then $aX_t + bY_t$ is a martingale. We’ll see an interesting generalization next time.
Important concepts from today: why is integration with respect to Brownian motion different from ordinary integration; $\langle dB_t \rangle = dt$; mean and variance of the integral; stochastic chain rule.

1. Review of last time

Conditional expectation is a random variable: $\mathbb{E}(X|Y)$ is a random variable that’s $\sigma(Y)$-measurable, and such that the integral of $E(X|Y)$ over a set $A \in \sigma(Y)$ is equal to the integral of $X$ over the same set $A$. The same goes for conditioning on a $\sigma$-algebra: $\mathbb{E}(X|\mathcal{F})$ is a $\mathcal{F}$-measurable random variable such that

$$\int_A \mathbb{E}(X|\mathcal{F}) d\mathbb{P} = \int_A X d\mathbb{P}, \text{ for all sets } A \in \mathcal{F}.$$

When $\sigma(Y)$, or $\mathcal{F}$, is generated by finitely many atoms of positive probability, we can easily compute the conditional expectation, as

$$\mathbb{E}[X|Y = y] = \text{average of } X \text{ over just those elements of } \omega \text{ where } Y(\omega) = y = \frac{\text{weighted sum of those } X(\omega) \text{ for which } Y(\omega) = y}{\mathbb{P}(Y = y)}.$$

This definition isn’t useful when $\mathbb{P}(Y = y) = 0$, but we have lots of properties to help us compute conditional expectation:

1. $\mathbb{E}(X|\mathcal{F})$ is a $\mathcal{F}$-measurable random variable. $\mathbb{E}(X|Y) = \mathbb{E}(X|\sigma(Y))$ is a $\sigma(Y)$-measurable random variable, i.e. a function of $Y$.
2. $\int_A \mathbb{E}(X|\mathcal{F}) d\mathbb{P} = \int_A X d\mathbb{P}$, if $A \in \mathcal{F}$.
3. In particular, if $A = \Omega$, $\mathbb{E}[\mathbb{E}(X|\mathcal{F})] = \mathbb{E}[X]$. This is sometimes called the law of total probability, and it says that to compute the average of $X$ you compute its averages along certain subsets, and then take a weighted average of those values.
4. Conditional expectation is linear: $\mathbb{E}(c_1X_1 + c_2X_2|\mathcal{F}) = c_1\mathbb{E}(X_1|\mathcal{F}) + c_2\mathbb{E}(X_2|\mathcal{F})$.
5. If $X$ is $\mathcal{F}$-measurable, $\mathbb{E}(X|\mathcal{F}) = X$.
6. If $X$ is $\mathcal{F}$-measurable, $\mathbb{E}(XZ|\mathcal{F}) = X\mathbb{E}(Z|\mathcal{F})$.
7. If $X$ and $Y$ are independent, then $\mathbb{E}(X|Y) = \mathbb{E}[X]$ is the constant random variable. If $X$ is independent of the $\sigma$-algebra $\mathcal{F}$, then $\mathbb{E}(X|\mathcal{F}) = \mathbb{E}[X]$.
8. If $X$ is independent of $\mathcal{F}$, then $\mathbb{E}(h(X,Z)|\mathcal{F}) = \mathbb{E}(\mathbb{E}_X[h(X,Z)]|\mathcal{F})$. **This doesn’t require $Z$ to be $\mathcal{F}$-measurable**, I was mistaken last time. Here, $\mathbb{E}_X[h(X,Z)]$ means “average the function over $X$, leaving behind a function of $Z$ alone”. In particular, if $Z$ is $\mathcal{F}$-measurable, the answer will be $\mathbb{E}(h(X,Z)|\mathcal{F}) = \mathbb{E}_X[h(X,Z)]$, the function of $Z$ obtained by taking the average of $h(X,Z)$ over $X$. For example, if $X \sim \mathcal{N}(0,1)$ and $Z$ takes on positive integer values, then $\mathbb{E}_X[X^2]$ is a random variable.

*Date: 10/30/2014.*
variable that takes on the value 0 when $Z$ is odd, and $(2m-1)(2m-3)\ldots 5\cdot 3\cdot 1$ if $Z = 2m$ is even; notice that $X$ isn’t around anymore, just $Z$.

(9) **If $\mathcal{F} \subseteq \mathcal{G}$, then $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{F}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{F}) = \mathbb{E}(X|\mathcal{F})$ (just one conditional, on the smallest amount of information available). This does not work for two generic $\sigma$-algebras, only when one is a subset of the other!**

**Note:** $\mathbb{E}(X|Y)$ is $\sigma(Y)$-measurable (or $\mathbb{E}(X|\mathcal{F})$ is $\mathcal{F}$-measurable), but it doesn’t have to be $\sigma(X)$-measurable! Here’s an example:

| $\omega$ | $X$ | $Y$ | $\mathbb{E}(X|Y)$ |
|----------|-----|-----|------------------|
| 1        | 1   | 0   | 2                |
| 2        | 3   | 0   | 2                |
| 3        | 1   | 1   | 3                |
| 4        | 5   | 1   | 3                |

Notice that $\mathbb{E}(X|Y)$ is constant where $Y$ is constant, not where $X$ is constant. (Although it’s often the case that we condition on a $\sigma(X)$-measurable random variable $Y$, and then the outcome will automatically end up being $\sigma(X)$-measurable.)

We also mentioned that $\mathbb{E}(X|\mathcal{F})$ is the $\mathcal{F}$-measurable random variable which minimizes the expected square deviation $\mathbb{E}([X - Y]^2)$ among all the $\mathcal{F}$-measurable $Y$’s. That is, it’s a best approximation of $X$ by a $\mathcal{F}$-measurable random variable.

We then moved on to define a **filtration**, an increasing collection of $\sigma$-algebras: $\mathcal{F}_s \subseteq \mathcal{F}_t$ for $s \leq t$. Then a stochastic process $X_t$ is **martingale** with respect to a filtration $\mathcal{F}_t$ if three things happen:

1. $X_t$ is $\mathcal{F}_t$-measurable. Equivalently, $\mathbb{E}(X_t|\mathcal{F}_t) = X_t$.
2. $\mathbb{E}(X_s|\mathcal{F}_s) = X_s$ for $s \leq t$. Notice that this is equality of random variables, they should coincide on all $\omega \in \Omega$.
3. $\mathbb{E}(|X_t|) < \infty$ for all $t$. This condition rarely fails, and will never fail in our course. But you should start worrying about it if you see words like “Cauchy” or “power law” or “heavy-tailed” near your random variables.

The examples we’ve seen in last lecture and on the homework are:

1. Brownian motion $B_t$ is a martingale with respect to its natural filtration $\mathcal{F}_t = \sigma(B_s, s \leq t)$. This is the $\sigma$-algebra containing all information about all $B_s$ for all times $s \leq t$ (not just time $t$, but also the past).
2. $B^2_t - t$ is a martingale with respect to the filtration of the Brownian motion, and therefore with respect to its own natural filtration as well.
3. $B^3_t - 3tB_t$ is a martingale with respect to the filtration of the Brownian motion, hence its own natural filtration as well. (This was on the homework.)
4. Geometric brownian motion is $X_t = \exp(\mu t + \sigma B_t)$, and we saw that

$$Y_t = e^{-t/2}X_t$$

is a martingale with respect to the filtration of the Brownian motion (which coincides with its natural filtration). Equivalently, $e^{-\sigma^2/2+\sigma B_t}$ is a martingale. To see this, we needed to compute $\mathbb{E}[e^Z]$ where $Z \sim N(\mu, \sigma)$ is a normal random variable, and we had the identity

$$\mathbb{E}[e^Z] = e^{\mu + \frac{1}{2}\sigma^2}, \quad Z \sim N(\mu, \sigma).$$

5. Taking $\sigma = i$, we have $e^{i/2+itB_t}$ is a martingale with respect to the filtration of Brownian motion, and taking the real part, $e^{i/2}\cos(B_t)$ is a martingale with respect to the filtration of Brownian motion (and hence also with respect to its natural filtration). (This was on the homework.)
(6) If \( Z \) is a random variable and \( F \) is a filtration, then \( X_t = \mathbb{E}(Z|F_t) \) is a martingale with respect to \( F \).

(7) Linear combination of martingales is a martingale.

Let’s put martingales and conditional expectations into the context of the course. In lecture 1, we had the following approach to computing prices of financial derivatives:

1. Divide by powers of \( 1 + r \) to make cash be worth a constant amount at all times.
2. Find \( p \) and \( 1 - p \) that turn the stock prices \( S_t \) into a martingale.
3. All contracts should be martingales with respect to this measure that uses the risk-neutral \( p \) and \( 1 - p \); that will make sure that all portfolios are martingales, ensuring that you can’t make money from nothing (no arbitrage).
4. If a contract takes the form \( \Phi(S_T) \) at time \( T \), its price at time \( t \leq T \) should be \( \mathbb{E}(\Phi(S_T)|F_t) \), where \( F_t = \sigma(S_u, u \leq t) \) takes account of all the stock prices up to time \( t \). (For European options, it turned out that we only needed the stock prices at time \( t \).)

Our goal is to be able to do the same analysis when the stock prices are modeled by \( S_t = \exp(\mu t + \sigma B_t) \). That means that we need to find the change of measure that turns this process into a martingale, and we then need to be able to compute conditional expectations of functions of geometric Brownian motion with respect to this measure. In particular, for European calls the function is still \( \Phi(S_T) = S_T - K \), so to price a European call we need to be able to compute

\[
\mathbb{E}(e^{\mu T + \sigma B_T} - K | F_t), \quad F_t = \sigma(B_s, s \leq t),
\]

with respect to the risk-neutral measure.

To price other financial derivatives, we may have other functions inside the conditional expectation. A lot of the time we will be able to do this by writing a stochastic differential equation that the value of the contract satisfies.

2. Martingale transform

We will start today by defining yet another set of martingales. Let’s switch to discrete time for a moment, and let \( X_n, \ n \geq 0 \) be a martingale with respect to a filtration \( F \) (\( n \) is an integer index). Let \( C_n \) be a process (i.e. sequence of random variables) that’s predictable: \( C_n \) is \( F_{n-1} \)-measurable. That is, you know the value of \( C_n \) on day \( n - 1 \). Then the process given by

\[
Y_0 = 0, \quad Y_n - Y_{n-1} = C_n(X_n - X_{n-1})
\]

is a \( F \)-martingale. It’s called the martingale transform of \( X \) by \( C \). (We need an additional assumption to control expectations: for example, \( \mathbb{E}[C_n^2] < \infty \) and \( \mathbb{E}[X_n^2] < \infty \) for all \( n \).) Let’s check this.

First, \( Y_n \) is \( F_n \)-measurable, because if you know \( F_n \) then you know \( X_0, \ldots, X_n \), and therefore you can compute \( Y_0, \ldots, Y_n \). Second,

\[
\mathbb{E}(Y_n - Y_k | F_k) = \mathbb{E}((Y_n - (Y_{n-1})) + (Y_{n-1} - Y_{n-2}) + \ldots + (Y_{k+1} - Y_k)) | F_k)
\]

\[
= \mathbb{E}(C_n(X_n - X_{n-1}) | F_k) + \ldots + \mathbb{E}(C_{k+1}(X_{k+1} - X_k) | F_k)
\]

\[
= \mathbb{E}(\mathbb{E}(C_n(X_n - X_{n-1}) | F_{n-1}) | F_k) + \ldots + C_{k+1} \mathbb{E}(X_{k+1} - X_k | F_k)
\]

\[
= \mathbb{E}(C_n \mathbb{E}(X_n - X_{n-1} | F_{n-1}) | F_k) + \ldots + C_{k+1} \mathbb{E}(X_{k+1} - X_k | F_k)
\]

\[
= \mathbb{E}(0 | F_k) + \ldots + 0 = 0.
\]
3. **Itô integral: introduction**

Our goal for the rest of today is to make sense of expressions of the form

\[ X_t = \int_0^t Y_s dB_s, \quad B \text{ is a Brownian motion.} \]

A particular example will be for \( Y_s = B_s \), i.e. \( \int_0^t B_s dB_s \). We will find that there are good reasons why we should have

\[ \int_0^t B_s dB_s \neq \frac{1}{2} B_t^2, \quad \int_0^t B_s dB_s = \frac{1}{2} B_t^2 - \frac{1}{2} t. \]

Why is there a problem at all? If \( B \) had differentiable sample paths, we would write

\[ \int_0^t Y_s dB_s = \int_0^t Y_s \frac{dB}{ds} ds. \]

Unfortunately, the derivative in there is not defined. You can get around this for sufficiently nice functions \( Y_s \), but not for \( Y_s = B_s \).

Let’s think about the integral as a Riemann sum. Partition of the interval \([0, t]\):

\[ 0 = t_0 < t_1 < t_2 \ldots < t_{n-1} < t_n = t. \]

Consider

\[ S_n = \sum_{i=1}^n B_{t_{i-1}} \Delta_i B, \quad \Delta_i B = B_{t_i} - B_{t_{i-1}}. \]

One could also consider \( B_{t_i} \Delta_i B \), but using \( B_{t_{i-1}} \) is better, because then this is a martingale transform.

Because \( S_n \) is an approximation for \( \int_0^t B_s dB_s \), we expect \( \frac{1}{2} B_t^2 \) to be part of the result. Rewrite \( S_n \) in a way that \( \frac{1}{2} B_t^2 \) makes an appearance:

\[ S_n = B_0(B_{t_1} - B_0) + B_{t_1}(B_{t_2} - B_{t_1}) + \ldots + B_{t_{n-1}}(B_t - B_{t_{n-1}}) \]
\[ = \left( \frac{1}{2} B_{t_1}^2 - \frac{1}{2} B_0^2 - \frac{1}{2} (B_{t_1} - B_0)^2 \right) + \left( \frac{1}{2} B_{t_2}^2 - \frac{1}{2} B_{t_1}^2 - \frac{1}{2} (B_{t_2} - B_{t_1})^2 \right) \]
\[ + \ldots + \left( \frac{1}{2} B_t^2 - \frac{1}{2} B_{t_{n-1}}^2 - \frac{1}{2} (B_t - B_{t_{n-1}})^2 \right) \]
\[ = \frac{1}{2} B_t^2 - \frac{1}{2} B_0^2 - \frac{1}{2} \sum_{i=1}^n (\Delta_i B)^2. \]

In Riemann integration, we would now argue that \( Q_n \to 0 \) as \( n \to \infty \) and the partition becomes finer. For Brownian motion, \( Q_n \not\to 0 \). Instead, we have

\[ \mathbb{E} Q_n = \sum_{i=1}^n \mathbb{E}[(B_{t_i} - B_{t_{i-1}})^2] = \sum_{i=1}^n (t_i - t_{i-1}) = t, \]
and also, because increments of Brownian motion are independent,

\[
\text{Var} Q_n = \sum_{i=1}^{n} \text{Var}((B_{t_i} - B_{t_{i-1}})^2) = \sum_{i=1}^{n} \mathbb{E}[\Delta_i(B)]^4 - \sum_{i=1}^{n} (\mathbb{E}[\Delta_i(B)^2])^2
\]

\[
= \sum_{i=1}^{n} \frac{(t_i - t_{i-1})^4 \mathbb{E}[N(0, 1)^4]}{3} - \sum_{i=1}^{n} (t_i - t_{i-1})^2
\]

\[
= 2 \sum_{i=1}^{n} (t_i - t_{i-1})^2.
\]

As the partition becomes finer,

\[
\text{Var} Q_n \leq \left( \sum_{i=1}^{n} (t_i - t_{i-1}) \right) \max_i (t_i - t_{i-1}) \to 0, \quad \text{as} \quad \max_i (t_i - t_{i-1}) \to 0.
\]

So \( Q_n \to t \), and

\[
S_n \to \int_{0}^{t} B_s dB_s = \frac{1}{2} B_t^2 - \frac{1}{2} t.
\]

Notice that

\[
Q_n = \sum_{i=1}^{n} (\Delta_i B)^2 \quad \text{should converge to} \quad \int_{0}^{t} (dB_s)^2 = t.
\]

This gives us the very good heuristic:

\[
(dB_t)^2 = dt, \quad (dt)^2 = 0.
\]

(The second one comes from ordinary calculus.)

Using this heuristic and Taylor expansion, we can write

\[
B_{t+dt}^2 \approx B_t^2 + 2B_t \quad dB_t + \quad \frac{1}{2} \quad \frac{2}{n} \quad \text{second derivative} \quad \cdot \quad (dB_t)^2
\]

\[
\approx B_t^2 + 2B_t dB_t + dt,
\]

which means

\[
\frac{d(B_t^2)}{dt} = 2B_t dB_t + dt, \quad B_t^2 = \int_{0}^{t} 2B_s dB_s + \int_{0}^{t} ds .
\]

4. \textsc{Itô Integral: Definition}

We can define more general integrals \( \int f(s) dB_s \) for \textit{predictable} functions \( f \), meaning that \( f(t) \) is measurable with respect to \( \sigma(B_u, u \leq t) \). The idea is as follows:

1. Approximate the integral by Riemann sums. If you think about it, this is the same as defining the integral exactly for predictable \textit{piecewise-constant} functions.
2. Show that as the size of the partition shrinks, the integral converges to some random variable (in the sense that the variance of the difference converges to 0).
   This random variable is then the value of the integral.

The integrand \( f(s) \) needs to satisfy two conditions:
1. $f$ is adapted to the filtration of Brownian motion, or equivalently, $f(t) \in \sigma(B_u, u \leq t)$, for all $t$. This happens automatically if $f$ is either non-random, or a function of $t$ and $B_t$ (but we also allow functions of the entire trajectory).

2. We need $\int_0^t E[f(s)^2]ds < \infty$.

Properties of Itô integral $\int_0^t f(s)dB_s$:

1. The integral is a martingale.
2. Integral is still linear: $\int_a^b (f(s) + g(s))dB_s = \int_a^b f(s)dB_s + \int_a^b g(s)dB_s$
3. Also linear in the domain: $\int_a^b f(s)dB_s + \int_a^c f(s)dB_s = \int_a^d f(s)dB_s$
4. The Itô isometry:
$$E[(\int_0^t f(s)dB_s)^2] = \int_0^t (E[f(s)^2])dt.$$ Notice that on the right we have an ordinary Riemann integral of a non-random function, and we’ve assumed it’s finite.
5. The integral is continuous, i.e. has continuous sample paths.

5. ITÔ LEMMA: CHAIN RULE OF STOCHASTIC CALCULUS

Recall the classical chain rule for ordinary functions:
$$f(g(x))' = f'(g(x)) \cdot g'(x).$$ For example, for an ordinary function $b(t)$, we have $\frac{d}{dt}b^2(t) = 2b(t)\frac{db(t)}{dt}$, or $d(b^2(t)) = 2b(t)db(t)$. We saw that this doesn’t hold for Brownian motion: there we had $d(B^2_t) = 2B_tdB_t + dt$. Let’s try to understand why this is happening.

In integral form, ordinary chain rule says
$$(\text{chain rule}) \quad f(g(b)) - f(g(a)) = \int_a^b f'(g(t))g'(t)dt = \int_a^b f'(g(t))dg(t).$$

An intermediate step between the two rules is the first-order Taylor expansion:
$$f(g(t + dt)) = f(g(t)) + g'(t)dt + O(dt^2) \quad \text{in ordinary calculus}$$
$$= \left( f(g(t)) + g'(t)dy + O(\|dy\|^2) \right) \approx f(g(t)) + g'(g(t))g'(t)dt \quad \text{in ordinary calculus}.$$ 

When we’re working with functions of Brownian motion, we cannot just do the first-order Taylor expansion; we must take the second-order term into account as well. So:
$$f(B_{t+dt}) = f(B_t + dB_t) \approx f(B_t) + f'(B_t)dB_t + \frac{1}{2} f''(B_t)(dB_t)^2$$
$$= f(B_t) + f'(B_t)dB_t + \frac{1}{2} f''(B_t)dt.$$

Lemma 5.1 (Itô lemma, or stochastic chain rule).
$$\boxed{df(B_t) = f'(B_t)dB_t + \frac{1}{2} f''(B_t)dt, \quad \text{differential form}}$$
$$\boxed{f(B_b) - f(B_a) = \int_a^b f'(B_t)dB_t + \frac{1}{2} \int_a^b f''(B_t)dt, \quad \text{integral form}}.$$ 

Notice that $\int_a^b f''(B_t)dt$ is the ordinary Riemann integral, but possibly of a random function. We will see how to generalize the Itô lemma to more complicated functions, particularly functions of several variables.
Important concepts from today: Itô lemma for more involved functions, $dBdt = 0$, Black-Scholes pricing of a European call by a lot of differentiation, Black-Scholes pricing of a European call by a risk-neutral measure.

1. Review of last time

We defined the martingale transform: if $X_n$ is a martingale with respect to filtration $\mathcal{F}$, and $C_n$ is predictable, meaning $C_n$ is $\mathcal{F}_{n-1}$-measurable, then

$$Y_n = \sum_{k=1}^{n} C_k (X_k - X_{k-1})$$

is a martingale also adapted to $\mathcal{F}$. (We need an additional condition, e.g. $\mathbb{E}[X_n^2] < \infty, \mathbb{E}[C_n^2] < \infty$; we won’t worry about it.)

Taking this concept into continuout time, we defined the stochastic integral,

$$\int_{0}^{t} f(s) dB_s,$$

where $f$ is predictable, that is, measurable with respect to $\sigma(B_u, u \leq t)$. We showed that this is different from the ordinary Riemann integral; for example,

$$\int_{0}^{t} B_s dB_s = \frac{1}{2} B_t^2 - \frac{1}{2} t,$$

and not $\frac{1}{2} B_t^2$.

In addition to predictability, we need a condition of the form

$$\int_{0}^{t} \mathbb{E}[f(s)^2] ds < \infty,$$

but we won’t ever run into this not being the case.

We showed some properties of the stochastic integral:

1. The stochastic integral $X_t = \int_{0}^{t} f(s) dB_s$ is a martingale. In particular, $\mathbb{E}[X_t] = \mathbb{E}[X_0] = 0$. A good way to see this is that $\mathbb{E}[\int_{0}^{t} f(s) dB_s] = \int_{0}^{t} \mathbb{E}[f(s) dB_s] = \int_{0}^{t} \mathbb{E}[f(s)] \mathbb{E}[dB_s]$ = 0, since the Brownian motion increments have mean zero.

2. Integral is still linear: $\int_{a}^{b} (f(s) + g(s)) dB_s = \int_{a}^{b} f(s) dB_s + \int_{a}^{b} g(s) dB_s$

3. Also linear in the domain: $\int_{a}^{b} f(s) dB_s + \int_{a}^{c} f(s) dB_s = \int_{a}^{c} f(s) dB_s$

4. The Itô isometry lets us compute the variance of the integral:

$$\text{Var} \left( \int_{0}^{t} f(s) dB_s \right) = \mathbb{E} \left( \int_{0}^{t} f(s) dB_s \right)^2 = \int_{0}^{t} (\mathbb{E}[f(s)^2]) ds.$$
Notice that on the right we have an ordinary Riemann integral of a non-random function, and we’ve assumed it’s finite.

(5) The integral is continuous, i.e. has continuous sample paths.

If the integrand is an ordinary, non-random function, then
\[ \int_0^t f(s) dB_s \sim N(0, \int_0^t f(s)^2 ds) \]
will be a normal random variable, because it’s a linear transformation of the increments of Brownian motion. This doesn’t have to be true if the integrand is random, for example
\[ \int_0^t B_s dB_s = \frac{1}{2} B_t^2 - \frac{1}{2} t \]
is definitely not normal because it can’t be smaller than \(-t/2\).

We derived the fundamental relationships
\[ dt^2 = 0, \quad dB_t^2 = dt \]
and used them to show that for stochastic calculus,
\[ d(f(B_t)) = f'(B_t) dB_t + \frac{1}{2} f''(t) dt, \quad f(B_t) - f(B_0) = \int_0^t f'(B_s) dB_s + \frac{1}{2} \int_0^t f''(B_s) ds. \]
(The first of these terms is what you’d expect from ordinary chain rule, the second is particular to stochastic calculus.)

2. More general Itô rules

We begin by deriving more rules of differentiation. So far we know how to differentiate \( f(B_t) = B_t^2 \), but not \( f(t, B_t) = tB_t \). Let’s fix this:
\[ dBdt = 0 \]

Now if we need to work with functions of many variables, we just do the Taylor expansion in all variables, until terms start vanishing:
\[ f(t + dt, B + dB_t) = f(t, B_t) + dt \frac{\partial}{\partial t} f(t, B_t) + dB_t \frac{\partial}{\partial B_t} f(t, B_t) \]
\[ + \frac{1}{2} \sum_{j=0}^1 \frac{\partial^2}{\partial t^j \partial B^{2-j}} f(t, B_t) + \frac{1}{2} \sum_{j=0}^\infty dBdBt \frac{\partial^j}{\partial B^j \partial dt} f(t, B_t) + \ldots \]
\[ \approx f(t, B_t) + \partial_B f(t, B_t) dB + \left( \partial_t f(t, B_t) \frac{1}{2} \partial_{BB} f(t, B_t) \right) dt. \]

Here, \( \partial \) refers to partial differentiation, and \( \partial_{BB} \) means “take the second partial derivative with respect to \( B \)”. That is,
\[ df(t, B_t) = \partial_B f(t, B_t) dB + \left( \partial_t f(t, B_t) \frac{1}{2} \partial_{BB} f(t, B_t) \right) dt, \]
or
\[ f(t, B_t) - f(0, B_0) = \int_0^t \partial_B f(s, B_s) ds + \int_0^t \left( \partial_s f(s, B_s) + \frac{1}{2} \partial_{BB} f(s, B_s) \right) ds. \]

Remark 1. When you’re taking partial derivatives, you get to pretend that variables don’t depend on each other. That is, \( f \) is just a function of two variables that have nothing to do with each other, and you’re taking the partial derivative of \( f \) with respect to its first input or its second input. You just happen to be evaluating it at a point where \( B_t \) depends on \( t \).
For example, if we take \( f(t, B_t) = tB_t \), then
\[
df(t, B_t) = t dB_t + B_t dt + 0, \quad tB_t = \int_0^t s dB_s + \int_0^t B_s ds.
\]

**Example 1.** Show that \( B_t^3 - 3tB_t \) is a martingale.

**Solution:** Differentiate:
\[
d(B_t^3 - 3tB_t) = 3B_t^2 dB_t + 3B_t dt - 3t dB_t - 3B_t dt = 3(B_t^2 - t) dB_t,
\]
or
\[
B_t^3 - 3tB_t = 3 \int_0^t (B_s^2 - s) dB_s.
\]

Since stochastic integrals are martingales, we see that \( B_t^3 - 3tB_t \) is a martingale.

**Example 2.** Determine whether \( f(t, B_t) = e^{-t/2+B_t} \) is a martingale.

**Solution:** Differentiate:
\[
df(t, B_t) = \frac{1}{2} e^{-t/2+B_t} \frac{\partial}{\partial t} dt + e^{-t/2+B_t} dB_t + \frac{1}{2} e^{-t/2+B_t} \frac{\partial}{\partial B} dt = e^{-t/2+B_t} dB_t.
\]

Consequently, \( f(t, B_t) = f(0, B_0) + \int_0^t e^{-s/2+B_s} dB_s \) is a martingale.

For the particular function \( f(t, B_t) = e^{-t/2+B_t} \) that we just looked at,
\[
f(t, B_t) - f(s, B_s) + \int_s^t f(u, B_u) dB_u, \quad df(t, B_t) = df(t, B_t).
\]

Ordinarily, this is the equation satisfied by the exponential function: \( f'(x) = f(x) \) means \( f(x) = Ce^x \). In the stochastic world, the solution to the stochastic differential equation \( df(t, B_t) = f(t, B_t) \) is \( f(t, B_t) = Ce^{-t/2+B_t} \). This is sometimes called the Itô exponential.

**Example 3** (Geometric Brownian motion). Let \( X_t = f(t, B_t) = \exp((c - \frac{1}{2}\sigma^2)t + \sigma B_t) \).

Then:
\[
dX_t = \sigma X_t dB_t + (c - \frac{1}{2}\sigma^2)X_t dt + \frac{1}{2}\sigma^2 X_t dt = \sigma X_t dB_t + cX_t dt.
\]

That is, \( S \) satisfies the stochastic differential equation
\[
X_t - X_0 = c \int_0^t X_s ds + \sigma \int_0^t X_s dB_s.
\]

We have just seen a couple of stochastic differential equations, or SDEs (one satisfied by the Itô exponential, one satisfied by more general geometric Brownian motions). We’ll discuss SDEs more next time, but for now, let’s see how we can deal with solutions to them.

Suppose
\[
dX_t = A_1(t) dt + A_2(t) dB_t, \quad \text{or equivalently,} \quad X_t = X_0 + \int_0^t A_1(s) ds + \int_0^t A_2(s) dB_s.
\]

Consider a function \( f(t, X_t) \). The rule is the same as always: take the second-order Taylor expansion, but only keep second-order terms if \( dB^2 \) appears in them (because
\[ dt^2 = dBdt = 0: \]
\[
df(X_t) = \partial_t f(t, X_t)dt + \partial_X f(t, X_t)dX_t + \frac{1}{2} \partial_{XX} f(t, X_t) (dX_t)^2 = A^2_t dB_t^2 = A'_t dt
\]
\[
= \left( \partial_t f(t, X_t) + \partial_X f(t, X_t)A_1 + \frac{1}{2} \partial_{XX} f(t, X_t)A_2^2 \right) dt + (\partial_X f(t, X_t)A_2(t)) dB_t.
\]
If we wanted to write this in integral form, we would write
\[
X_t - X_0 = \int_0^t A_1(s)ds + \int_0^t A_2(s)dB_s \implies
f(X_t) - f(X_0) = \int_0^t \left( \partial_s f(s, X_s) + \partial_X f(s, X_s)A_1(s) + \frac{1}{2} \partial_{XX} f(s, X_s)A_2(s)^2 \right) ds
+ \int_0^t \partial_X f(s, X_s)A_2(s)dB_s.
\]
Notice that while the \(dt\)-term gets quite complicated, the \(dB\) term changes in a very simple way. This is sometimes handy.

Finally, we may want to differentiate products of two (or more) solutions of SDEs, or other functions of multiple solutions of SDEs (or just a function of a solution of an SDE and a Brownian motion). Suppose
\[
dX_t = A_1(t)dt + A_2(t)dB_t, \quad dY_t = A_3(t)dt + A_4(t)dB_t.
\]
Consider \(f(t, X_t, Y_t)\): we again do a second-order Taylor expansion, and then simplify terms, using the equations \(dt^2 = 0, dBdt = 0, dB^2 = dt\):
\[
df(t, X_t, Y_t) = \partial_t(f)dt + \partial_X(f)dX_t + \partial_Y(f)dY_t
+ \frac{1}{2} \partial_{XX}(f)(dX_t)^2 + \frac{1}{2} \partial_{YY}(f)(dY_t)^2 + \underbrace{\partial_{XY}(f)}_{\text{this is } \frac{1}{2}(\partial_{XY} + \partial_{YX})} dX_t dY_t
= \partial_t(f)dt + \partial_X(f) \underbrace{(A_1(t)dt + A_2(t)dB_t)}_{dX_t} + \partial_Y(f) \underbrace{(A_3(t)dt + A_4(t)dB_t)}_{dY_t}
+ \frac{1}{2} \partial_{XX}(f) \underbrace{(A_2(t)^2dt)}_{dX_t^2} + \frac{1}{2} \partial_{YY}(f) \underbrace{(A_4(t)^2dt)}_{dY_t^2}
+ \underbrace{\partial_{XY}(f)}_{\frac{1}{2}(\partial_{XY} + \partial_{YX})} \underbrace{(A_2(t)A_4(t)dt)}_{dX_t dY_t}
\]
\[
= \left( \partial_t(f) + \partial_X(f)A_1(t) + \partial_Y(f)A_3(t) + \frac{1}{2} (\partial_{XX}(f)A_2(t)^2 + 2\partial_{XY}(f)A_2(t)A_4(t) + \partial_{YY}(f)A_4(t)^2) \right) dt
+ \partial_X(f)A_2(t)dB_t + \partial_Y(f)A_4(t)dB_t.
\]

**Example 4.** Consider the product of two such stochastic processes:
\[
d(X_t Y_t) = Y_t dX_t + X_t dY_t + \underbrace{dX_t dY_t}_{A_2(t)A_4(t)dt}
= (Y_t A_1(t) + X_t A_3(t) + A_2(t)A_4(t)) dt + (X_t A_2(t) + Y_t A_4(t)) dB_t
\]
The \(dX_t dY_t\) the stochastic calculus correction to the classical product rule.

**Example 5 (A case of integration by parts).** Consider \(X_t = f(t)\) a function of time only, and \(Y_t = B_t\). Then
\[
d(f(t)B_t) = B_t f'(t)dt + f(t)dB_t + 0,
\]
so \( f(t)B_t - f(0)B_0 = \int_0^t B_s f'(s)ds + \int_0^t f(s)dB_s \). Rearranging, we get the integration by parts formula:

\[
\int_0^t f(s)dB_s = f(t)B_t - \int_0^t B(s)f'(s)ds
\]

This looks just like the ordinary integration by parts, because \( f \) was an ordinary differentiable function of time (non-random). If \( f \) were random, we would have an extra correction term, \( \int_0^t \partial_B f(s, B_s)ds \).

3. BLACK-SCHOLES OPTION PRICING

We have a market with a stock and a bond, and we assume that the price of a stock is given by a geometric Brownian motion:

\[
S_t = f(t, B_t) = X_0e^{(c-\frac{1}{2}\sigma^2)t+\sigma B_t},
\]

satisfying the stochastic differential equation

\[
dS_t = cS_tdt + \sigma S_t dB_t, \quad S_t = S_0 + c \int_0^t S_s ds + \sigma \int_0^t S_s dB_s.
\]

Here, \( \sigma \) is the volatility or the stock, and \( c \) is the mean rate of return.

We also have a non-risky asset on the market, with price \( \beta_t = \beta_0e^{rt} \). Here, \( r \) is the interest rate, and interest is being compounded continuously. Of course, \( \beta \) satisfies the deterministic equation

\[
\dot{\beta}_t = \beta_0 + r \int_0^t \beta_s ds.
\]

Consider a portfolio with \( a_t \) shares of stock, and \( b_t \) shares of the bond. The value of the portfolio at time \( t \) is

\[
V_t = a_t S_t + b_t \beta_t.
\]

(This is, of course, a random quantity.) If \( a_t > 0 \), this means you’ve bought stock; if \( a_t < 0 \), you’re short-selling stock (which means you’ll need to buy it later). Similarly, if \( b_t < 0 \) then you’re borrowing money from the bank (and will need to repay it later). We’ll assume that \( a_t \) and \( b_t \) are unbounded, and that you aren’t taking money out of the portfolio or adding it in. (And also that there is no cost to transactions: buying and selling is free.)

The last condition is called self-financing, and it should mean that

\[
dV_t = a_t dS_t + b_t dB_t = a_t(cS_t dt + \sigma S_t dB_t) + b_t(r \beta_t dt)
\]

changes in the value of the portfolio come only from the changes in the prices of the underlying stocks and bonds.

Let’s think about this in discrete time, where you decide on \( a_{n+1}, b_{n+1} \), then the changes happen to the price, and then you can reallocate your wealth:

\[
V_{n+1} = a_{n+1}S_{n+1} + b_{n+1} \beta_{n+1} = a_{n+2}S_{n+1} + b_{n+2} \beta_{n+1}
\]

reallocating existing wealth into new combination of stock and bond

\[
V_n = a_{n+1}S_n + b_{n+1} \beta_n
\]

\[
\delta V = a_{n+1} \delta S + b_{n+1} \delta \beta
\]

Now consider an option, such as a European call, which at time \( T \) will be paying \( (S_T - K)_+ = \max(S_T - K, 0) \). Let’s try to find a replicating portfolio for this option: that is, there are some functions \( a_t, b_t \) that satisfy the conditions of a self-financing portfolio, and we need the final value to be equal to \( (S_T - K)_+ \).
Suppose that the value of this replicating portfolio is given by

\[ V_t = u(T - t, S_t), \quad u \text{ is a smooth deterministic function.} \]

That is, we assume that all we need to know to price the call option is the time left to exercise it, and the current stock price (not, say, the history of how the stock has been doing to date). The final value gives as the initial condition on \( u \):

\[ u(0, S_T) = (S_T - K)_+. \]

We’ll now derive a partial differential equation that \( u \) satisfies, by applying Itô lemma to it and comparing to the definition of a self-financing portfolio. The goal is to have \( a_t \) and \( b_t \) fall out, and have an equation that involves just the value \( u \), the time \( t \), and the stock price \( S \).

Apply Itô lemma to \( V \):

\[
dV_t = -\partial_t u(T - t, S_t)dt + \partial_S u(T - t, S_t) dS_t + \frac{1}{2} \partial_{SS} u(T - t, S_t) (dS_t)^2
\]

\[ = -\partial_t u(T - t, S_t) dt + \partial_S u(T - t, S_t) (cS_t dt + \sigma S_t dB_t) + \frac{1}{2} \partial_{SS} u(T - t, S_t) (\sigma^2 S_t^2 dt)
\]

\[ = \left( -\partial_t u(T - t, S_t) + cS_t \partial_S u(T - t, S_t) + \frac{1}{2} \sigma^2 S_t^2 \partial_{SS} u(T - t, S_t) \right) dt
\]

\[ + (\sigma S_t \partial_S u(T - t, S_t)) dB_t.\]

Here’s a fact:

If \( dV_t \) has a representation as an SDE, \( dV_t = A_1(t) dt + A_2(t) dB_t \), then the representation is unique.

In particular, we now have a system of two partial differential equations

\[
\begin{align*}
-\partial_t u(T - t, S_t) + cS_t \partial_S u(T - t, S_t) + \frac{1}{2} \sigma^2 S_t^2 \partial_{SS} u(T - t, S_t) &= ca_t S_t + rb_t \beta_t, & \text{dt terms} \\
\sigma S_t \partial_S u(T - t, S_t) &= \sigma a_t S_t, & \text{dB terms}
\end{align*}
\]

From the \( dB \) term, we see \( a_t = \partial_S u(T - t, S_t) \). Substituting this into the \( dt \) term,

\[
-\partial_t u(T - t, S) + ca_T S_T + \frac{1}{2} \sigma^2 S_T^2 \partial_{SS} u(T - t, S_T) = ca_T S_T + rb_T \beta_T = r(u(T - t, S_t) - a_t S_t)
\]

recall that \( u(T_t, S_t) = a_t S_t + b_t \beta_t \)

Equivalently, for all \( t > 0 \) and \( s > 0 \),

\[
\partial_t u(t, s) = \frac{1}{2} \sigma^2 s^2 \partial_{ss} u(t, s) + rs \partial_s u(t, s) - ru(t, s), \quad u(0, s) = (s - K)_+
\]

this is the \( ra_t S_t \) term from last line

We can go from random stock price \( S_t \) to just a random variable \( s \) because the PDE is satisfied at all possible values of the stock price, i.e. at all positive numbers.

That is: we started from the SDE that the value of the portfolio needed to satisfy, assumed that it was parametrized as \( u(T - t, S_t) \), and arrived at the PDE satisfied by \( u \). We haven’t used the initial condition yet. The PDE alone (without any initial conditions) is known as the Black-Scholes PDE, and it has to be satisfied by any price of a portfolio or contract in the Black-Scholes market. The initial condition is what determines the particular solution or contract.
In the case of \( u(0, s) = (s - K)_+ \), the combination of PDE and initial conditions has an explicit solution:

\[
u(t, s) = s \Phi \left( \frac{\ln(s/K) + (r + \frac{1}{2} \sigma^2 t)}{\sigma \sqrt{t}} \right) - Ke^{-rt} \Phi \left( \frac{\ln(s/K) + (r + \frac{1}{2} \sigma^2 t)}{\sigma \sqrt{t}} - \sigma \sqrt{t} \right),\]

where \( \Phi \) is the normal CDF. Make sure you can check that this function \( u(t, s) \) satisfies the PDE! We’ll see a way to derive this answer when we think about that calculation from a martingale point of view.

The replicating portfolio is given by

\[
a_t = \partial_S u(T - t, S_t), \quad b_t = \frac{u(T - t, S_t) - a_t S_t}{\beta_t}.
\]

If we hadn’t managed to solve the final PDE explicitly, we could still compute the solution to it numerically on the computer, by picking a grid of time increments and prices. This is what you end up doing if you can’t get a neat answer for a particular contract.

4. A change-of-measure approach to Black-Scholes

Our approach in the previous section to assume that there is a replicating portfolio for a European call, which was based only on the current stock price and the time remaining until execution. (Note that in real life we may well have more information, like the weather forecast for tomorrow and the prices of all the other stocks in the market. I hope so say something about multidimensional models later in the course.) We then used the SDEs satisfied by \( S_t \) and \( \beta_t \) to write down a non-random partial differential equation satisfied by the value \( u \), and we pulled a solution of that PDE out of a hat.

As we saw in the first lecture, there’s a different approach to arbitrage-free pricing in the market, and it’s based on the following observation:

1. Start by normalizing all prices so that the price of bonds is constant, i.e. a martingale.
2. Change measures to make the discounted price of stocks into a martingale. If we look at the answer we got by our first approach, we see that the mean rate of return \( c \) didn’t come into it at all: option price does not care about the mean rate of return, only about volatility. So an intelligent guess would be that we can make the discounted price of stocks look like \( \exp(-\sigma^2 t/2 + \sigma \tilde{B}_t) \), the geometric Brownian motion that’s a martingale.
3. In the world where bonds and stocks are martingales, our financial derivative should be a martingale as well: then any portfolio will be a martingale, and therefore its expected value will be the same always. This implies that if you have a chance of profiting, you also have a chance of losing, and that’s exactly what “no arbitrage” is about.

So suppose that

\[
e^{-rt} S_t = e^{-\frac{\sigma^2}{2} t + \sigma \tilde{B}_t}
\]

for some Brownian motion \( \tilde{B}_t \). To make the discounted European call into a martingale, we will define

\[
e^{-rt} V_t = \mathbb{E} \left( e^{-rT} (S_T - K)_+ | \mathcal{F}_t \right),
\]

where \( \mathcal{F}_t = \sigma(S_u, u \leq t) \) gives the information available from the stock prices up to time \( t \). We saw that conditioning a random variable on a filtration defines a martingale, which
tells us that this is one right thing to do. (It turns out that it’s the only right thing to do, although we won’t show this.)

Move the exponential over to the other side, and recall that **discounted** stock price follows a geometric Brownian motion:

\[ e^{-rT} S_T = e^{-rt} S_t e^{-\sigma^2/2(T-t)} + \sigma (\tilde{B}_T - \tilde{B}_t) \]

Use that to split up the stock price into a part that’s independent of \( F_t \) and a part that’s \( F_t \)-measurable:

\[ V_t = e^{-r(T-t)} \mathbb{E} \left( \underbrace{S_t e^{-(r-\sigma^2/2)(T-t)+\sigma (\tilde{B}_t - \tilde{B}_t)}}_{S_T} - K \right)_{+} \mid F_t \]

The point here is, as usual, that \( S_t \) is \( F_t \)-measurable, whereas \( \sigma (\tilde{B}_T - \tilde{B}_t) \) is a normal random variable with mean 0, variance \( \sigma^2 (T-t) \), and independent of \( F_t \).

Recall the rule that said “if you have something independent inside a conditional expectation, you may take the average over it first”. When we do that to \( (\tilde{B}_T - \tilde{B}_t) \), we’ll get something that is \( F_t \)-measurable. This means that

\[ V(t, S_t) = e^{-r(T-t)} \mathbb{E}_{N(0,\sigma^2(T-t))} \left[ (S_t e^{-(r-\sigma^2/2)(T-t)} e^{N(0,\sigma^2(T-t))} - K)_{+} \right]. \]

Rewriting as an integral (notice that we’re integrating over \( u \) below, which means that we treat \( S_t \) as a constant!)

\[ V(t, X_t) = e^{-r(T-t)} \frac{1}{\sqrt{2\pi} \sigma \sqrt{T-t}} \int_{-\infty}^{\infty} \left( S_t e^{-(r-\sigma^2/2)(T-t)} e^u - K \right) + e^{-\frac{1}{2} \sigma^2 u^2} du. \]

It’s now clear why the answer should be a complicated expression involving the normal CDF \( \Phi \): you change variables to get something that’s \( N(0,1) \), then you figure out the bounds of integration from the fact that you’re taking the positive part.

This approach is simpler because we didn’t have to solve a PDE. However, what we haven’t explained yet is why you can change measure (in a way that doesn’t change the filtration) and transform a generic geometric Brownian motion into an Itô integral (which is a martingale). This is the content of Girsanov’s theorem, and we’ll talk about it next time.
Important concepts from today: solving linear ODEs, solving linear SDEs, converting linear SDE into ODEs satisfied by the mean and variance of the solution.

1. Review of last time

Summary of Itô calculus in one dimension:

\[ dt^2 = 0, \quad dB dt = 0, \quad dB^2 = dt \]

If you want to find the differential of anything, take the second-order Taylor expansion and leave only the non-zero terms from the above equation. For example:

1. If \( f(B_t) \) is a function of only Brownian motion, then

\[
\begin{align*}
    d(f(B_t)) &= f'(B_t)dB_t + \frac{1}{2}f''(B_t)dt, \\
    f(B_t) - f(B_0) &= \int_0^t f'(B_s)dB_s + \int_0^t f''(B_s)ds
\end{align*}
\]

2. If \( f(t, B_t) \) is a function of time and Brownian motion, then

\[
\begin{align*}
    d(f(t, B_t)) &= \partial_B f(t, B_t)dB_t + \partial_t f(t, B_t)dt + \frac{1}{2}\partial_{BB} f(t, B_t)dt \\
    f(t, B_t) - f(0, B_0) &= \int_0^t f(s, B_s)dB_s + \int_0^t \left( \partial_s f(s, B_s) + \frac{1}{2}\partial_{BB} f(s, B_s) \right)ds
\end{align*}
\]

Remember that in taking partial derivatives you get to ignore the fact that \( B_t \) depends on \( t \), and just treat them as two unrelated variables. You just happen to be evaluating the partial at a point where the two variables depend on one another.

3. If \( X_t \) satisfies an SDE, \( dX_t = A_1(t)dt + A_2(t)dB_t \), where \( A_1(t) \) and \( A_2(t) \) are possibly random functions adapted to the filtration of Brownian motion, then

\[
\begin{align*}
    d(f(t, X_t)) &= \partial_t f(t, X_t)dt + \partial_X f(t, X_t)dX_t + \frac{1}{2}\partial_{XX} f(t, X_t)(dX_t)^2 \\
    &= \partial_t f(t, X_t)dt + \partial_X f(t, X_t)\left( A_1(t)dt + A_2(t)dB_t \right) + \frac{1}{2}\partial_{XX} f(t, X_t)A_2(t)^2dt
\end{align*}
\]

and you can write down an integral formulation as well.

4. If you have two solutions of SDEs, \( dX_t = A_1(t)dt + A_2(t)dB_t \) and \( dY_t = A_3(t)dt + A_4(t)dB_t \), then

\[
\begin{align*}
    d\left( f(t, X_t, Y_t) \right) &= \partial_t f(t, X_t, Y_t)dt + \partial_X f(t, X_t, Y_t)\left( dX_t \right) + \partial_Y f(t, X_t, Y_t)\left( dY_t \right) \\
    &\quad + \frac{1}{2}\partial_{XX} f(t, X_t, Y_t)(dX_t)^2 + \frac{1}{2}\partial_{YY} f(t, X_t, Y_t)(dY_t)^2 + \partial_{XY} f(t, X_t, Y_t)(dX_t dY_t)
\end{align*}
\]

Date: 11/13/2014.
Notice that there’s no 1/2 on the $\partial_{XY}$. The reason is that $\partial_{XY} = \partial_{YX}$ provided $f$ has continuous second partials, so this term is actually $1/2(\partial_{XY} + \partial_{YX})$.

This gives the stochastic product rule:

$$d(X_tY_t) = X_t dY_t + Y_t dX_t + dX_t dY_t.$$  

It also gives the integration-by-parts formula: if, say, $Y_t$ is actually nonrandom, then $dX_t dY_t$ will consist of $dt^2$ or $dB dt$ terms, which are all zero. So, if $Y_t = f(t)$ is nonrandom, and $X_t = B_t$, then

$$d(f(t)B_t) = B_t f'(t) dt + f(t) dB_t,$$

$$\int_0^t f(t) dB_t = f(t) B_t - f(0) B_0 - \int_0^t B_s f'(s) ds.$$  

We then discussed Black-Scholes pricing of derivatives in a market with stock price $S_t = S_0 \exp((c - \frac{1}{2} \sigma^2)t + \sigma B_t)$ is a geometric Brownian motion ($c$ is the mean rate of return, $\sigma$ is the volatility), and bond price $\beta_t = \beta_0 \exp(rt)$ (i.e. interest rate compounded continuously).

1. One approach was to assume that a contract is given by a replicating portfolio. A self-financing portfolio satisfied

$$V_t = a_t S_t + b_t \beta_t, \quad dV_t = a_t dS_t + b_t dB_t,$$

where

$$dS_t = c S_t dt + \sigma S_t dB_t, \quad d\beta_t = r \beta_t dt.$$  

Assuming that $V_t = u(T - t, S_t)$ depends only on time remaining until execution and current stock price, we wrote down the Itô formula for $dV_t$, set it equal to the line above, and derived the Black-Scholes PDE

$$\partial_t u(t, s) = \frac{1}{2} \sigma^2 s^2 \partial_{ss} u(t, s) + rs \partial_s u(t, s) - ru(t, s).$$  

(This required setting the $dt$ and $dB_t$ portions of two SDEs equal to each other; we haven’t quite justified that step.) The initial condition on $u$ is what the contract pays out at time $T$, of the form $u(0, S_T) = f(S_T)$, and we can sometimes solve this for the value of the portfolio. We also saw how to extract the replicating portfolio from the value:

$$a_t = \partial_S u(t, S_t) = \partial_s u(t, s)|_{s=S_t}, \quad b_t = \beta_t^{-1} (u(t, S_t) - a_t S_t).$$  

In the case of the European call $u(0, s) = (s - K)_+$, we pulled a solution out of the hat:

$$u(t, s) = s \Phi \left( \frac{\ln(s/K) + (r + \frac{1}{2} \sigma^2 t)}{\sigma \sqrt{t}} \right) - Ke^{-rt} \Phi \left( \frac{\ln(s/K) + (r + \frac{1}{2} \sigma^2 t)}{\sigma \sqrt{t}} - \sigma \sqrt{t} \right),$$

where $\Phi$ is the normal CDF.

2. Noticing that the final answer did not depend on the mean rate of return, we said “let’s pretend that $c = r$, i.e. $e^{-rt} S_t$ is the geometric Brownian motion that’s a martingale”. This should be a change of measure, although we haven’t justified this part fully. Because discounted bond price is a constant (hence a martingale), and discounted stock price is a martingale also, we expect discounted value of anything to be a martingale. This lets us price contracts via

$$e^{-rt} V_t = \mathbb{E} \left( e^{-rT} V_T | \mathcal{F}_t \right),$$

where in computing $\mathbb{E}$ we use the fact that $e^{-rT} S_T = e^{-rt} S_t \cdot e^{-\frac{1}{2} \sigma^2 + \sigma (B_T - B_t)}$ for a Brownian motion $B$.  

2
In the context of final value being $V_T = (S_T - K)_+$, we argued that $\tilde{B}_T - \tilde{B}_t \sim N(0, T - t)$ is independent of $\mathcal{F}_t$, and everything else is $\mathcal{F}_t$-measurable, which let us replace the conditional expectation by taking the expectation with respect to this $N(0, T - t)$ random variable. This gave the answer

$$V(t, X_t) = e^{-r(T-t)} \frac{1}{\sqrt{2\pi \sigma \sqrt{T-t}}} \int_{-\infty}^{\infty} \left( S_t e^{(r-\sigma^2/2)(T-t)} e^u - K \right) e^{-\frac{1}{2} \frac{u^2}{2\sigma^2 (T-t)}} du,$$

and we decided that we could compute this integral if necessary (change of variables to get a standard normal CDF, and replace the positive part by bounds of integration), and the answer should look something like the Black-Scholes solution.

2. **Stochastic differential equations**

Consider the stochastic differential equation

$$dX_t = A_1(t, X_t)dt + A_2(t, X_t)dB_t, \quad X_t = X_0 + \int_0^t A_1(s, X_s)ds + \int_0^t A_2(s, X_s)dB_s.$$

We are interested in solutions where $X_t$ is adapted to the filtration of Brownian motion; these are called strong solutions. This means that $X_t$ is a function of the path of the Brownian motion up to time $t$. (It’s possible also to look for distributional equality; these are weak solutions.) Any solution of an SDE is called a diffusion.

Note that a Brownian motion is a diffusion, because it solves the SDE $dX_t = 0 \, dt + dB_t$.

**Theorem 2.1** (Existence and uniqueness of strong solutions). There exists a unique strong solution to an SDE (with initial conditions) if:

1. The initial condition $X_0$ is independent of the Brownian motion;
2. The coefficient functions $A_1(t, X_t)$ and $A_2(t, X_t)$ are continuous (in each variable);
3. The coefficient functions are Lipschitz in $X_t$.

Lipschitz formally means:

$$|a(t, y) - a(t, x)| \leq C |y - x|, \quad \forall x, y.$$

A good way of thinking about it is that it’s continuous, differentiable, and the derivative is bounded.

**Example 1.**

All linear functions are Lipschitz.

The functions $|x|$, $(x - K)_+$ are Lipschitz.

The function $f(x) = x^2$ is Lipschitz if $x$ is bounded, but not on $\mathbb{R}$ (the derivative tends to $\infty$ as $x \to \infty$). It would work as the coefficient function if $|X_t|$ is guaranteed to be bounded for some reason.

The function $f(x) = \sqrt{x}$ is Lipschitz on $[\epsilon, \infty)$ for any $\epsilon$, but is not Lipschitz on $[0, 1]$. It would work as the coefficient function if $X_t$ is guaranteed to be not too small.

The most general SDE that we will learn to solve is the general linear SDE:

$$dX_t = (c_1(t)X_t + c_2(t))dt + (\sigma_1(t)X_t + \sigma_2(t))dB_t.$$

This covers many interesting financial processes. In addition, we will figure out how to write the (ordinary) differential equation satisfied by the mean and variance of a solution to an SDE.
2.1. Interlude on solving linear ODEs. Before dealing with stochastic SDEs, let’s look at the linear ODE,
\[ \frac{dx}{dt} = c_1(t)x + c_2(t). \]
The general plan of solution is as follows:

1. Let \( p(t) = -\int_0^t c_1(s)ds \). Make sure to get the sign right!
2. The integrating factor is \( y(t) = e^{p(t)}. \)
3. The point is that
\[ \frac{d}{dt} \left( e^{p(t)}x \right) = e^{p(t)} \left( c_1(t)x + c_2(t) \right) + (-c_1(t)) e^{p(t)}x = e^{p(t)}c_2(t). \]
4. Therefore,
\[ e^{p(t)}x(t) - \frac{p(0)}{p(t)} x(0) = \int_0^t e^{p(s)}c_2(s)ds, \quad x(t) = e^{-p(t)} \left( x(0) + \int_0^t e^{p(s)}c_2(s)ds \right). \]

2.2. Solving the linear SDE. The trick with integrating factors above won’t work in quite the same way on SDEs, since the stochastic product rule involves more terms. That said, it does work when there’s no \( \sigma_1(t) \), i.e. when the noise part doesn’t involve \( X \) at all.

Example 2 (Additive noise). Suppose
\[ dX_t = (c_1(t)X_t + c_2(t))dt + \sigma(t)dB_t. \]
Let \( p(t) = -\int_0^t c_1(s)ds \) as before, then
\[ d(e^{p(t)}X_t) = e^{p(t)}((c_1(t)X_t + c_2(t))dt + \sigma(t)dB_t) - c_1(t)e^{p(t)}X_tdt \]
There are no extra terms, because this is linear in \( X \) (so no second derivatives in \( X \)), and there are no mixed terms because \( dp(t) \) involves only \( dt \) (no \( dB \)).
Consequently,
\[ e^{p(t)}X_t - X_0 = \int_0^t e^{p(s)}c_2(s)ds + \int_0^t e^{p(s)}\sigma(s)dB_s. \]

Example 3 (Vasicek interest rate model). In the Vasicek interest rate model,
\[ dr_t = C(\mu - r_t)dt + \sigma B_t. \]
When \( r_t \) gets away from \( \mu \), it tends to come back to it at rate proportional to \( \mu - r_t \), but there are fluctuations of order \( \sigma \) (the volatility).
Let’s solve: \( p(t) = \int_0^t Cds = Ct \), so
\[ e^{Ct}r_t - r_0 = \int_0^t e^{Cs}C\mu ds + \int_0^t e^{Cs}\sigma dB_s = \mu(e^{Ct} - 1) + \sigma \int_0^t e^{Cs}dB_s, \]
or
\[ r_t = e^{-Ct}r_0 + \mu(1 - e^{-Ct}) + \sigma e^{-Ct} \int_0^t e^{Cs}dB_s. \]
Let’s compute the mean and variance of the solution, using the fact that the mean of an integral \( dB \) is 0, and the variance of the deterministic part is also zero.
\[ \mathbb{E}r_t = e^{-Ct}\mathbb{E}r_0 + \mu(1 - e^{-Ct}), \]
and
\[ \text{Var}(r_t) = \text{Var} \left( \sigma e^{-Ct} \int_0^t e^{Cs} dB_s \right) = \sigma^2 e^{-2Ct} \int_0^t e^{2Cs} ds = \sigma^2 \frac{1}{2C} (1 - e^{-2Ct}). \]

Notice that the mean and variance tend to a limit as \( t \to \infty \); this is because \( r_t \) converges to a (normal) random variable in the limit of \( t \to \infty \).

Let’s now do an example where \( X_t \) appears in the \( dB \) and \( dt \) terms, but without any extra additive noise:

**Example 4 (Multiplicative noise).** Suppose
\[ dX_t = c(t) X_t dt + \sigma(t) X_t dB_t. \]
We expect the solution to be exponential, so we try \( Y_t = \ln X_t \). Then
\[
\begin{align*}
    dY_t &= \frac{1}{X_t} dX_t + \frac{1}{2} \frac{1}{X_t^2} (dX_t)^2 \\
    &= c(t) dt + \sigma(t) dB_t - \frac{1}{2} \sigma(t)^2 dt,
\end{align*}
\]
so
\[
\ln X_t = \ln X_0 + \int_0^t c(s) - \frac{1}{2} \sigma(s)^2 ds + \int_0^t \sigma(s) dB_s,
\]
or
\[
X_0 = X_0 \exp \left( \int_0^t c(s) - \frac{1}{2} \sigma(s)^2 ds + \int_0^t \sigma(s) dB_s \right),
\]
a variant on geometric Brownian motion.

3. **Mean and variance of solution of linear SDE**

Although we got the solution of a linear SDE somewhat explicitly, it may not be a terribly useful representation, and we often want to know things like the mean and variance of the solution. We could in principle find these from the representation of the solution, but there’s a simpler way.

Suppose
\[ dX_t = (c_1(t) X_t + c_2(t)) dt + (\sigma_1(t) X_t + \sigma_2(t)) dB_t, \]
meaning
\[ X_t = X_0 + \int_0^t (c_1(s) X_s + c_2(s)) ds + \int_0^t (\sigma_1(s) X_s + \sigma_2(s)) dB_s. \]
The key thing to notice is that the expected value of the stochastic integral is zero, so
\[ \mathbb{E} X_t = \mathbb{E} X_0 + \mathbb{E} \int_0^t (c_1(s) X_s + c_2(s)) ds. \]
Under some assumptions, expectation and integral commute, so in fact
\[ \mathbb{E} X_t = \mathbb{E} X_0 + \int_0^t (c_1(s) \mathbb{E} X_s + c_2(s)) ds. \]
This means that the mean of the SDE satisfies an ordinary differential equation,
\[ m(t) = m(0) + \int_0^t (c_1(s) m(s) + c_2(s)) ds, \quad m'(t) = c_1(t) m(t) + c_2(t). \]
This is a general linear ODE, which we know how to solve (find the integrating factor...)
Now let’s find $\mathbb{E}[X_t^2]$: clearly, knowing $\mathbb{E}[X_t^2]$ and $\mathbb{E}[X_t]$ is enough to find the variance. Using Itô lemma, $X^2$ satisfies its own SDE:

$$d(X_t^2) = 2X_t dX_t + (dX_t)^2 = \ldots$$

and so we can find its expected value in the same way as we found $\mathbb{E}[X_t]$; it will also satisfy a linear ordinary differential equation, namely

$$q'(t) = (2c_1(t) + \sigma_1(t)^2)q(t) + 2(c_2(t) + \sigma_1(t)\sigma_2(t))m(t) + \sigma_2^2(t).$$

Notice that the function $m(t)$ comes into the ODE, but $m(t)$ is by this point “known”.

In the same way we could in principle find differential equations satisfied by $\mathbb{E}[X_t^3]$, $\mathbb{E}[X_t^4]$, and so on, although it becomes increasingly unlikely that we’ll be able to write down a nice solution of them.

**Example 5 (Vasicek interest rate model).** Recall

$$dr_t = C(\mu - r_t)dt + \sigma B_t$$

had solution

$$r_t = e^{-Ct}r_0 + \mu(1 - e^{-Ct}) + \sigma e^{-Ct} \int_0^t e^{Cs} dB_s.$$  

The ODE for the mean is

$$m'(t) = C(\mu - m(t))dt.$$  

This is a very nice equation whose solution is negative exponential: formally, let $u(t) = \mu - m(t)$, then $u'(t) = -Cu dt$, so $u(t) = u(0)e^{-Ct}$ and

$$m(t) = \mu - u(0)e^{-Ct} = \mu - (\mu - m(0))e^{-Ct} = e^{-Ct}m(0) + \mu(1 - e^{-Ct}).$$

Now, to find the variance, we write

$$d(r_t^2) = 2r_t dr_t + (dr_t)^2$$

$$= 2r_t (C(\mu - r_t)dt + \sigma dB_t) + \sigma^2 dt$$

$$= (-2Cr_t^2 + 2C\mu r_t + \sigma^2) dt + 2r_t \sigma dB_t.$$  

Setting $q(t) = \mathbb{E}r_t^2$, we’re allowed to kill the $dB$ term, and take expectations of the $dt$ term:

$$q'(t) = -2Cq(t) + (2C\mu m(t) + \sigma^2)$$

This means that $w(t) = e^{2Ct}q(t)$ has derivative

$$w'(t) = e^{2Ct} (2C\mu m(t) + \sigma^2),$$

which is a nice integrable thing involving a bunch of exponentials. You can finish the computation yourself; remember that this is $\mathbb{E}[r_t^2]$, not $\text{Var}(r_t)$!
Important concepts from today: Girsanov’s theorem gives a change of measure that lets you kill off drift from Brownian motion; volatility is important but hard to measure.

1. Review of last time

A stochastic differential equation
\[ dX_t = A_1(t, X_t, B_t)dB_t + A_2(t, X_t, B_t)dt \iff X_t = X_0 + \int_0^t A_1(t, X_t, B_t)dB_t + \int_0^t A_2(t, X_t, B_t)dt. \]

A strong solution is a process \( X_t \) for which the SDE holds (in the integral formulation), and such that \( X_t \) is measurable with respect to \( \sigma(X_0; B_u, u \leq t) \) for all \( t \geq 0 \). A weak solution doesn’t require measurability.

**Theorem 1.1** (Existence and uniqueness of strong solutions). There exists a unique strong solution to an SDE (with initial conditions) on a time interval \( t \in [0, C] \) if:

1. The initial condition \( X_0 \) is independent of the Brownian motion;
2. The coefficient functions \( A_1(t, X_t, B_t) \) and \( A_2(t, X_t, B_t) \) are continuous (in all variables);
3. The coefficient functions are Lipschitz in \( X_t \).

Lipschitz means that when you change the second input, the output of the function changes by a bounded amount. The right way to think about it is that the partial derivative with respect to \( X \) is defined almost everywhere, and bounded.

Some examples of Lipschitz coefficient functions:

1. \( \sin(x) \), because the derivative is between \(-1, 1\).
2. \( \sin(t x) \), because the partial derivative is between \(-t, t\), hence between \(-C, C\).
3. \( t^{17} \sin(tx) \), because the partial derivative is between \(-t^{18}, t^{18}\), so between \(-C^{18}, C^{18}\).
4. \( (tx - 7)_+ \), because the partial derivative is either \( t \) (when \( x > 7/t \)) or 0, so between 0 and \( C \).
5. \( \sqrt{1 + x^2} \), because the partial derivative is \( x/\sqrt{1 + x^2} \) which is bounded (it tends to \(-1\) as \( x \to -\infty \), to 1 as \( x \to \infty \), and is defined and continuous everywhere, so it will be bounded on the entire real line.)

And here are some examples of non-Lipschitz functions:

1. \( x^2 \), because the derivative tends to \( \infty \) as \( x \to \infty \). (Remember, time is in a bounded interval; \( X_t \) is not.)
2. \( \ln(x) \), because the derivative tends to \( \infty \) as \( x \to 0 \).
3. \( 1/x \), because the derivative tends to \( \pm \infty \) as \( x \to 0 \).
Sometimes there is a unique strong solution even when the conditions of the theorem aren’t satisfied.

We then discussed how to solve linear ordinary differential equations, 

\[ x'(t) = c_1(t)x(t) + c_2(t), \quad \text{or} \quad dx_t = (c_1(t)x_t + c_2(t)) \, dt. \]

(1) Let \( p(t) = -\int_0^t c_1(s) \, ds \). Make sure to get the sign right!
(2) The integrating factor is \( y(t) = e^{p(t)} \).
(3) The point is that

\[
\frac{d}{dt} (e^{p(t)}x) = e^{p(t)} \left( c_1(t)x + c_2(t) + (-c_1(t)) e^{p(t)}x \right) = e^{p(t)}c_2(t).
\]

(4) Therefore,

\[
e^{p(t)}x(t) - \frac{e^{p(0)}x(0)}{p(0)=0} = \int_0^t e^{p(s)}c_2(s) \, ds, \quad x(t) = e^{-p(t)} \left( x(0) + \int_0^t e^{p(s)}c_2(s) \, ds \right).
\]

This method translates directly to the SDE with additive noise:

\[ dX_t = (c_1(t)X_t + c_2(t)) \, dt + \sigma(t)dB_t. \]

(1) Let \( p(t) = -\int_0^t c_1(s) \, ds \) as before.
(2) Then

\[
d(e^{p(t)}X_t) = e^{p(t)} ((c_1(t)X_t + c_2(t)) \, dt + \sigma(t)dB_t) - c_1(t)e^{p(t)}X_t \, dt
\]

(3) Consequently,

\[
e^{p(t)}X_t - X_0 = \int_0^t e^{p(s)}c_2(s) \, ds + \int_0^t e^{p(s)}\sigma(s)dB_s,
\]

or

\[
X_t = e^{-p(t)} \left( X_0 + \int_0^t e^{p(s)}c_2(s) \, ds + \int_0^t e^{p(s)}\sigma(s)dB_s \right).
\]

When the SDE has multiplicative noise, we use a different idea:

\[ dX_t = c(t)X_t \, dt + \sigma(t)X_tdB_t. \]

Because \( dX_t \) is proportional to \( X_t \), we expect the solution to be exponential, so let \( Y_t = \ln X_t \). Then

\[
y_t = \frac{1}{X_t}dX_t + \frac{-1}{2X_t^2} (dX_t)^2
\]

\[
= c(t)dX_t + \sigma(t)dB_t - \frac{1}{2} \sigma(t)^2 dt
\]

so

\[
\ln X_t = \ln X_0 + \int_0^t c(s) - \frac{1}{2} \sigma(s)^2 ds + \int_0^t \sigma(s)dB_s,
\]

or

\[
X_0 = X_0 \exp \left( \int_0^t c(s) - \frac{1}{2} \sigma(s)^2 ds + \int_0^t \sigma(s)dB_s \right),
\]

a variant on geometric Brownian motion.

Let’s now combine those two to get the general linear SDE:
Example 1 (General linear SDE). Suppose

\[ dX_t = (c_1(t)X_t + c_2(t))dt + (\sigma_1(t)X_t + \sigma_2(t))dB_t. \]

Recall that multiplying by the correct negative exponential could kill the \( c_1(t)X_tdt \) term. Similarly, multiplying by the right \( \text{Itô exponential} \) will kill \( c_1(t)X_tdt + \sigma_1(t)X_tdB_t \). Specifically: let

\[ dY_t = c_1(t)Y_tdt + \sigma_1(t)Y_tdB_t \]

and consider

\[
\begin{aligned}
&d \left( \frac{X_t}{Y_t} \right) \bigg|_{f(x,y)=x/y} = \frac{1}{Y_t} \frac{\partial}{\partial x} X_t - \frac{X_t}{Y_t^2} \frac{\partial}{\partial y} Y_t \\
&+ 0dX_t^2 + \frac{1}{2} \left[ \frac{2X_t}{Y_t^2} \frac{\partial}{\partial y} \sigma_1(t)^2Y_t^2dt + \frac{1}{Y_t^2} \frac{\partial}{\partial y} \frac{dX_tdB_t}{\sigma_1(t)^2Y_t^2} \right] \\
&= \frac{1}{Y_t} \left( c_2(t) - \sigma_1(t)\sigma_2(t) \right) dt + \frac{1}{Y_t} \sigma_2(t)dB_t.
\end{aligned}
\]

I am skipping some algebra where I expand terms and cancel them; you should check for yourself that it actually works out!

This is now an equation that we know how to solve:

\[
X_t = X_0 + \int_0^t \frac{1}{Y_s} \left( c_2(s) - \sigma_1(s)\sigma_2(s) \right) ds + \int_0^t \frac{1}{Y_s} \sigma_2(t)dB_t,
\]

where

\[
Y_t = \frac{Y_0}{\exp \left( \int_0^t \left( c_1(s) - \frac{1}{2}\sigma_1(s)^2 \right) ds + \int_0^t \sigma_1(s)dB_s \right)}.
\]

We also discussed how to set up an ODE for the mean and variance of a solution to a linear SDE. If

\[ dX_t = (c_1(t)X_t + c_2(t))dt + (\sigma_1(t)X_t + \sigma_2(t))dB_t, \]

then \( m(t) = \mathbb{E}[X_t] \) satisfies the linear ODE

\[ dm(t) = (c_1(t)m(t) + c_2(t))dt, \quad m'(t) = c_1(t)m(t) + c_2(t). \]

(We’ve replaced \( X_t \) by its mean on the right-hand side; the expectation of the Brownian term is zero.) This we already know how to solve. Moreover, we can write down the ODE for \( X_t^2 \):

\[ d(X_t^2) = 2X_tdB_t + (dX_t)^2 \]

This will not be a linear SDE, but when we take expectations (\( q(t) = \mathbb{E}[X_t^2] \)), we will get

\[ q'(t) = (2c_1(t) + \sigma_1(t)^2)q(t) + 2(c_2(t) + \sigma_1(t)\sigma_2(t))m(t) + \sigma_2^2(t). \]

This is a linear ODE for \( q(t) \), although it uses the function \( m(t) \) that we had to find above. We then find

\[ \text{Var}[X_t] = \mathbb{E}[X_t^2] - (\mathbb{E}[X_t])^2 = q(t) - m(t)^2. \]

The advantage of this method was that it gave us more readily implementable expressions for the mean and variance of a solution of an SDE, without having to first solve the SDE itself (it’s not necessarily obvious what the mean and variance of a solution are, even when we write down the solution).
2. Girsanov’s theorem

In our discussion of solving the Black-Scholes model we mentioned a change of measure that let us assume that discounted stock price is a martingale. Let’s discuss this in some more detail.

Theorem 2.1. Let $B_t$ be a Brownian motion, defined on the probability space $(\Omega, \mathbb{P})$. Fix a time interval $[0, T]$. There exists a different measure $Q$ on $\Omega$ such that the process $\tilde{B}_t = B_t + qt, \ 0 \leq t \leq T$ is a Brownian motion with respect to $Q$. The process $\tilde{B}_t$ is (obviously) adapted to the filtration of the original Brownian motion.

The measure $Q$ is given by

$$Q(A) = \int_A e^{-qB_T(\omega) - \frac{1}{2}q^2T}d\mathbb{P}(\omega).$$

Equivalently,

$$\frac{dQ}{d\mathbb{P}}(\omega) = e^{-qB_T(\omega) - \frac{1}{2}q^2T}.$$

We’ll use a slightly different characterization of Brownian motions than we had before. Specifically, we will show the following:

1. $\tilde{B}_0 = 0$ – obviously true;
2. Sample paths of $\tilde{B}$ are continuous – true, since true for $B$;
3. $\tilde{B}_t$ is a martingale (with respect to its natural filtration, which coincides with $B$’s natural filtration), and
   $$\mathbb{E}_Q[(\tilde{B}_t - \tilde{B}_s)^2] = t - s, \quad s < t.$$

That is, we need to show

$$\int_\Omega (\tilde{B}_t(\omega) - \tilde{B}_s(\omega))^2dQ(\omega) = t - s$$

and also that

$$\int_A \tilde{B}_t(\omega)1_A(\omega)dQ(\omega) = \int_A \tilde{B}_s(\omega)1_A(\omega)dQ(\omega)$$

whenever $A \in \mathcal{F}_s$.

We will also need to check that $Q$ is a measure, i.e. that $Q(\Omega) = 1$.

Let’s begin with a simpler calculation.

Example 2. Let $X : \Omega \rightarrow \mathbb{R}$ be a standard normal random variable, defined on a probability space $(\Omega, \mathbb{P})$. (There’s also a $\sigma$-algebra lurking around, but it will be the same throughout.) The density of $X$ is given by

$$\mathbb{P}(X \in (x, x + dx)) \approx f_X(x)dx = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}dx.$$

Now consider a different measure on $\Omega$, given by

$$\frac{dQ}{d\mathbb{P}}(\omega) = e^{\mu X(\omega) - \frac{1}{2}\mu^2}, \quad Q(A) = \int_A e^{\mu X(\omega) - \frac{1}{2}\mu^2}d\mathbb{P}(\omega).$$

First, let’s check this is a probability measure:

$$Q(\Omega) = \int_{\Omega} e^{\mu X(\omega) - \frac{1}{2}\mu^2}d\mathbb{P}(\omega) = \mathbb{E}[e^{\mu X - \frac{1}{2}\mu^2}] = \mathbb{E}[e^{N(-\frac{1}{2}\mu^2\mu^2)}] = e^0 = 1.$$

Under $Q$, the density of $X$ is

$$Q(X \in (x, x + dx)) \approx e^{\mu x - \frac{1}{2}\mu^2}P(X \in (x, x + dx)) = \frac{1}{\sqrt{2\pi}}e^{\mu x - \frac{1}{2}\mu^2}e^{-\frac{1}{2}x^2} = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu)^2}.$$
So, under $Q$, $X$ still has variance 1, but it now has mean $\mu$. That is:

The change of measure that turns $X \sim N(0, 1)$ into a $N(\mu, 1)$ random variable is

$$\frac{dQ}{dP}(\omega) = e^{\mu X(\omega) - \frac{1}{2} \mu^2}.$$

**Example 3.** Now suppose under $P$, $X \sim N(\mu, \sigma^2)$. How do we change measure to make it a $N(0, \sigma)$ random variable?

We know that $X = \sigma Y + \mu$, where $Y \sim N(0, 1)$. Then $Y = \sigma^{-1}(X - \mu)$, so making $X$ into a $N(0, \sigma^2)$ random variable is equivalent to making $Y$ into a $N(-\mu/\sigma, 1)$ random variable. We know the change of measure that does that: we need

$$\frac{dQ}{dP}(\omega) = e^{-\frac{\mu}{\sigma} Y(\omega) - \frac{1}{2} \frac{\mu^2}{\sigma^2}}.$$

Rewrite this in terms of $X$, then

$$\frac{dQ}{dP}(\omega) = e^{-\frac{\mu}{\sigma^2} X(\omega) + \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \frac{\mu^2}{\sigma^2}} = e^{-\frac{\mu}{\sigma^2} X(\omega) + \frac{\mu^2}{2\sigma^2}}.$$

We’ve already shown that this $Q$ is a probability measure.

Now let’s look at the more general case. We have $\tilde{B}_t = B_t + qt \sim N(qt, t)$. One change of measure that turns this into a $N(0, t)$ random variable would be

$$\frac{dQ}{dP}(\omega) = e^{-q\tilde{B}_t(\omega) + \frac{1}{2}q^2t} = e^{-q(B_t(\omega)+qt) + \frac{1}{2}q^2t} = e^{-qB_t(\omega) - \frac{1}{2}q^2t}.$$

This looks almost, but not quite right: here, the change of measure depends on $t$, and we want to show that the change of measure with $T$ will work at all times $t \leq T$. But one thing this argument shows is that all of these changes of measure do give a probability distribution $Q$. Another way to see that would be to notice that

$$Q(\Omega) = \int_{\Omega} e^{-qB_T(\omega) - \frac{1}{2}q^2T} = \mathbb{E}_P \underbrace{e^{-qB_T - \frac{1}{2}q^2T}}_{\text{martingale w.r.t. } T, \text{ at } T = 0 \text{ equal to 1}}.$$

Since the expected value of a martingale is the same for all $T$, we get that $Q(\Omega) = 1$ for all values of $T$.

Next, let’s show that $\tilde{B}_t$ is a martingale with respect to $Q$. This requires showing that

$$\mathbb{E}_Q \left( \tilde{B}_t | \mathcal{F}_s \right) = \tilde{B}_s,$$

which means that for every $\mathcal{F}_s$-measurable set $A$ we need to show that

$$\int_A \tilde{B}_t(\omega) dQ(\omega) = \int_A \tilde{B}_s(\omega) dQ(\omega).$$

Rewrite $dQ(\omega) = (dQ/dP(\omega)) \cdot dP(\omega)$, then we’re showing

$$\int_A \tilde{B}_t(\omega) \frac{dQ}{dP}(\omega) dP(\omega) = \int_A \tilde{B}_s(\omega) \frac{dQ}{dP}(\omega) dP(\omega),$$

or

$$\mathbb{E}_P \left( \tilde{B}_t \frac{dQ}{dP} | \mathcal{F}_s \right) = \mathbb{E}_P \left( \tilde{B}_s \frac{dQ}{dP} | \mathcal{F}_s \right).$$

On the right-hand side, we have

$$\mathbb{E}_P \left( \tilde{B}_s \frac{dQ}{dP} | \mathcal{F}_s \right) = \tilde{B}_s \mathbb{E}_P (e^{-qB_T - \frac{1}{2}q^2T}| \mathcal{F}_s) = \tilde{B}_s e^{-qB_s - \frac{1}{2}q^2s}.$$
On the left-hand side, we condition on \( \mathcal{F}_t \) first:
\[
\mathbb{E}_\mathbb{P}\left( \mathbb{E}_\mathbb{P}\left( \tilde{B}_t \frac{dQ}{d\mathbb{P}} | \mathcal{F}_t \right) | \mathcal{F}_s \right) = \mathbb{E}_\mathbb{P}\left( \tilde{B}_t e^{-qB_t - \frac{1}{2}q^2t} | \mathcal{F}_s \right).
\]

I claim that \( \tilde{B}_t e^{-qB_t - \frac{1}{2}q^2t} \) is a \( \mathbb{P} \)-martingale; we can check this:
\[
d((B_t + qt) e^{-qB_t - \frac{1}{2}q^2t}) = (dB_t + qdt) e^{-qB_t - \frac{1}{2}q^2t} + (B_t + qt) \left(-q e^{-qB_t - \frac{1}{2}q^2t} \cdot dB_t\right)
\]
Itô exponential is a martingale
\[
= \text{something} \cdot dB_t.
\]

Consequently, the left-hand side is in fact equal to \( \tilde{B}_s \exp(-qB_s - \frac{1}{2}q^2s) \), as required.

Here’s a summary of what we observed:

\[\begin{align*}
\text{Because } \tilde{B}_s &= \mathbb{E}(\tilde{B}_s | \mathcal{F}_s) \text{ under any measure (i.e. because } \tilde{B}_s \text{ is adapted to the filtration, so is } \mathcal{F}_s \text{-measurable), I can use } \mathbb{E}(dQ/d\mathbb{P} | \mathcal{F}_s) \text{ as my change of measure that gives me the distribution of } \tilde{B}_s. \text{ That is, the single change of measure } e^{-q\tilde{B}_T - \frac{1}{2}q^2T} \text{ actually turns into the change of measure } e^{-qB_t - \frac{1}{2}q^2s} \text{ at time } s.
\end{align*}\]

This shows in particular that \( \tilde{B}_s \sim \mathcal{N}(0, s) \) under \( Q \) at all times, but we still need to show that \( \mathbb{E}[(\tilde{B}_T - \tilde{B}_s)^2] = t - s \).

Let’s do this again by conditioning:
\[
\mathbb{E}_Q[(\tilde{B}_T - \tilde{B}_s)^2] = \mathbb{E}_Q[\mathbb{E}_Q((\tilde{B}_T - \tilde{B}_s)^2 | \mathcal{F}_s)] = \mathbb{E}_Q[\mathbb{E}_Q(\tilde{B}_s^2 | \mathcal{F}_s) - \tilde{B}_s^2]
\]
Here I’ve expanded \( (\tilde{B}_T - \tilde{B}_s)^2 \) and have used the fact that \( \tilde{B} \) is a \( Q \)-martingale: \(-2\mathbb{E}(\tilde{B}_t | \mathcal{F}_s)\tilde{B}_s + \tilde{B}_s^2 = -\tilde{B}_s^2\).

Now, we already know \( \mathbb{E}_Q[-\tilde{B}_s^2] = -s \), since we’ve just shown \( \tilde{B}_s \sim \mathcal{N}(0, s) \), and the other term is
\[
\mathbb{E}_Q[\mathbb{E}_Q(\tilde{B}_s^2 | \mathcal{F}_s)] = \mathbb{E}_Q[\tilde{B}_s^2] = \mathbb{E}_Q[\mathbb{E}_Q(\tilde{B}_s^2 | \mathcal{F}_t)] = \mathbb{E}_Q[N(0, t)] = t.
\]
This finishes the demonstration that
\[
\mathbb{E}_Q[(\tilde{B}_T - \tilde{B}_s)^2] = t - s,
\]
which is what we wanted to show.

There is a stronger version of Girsanov’s theorem, which says that we can cancel not just linear drift, but in fact any deterministic drift:

**Theorem 2.2.** Suppose \( dX_t = u(t)dt + dB_t \) with \( X_0 = 0 \), where \( B \) is a Brownian motion under \( \mathbb{P} \). Then under the measure \( Q \) given by
\[
\frac{dQ}{d\mathbb{P}}(\omega) = e^{-\int_0^T u(s)dB_s - \frac{1}{2} \int_0^T u(s)^2ds},
\]
the process \( X_t \) is a Brownian motion.

We can even take away random drifts, not just deterministic ones; the only issue there is that integrability becomes harder to check in advance.

The summary of this is that we’ve shown how to convert the physical measure that describes how we think the world actually works into the equivalent martingale (or risk-neutral) measure. This is useful if you have good models (or simulations) for how the world will work, and want to transform them into statements about the risk-neutral situation in order to find values of contracts.

It’s reasonably clear that the converse to the Girsanov theorem can’t possibly be true. Certainly in the one-dimensional case, the Gaussian measure is absolutely continuous
with respect to all sorts of measures, so there are changes of measure that will turn into just about anything. Life is a little more constrained when we’re looking at entire trajectories, but even still, there should be lots and lots of measures that are absolutely continuous with respect to the given one. We do, however, get a converse if we assume that there’s nothing in the world except for the driving Brownian motion:

**Theorem 2.3.** If $B_t$ is a Brownian motion on $(\Omega, \mathcal{F}_t, \mathbb{P})$, and $Q \ll \mathbb{P}$ is an absolutely continuous measure defined on $\mathcal{F}_T$ for some fixed $T$. Let

$$L_t = \mathbb{E} \left( \frac{dQ}{d\mathbb{P}} \mid \mathcal{F}_t \right).$$

There exists an adapted process $\phi_t$ such that

$$dL_t = L_t \phi_t dB_t, \quad L_0 = 1,$$

or

$$L_t = \exp \left( \int_0^t \phi_t dB_t - \frac{1}{2} \int_0^t \phi_t^2 dt \right).$$

This shows that when we looked at Girsanov’s theorem, we’ve already considered all the allowed changes of measure: in particular, since Girsanov’s theorem didn’t change volatility, nothing changes volatility. **Volatility is a fundamental parameter in the Black–Scholes model.** This holds as long as we demand that everything should be adapted to the Brownian filtration.

The proof of this result has a lot to do with the following theorem, which we’ve also mentioned already:

**Theorem 2.4** (Martingale representation theorem). Let $B$ be a Brownian motion, and let $M$ be a martingale adapted to the Brownian filtration. Then there exists a unique adapted process $\phi_t$ such that

$$M_t = M_0 + \int_0^t \phi_t dB_t.$$

As always, we need some integrability conditions here, i.e. we need to require $\mathbb{E}[M_t^2] < \infty$ and then the stochastic integral is actually well-defined. The take-away message for today’s results is:

In the Brownian world, $M_t$ is a martingale $\iff dM_t = \text{something} \cdot dB_t$. Changes of measure can change the drift (the $dt$ term) arbitrarily, but if everything is adapted to the Brownian filtration, changing measure cannot affect the volatility (the $dB_t$ term).
Important concepts from today: markets with multiple assets; incompleteness; how to model (and try to compute things with) dividends.

1. Review of last time

Last time we stated Girsanov’s theorem:

**Theorem 1.1.** Let $B_t$ be a Brownian motion, defined on the probability space $(\Omega, \mathbb{P})$. Fix a time interval $[0, T]$. There exists a different measure $Q$ on $\Omega$ such that the process $\tilde{B}_t = B_t + q_t, \ 0 \leq t \leq T$ is a Brownian motion with respect to $Q$. The process $\tilde{B}_t$ is (obviously) adapted to the filtration of the original Brownian motion.

The measure $Q$ is given by

$$Q(A) = \int_A e^{-qB_T(\omega) - \frac{1}{2}q^2T} d\mathbb{P}(\omega).$$

Equivalently,

$$\frac{dQ}{d\mathbb{P}}(\omega) = e^{-qB_T(\omega) - \frac{1}{2}q^2T}.$$

More generally, suppose a stochastic process $X_t$ satisfies $dX_t = u(t)dt + dB_t$ with $X_0 = 0$. Then under the change of measure

$$\frac{dQ}{d\mathbb{P}}(\omega) = e^{-\int_0^T u(s)dB_s - \frac{1}{2} \int_0^T u(s)^2ds},$$

$X_t$ is distributed as a Brownian motion.

It’s possible to extend this result even further, in particular to random functions $u(t)$, but you need to make sure that the change of measure you get is actually a nice random variable, i.e. $u(t)$ is integrable. Homework problem 2c shows an example where $u(s) = 1/B_s$ turns out not to be integrable.

The interesting part of the statement of the theorem is that there is a single change of measure that works for all times $t \leq T$; changing measure to fix the mean at any one time is easy, but doing it at all times with one change of measure is the tricky part. The reason this worked was that when we look at $\tilde{B}_s$, which is an $\mathcal{F}_s$-measurable random variable, we only care about $\mathbb{E}(dQ/d\mathbb{P}|\mathcal{F}_s)$, not about what the change of measure does on the entire finer $\sigma$-algebra $\mathcal{F}_T$. It turns out that for the change of measure we’re considering, this conditional expectation is exactly the change of measure that removes the drift from $X_s$.

There was also a converse to Girsanov theorem for processes adapted to the Brownian filtration.
Theorem 1.2. Let $B_t$ be a Brownian motion on $(\Omega, \mathbb{P})$ with filtration $\mathcal{F}_t$, and let $Q \ll \mathbb{P}$ be an absolutely continuous measure defined on $\mathcal{F}_T$ for some fixed $T$. Then we can write

$$\mathbb{E}(\frac{dQ}{d\mathbb{P}} | \mathcal{F}_t) = \exp \left( \int_0^t \phi_s dB_s - \frac{1}{2} \int_0^t \phi_s^2 ds \right)$$

for some $\phi(s)$.

This looks exactly like the changes of measure in Girsanov’s theorem, with $\phi(s) = -u(s)$, which means that as long as we’re only interested in processes adapted to the Brownian filtration, all changes of measure have the same form as in Girsanov’s theorem. In particular, they all can change the drift, but not the volatility of a diffusion.

The take-away message was that a diffusion is a martingale if and only if it has no $dt$ term, and we can usually change a diffusion into a martingale by changing measures in a way that removes the $dt$ term. One of the things we’ll discuss today is when can you change multiple diffusions into martingales simultaneously.

2. IMPLIED VOLATILITY

The basic version of Girsanov’s change of measure turns $B_t + qt$ into a Brownian motion $\tilde{B}_t$, and therefore turns $e^{(\sigma q - \frac{1}{2} \sigma^2) t + \sigma B_t}$ into $e^{-\frac{1}{2} \sigma^2 + \sigma \tilde{B}_t}$, another geometric Brownian motion with the same volatility $\sigma$. The converse to Girsanov’s theorem can be used to show that you can’t change $\sigma$ if you insist on getting adapted processes. In particular,

In the Black–Scholes model, you really need to know the volatility $\sigma$.

How can you determine it? There are two approaches:

1. Historic volatility: look at historic stock prices, and determine the running average of $\sigma$ over a time period. The problem with this method is that $\sigma$ itself might change with time. Girsanov’s theorem tells us that we should be able to determine $\sigma$ from arbitrarily short time periods if we were really looking at trajectories of a geometric Brownian motion; but in real life we aren’t — over very short (seconds) time scales, stock price comes from limit order book dynamics and doesn’t look like a geometric Brownian motion.

2. Suppose you’re in a market where there are options being traded on the same underlying stock. If we assume that they were priced according to the Black–Scholes model, then we can use them to determine the market value of $\sigma$, or the implied volatility. To do this, we would look at the prices for which the options are being traded today, would compute those prices as functions of $\sigma$ in the Black–Scholes model, and would then try to fit a single value of $\sigma$. This allows you to get instantaneous values of $\sigma$. Unfortunately, if you do this in practice on many options (with different strikes or maturity times), you get multiple values of $\sigma$. Fixing execution time, there is a “volatility smile” for volatility as a function of strike price: at very high or very low strike prices, the volatility implied by the option prices is higher than at “moderate” strike prices. There is also a term structure of volatility, which describes how the implied volatility changes as a function of the execution time for options on the same stock.

Example 1. Suppose that the current stock price is $S_0 = 100$, interest rate is $r = 0.01$, and we have a call option with strike price $K = 100$ and execution time $T = 3.7$. (These numbers should look familiar from the homework.) The price of the option is an increasing
function of $\sigma$; it’s equal to about 3.63 for $\sigma \to 0$, and about 66.99 for $\sigma = 1$. Here’s a plot of it over a broader range. Because this curve is invertible, if we know the option price we can find the market value of $\sigma$.

![Figure 1. Price of a call option ($S_0 = 100$, $K = 100$, $r = 0.01$, $T = 3.7$) as a function of the volatility $\sigma$. We can invert this plot to find $\sigma$ from the price.]

3. Higher-dimensional models

In real life, markets have more than one stock in them. A more realistic model is this: we have one riskless bond, whose price is given by $d\beta_t = r(t)\beta_t dt$, and $n$ stocks whose prices satisfy

$$dS_i(t) = S_i(t)c_i(t)dt + S_i(t)\sum_{j=1}^{n} \sigma_{ij}(t)dB_j^i.$$  

Of course, we also have initial conditions for all of these. Here, $B^1, B^2, \ldots, B^n$ are independent Brownian motions. $r(t)$ is the (time-dependent instantaneous) volatility, $c_i(t)$ is the (instantaneous) mean rate of return, and $\sigma_{ij}(t)$ are dispersion coefficients. This is saying that the rate of growth of a stock is proportional to its current price, with noise that’s also proportional to price but may be correlated for different stocks; the best way to model that is to start with independent Brownian motions and put a covariance structure on top. In particular, each $S_i(t)$ is still a geometric Brownian motion here (with mean rate of return $c_i(t)$, and volatility $\sum_j \sigma_{ij}^2(t)$), but they are correlated among themselves. (See today’s homework!)

To deal with multidimensional models, we add one more rule to our arsenal of Itô rules:

$$dt^2 = 0, \quad dB_i dt = 0, \quad dB_i^2 = dt, \quad dB_i^i dB_j^j = 0 \text{ when } i \neq j.$$  

(To derive this: $\mathbb{E}[dB_i(t)dB_j(t)] = \mathbb{E}[dB_i(t)]\mathbb{E}[dB_j(t)] = 0$, and $\text{Var}(dB_i(t)dB_j(t)) = dt^2$; when you integrate, because increments of Brownian motion are independent across time, $\text{Var}(\int_0^t dB_i(t)dB_j(t)) = \int_0^t \text{Var}(dB_i(t)dB_j(t)) = O(dt)$. Here I’m treating the integral as a sum, and saying that the variance of the sum is the sum of variances.) To do Itô calculus with these objects, you take partial derivatives in all variables out to second order.
(remembering that $1/2$ goes in front of any second partial with respect to a single variable, and doesn’t go in front of mixed partials), and expand the products of differentials according to those rules. For example, if $dX_t = dt + dB_1(t) + dB_2(t)$, then

$$d(B_1^2(t) \sin(X_t)) = 2B_1(t) \sin(X_t)dB_1(t) + \left( dB_1(t)^2 \sin(X_t) + B_1(t)^2(t) \cos(X_t) dX_t - \frac{1}{2} B_1(t)^2(t) \sin(X_t) \right) dt = (dB_1(t))^2 + dB_1(t)dB_2(t) + dB_2(t)dB_1(t)$$

A portfolio in the multidimensional market consists of $n + 1$ numbers, $N_i(t)$, corresponding to the number of shares of each asset that I hold. I will take $N_0(t)$ to be the number of bonds. Thus,

$$V_t = N_0(t) \beta(t) + \sum_{i=1}^n N_i(t) S_i(t).$$

It’s self-financing if

$$dV_t = N_0(t) d\beta(t) + \sum_{i=1}^n N_i(t) dS_i(t).$$

The analog of Girsanov’s theorem in higher dimensions asserts that we can change to an equivalent risk-neutral measure in which all the mean rates of return are $r(t)$. The risk-neutral measure is given by

$$\frac{dQ}{dP}(\omega) = \exp \left( - \sum_{i=1}^n \int_0^T \theta_i(s) dB_i^t - \frac{1}{2} \sum_{i=1}^n \int_0^T \theta_i(s)^2 \right),$$

where

$$\theta_i(t) = \sum_{j=1}^n \left( \sigma(t)^{-1} \right)_{ij} (b_j(t) - r(t)).$$

Here, $\sigma(t)$ is the matrix with entries $\sigma_{ij}(t)$.

If we change to this measure, then we can price contingent claims as follows:

$$e^{-\int_0^T r(s) ds} V_T = \mathbb{E}_Q \left( e^{-\int_0^T r(s) ds} V_T | \mathcal{F}_t \right).$$

Here, we take $V_T$ as given, as a function of (all the) terminal prices $S_T$. We’ve discounted by the risk-free interest rate, and then taken conditional expectation, because the discounted value of the portfolio should be a martingale under $Q$.

It is also possible to set up a PDE for $V_i = u(T - t, S_1^t, \ldots, S_n^t)$ using Itô rules of differentiation, and then hope to solve the PDE. In the PDE version, we would use the fact that coefficients on $dt, dS_1^t, dS_2^t, \ldots, dS_n^t$ are all uniquely determined. This means that when you write down

$$dV(T - t, S_1(t), S_2(t), \ldots, S_n(t)) = N_0(t) d\beta(t) + \sum_{i=1}^n N_i(t) dS_i(t),$$

you have $n + 1$ equations here, which you can solve for $N_i(t)$ in terms of partial derivatives of $V$. The one remaining equation

$$V_t = N_0(t) \beta(t) + \sum_{i=1}^n N_i(t) S_i(t)$$

then gives you a PDE for $V$.

We casually wrote $\sigma(t)^{-1}$ in the Girsanov change-of-measure; this assumes that the matrix $\sigma(t)$ is invertible. What happens when it isn’t? The answer is that one of the stock prices is a (predictable) function of the other stocks, because one of the underlying
Brownian motions is a linear combination of the others. This means that very likely there’s arbitrage in the market, unless the one stock was priced exactly right. In practice, if you see covariance matrices that have very small eigenvalues (close to 0), you start looking for the linear combinations that have particularly low variance, and there might be arbitrage opportunities there.

4. INCOMPLETE MARKETS

We talked about what happens when you have fewer independent sources of information than tradeable assets; let’s now say something about the opposite case, a market where there are some extra non-tradeable sources of information. For example, you may have access to historical weather patterns, and you may be interested in pricing a contract based on this information (travel insurance). This increases the amount of information and the filtration, but doesn’t increase your hedging flexibility. For a simple example, suppose that in the binomial market model we had

\[
S_{t+1} = \begin{cases}
U \cdot S_t \\
M \cdot S_t \\
D \cdot S_t 
\end{cases}
\]

\( (M \text{ for middle}) \). Then there wouldn’t be a unique way to make the stock prices into a martingale, and equivalently, there wouldn’t be a way to exactly replicate some contracts using just stocks and bonds. For example, the contract that pays out \( S_t^2 \) at time 1 would be impossible to replicate: we cannot solve three equations

\[
a + bU = U^2, \quad a + bM = M^2, \quad a + bD = D^2
\]

for two unknowns \( a \) and \( b \).

The converse to this is that there are many values I could assign to the probabilities \( p, q, 1 - p - q \) or going Up, Middle, and Down that would make the discounted stock price into a martingale: indeed, I need to solve one equation with two unknowns, \( pU + qM + (1 - p - q)D = (1 + r) \). (Of course, I also need \( p > 0, q > 0, p + q < 1 \).) I could use any of these equivalent measures to derive arbitrage-free prices: there are many equivalent martingale measures, hence possibilities for pricing an option in an arbitrage-free way in this market.

On the other hand, as soon as I introduce some freely-traded options into this market, I start to be able to price options relative to each other. The mathematics in the Black-Scholes model gets a bit more involved, because option prices have different distributions than stock prices; but in a binomial market model it’s an entirely reasonable computation. A good example is that if I have a stock whose price I don’t tell you, and also a bond and an option on the stock, then I should be able to work out the price of the stock from the price of the option. (Think of option price as a function of \( s \) in the Black-Scholes model.)
Informal result on number of sources of information vs. number of traded assets:

<table>
<thead>
<tr>
<th># sources of info vs. # traded assets</th>
<th>what happens?</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; (non-invertible covariance)</td>
<td>there might be arbitrage in the market (if something was priced incorrectly)</td>
</tr>
<tr>
<td>&gt; (non-tradeable information)</td>
<td>many arbitrage-free prices possible, let the market decide (e.g. what maximizes your profit)</td>
</tr>
<tr>
<td>=</td>
<td>unique arbitrage-free price, unique replicating portfolio, you can figure out the correct price for all new contracts</td>
</tr>
</tbody>
</table>

5. Things we didn’t talk about

There are lots of contracts we didn’t discuss, some of which you should be seeing in FM 5021/5022:

(1) Options on options. I can put an option on any financial product, so why not on another option? It will have a value, but computing that value is an interesting problem! (Mainly because option price doesn’t evolve nearly as nicely as stock price.) But in the binomial market this isn’t hard to work out: first compute the value of the underlying option, then replicate a contract that has that payout when it’s executed. (You’ll have to keep track of two deadlines and two strike prices.)

(2) American options can be executed at any time until a deadline. Figuring out the price of an American option with a finite deadline is quite hard.

(3) Asian options have payout that depends on the average stock price until the deadline. This turns out to be easier to price, because the average stock price follows an SDE, and you just need to look at the system of SDEs for the value and this average.

(4) We didn’t talk about anything with dividends, i.e. the case when there are extra money flows (annuities where you get money from the contract, stocks with dividends, taking money out from the portfolio for consumption)

(5) Contracts on commodities are different from contracts based on financial products, because borrowing (short-selling) a commodity is harder than borrowing (short-selling) a financial product.

(6) If you’re playing on two markets – domestic and foreign – then the exchange rate will be coming into the picture as well.

(7) ...

Other books to read:

(1) John Hull, Options, futures, and other derivatives
(2) Steven Shreve, Stochastic calculus for finance, vol. 2
(3) Thomas Björk, Arbitrage theory in continuous time