Stochastic Calculus Notes

John R. Boccio

October 26, 2012

Contents

Chapter 1

Lecture 1

1.1 Basic terminology

Here are some basic definitions and ideas of probability. These might seem dry without examples. Be patient. Examples are coming in later sections. Although the topic is elementary, the notation is taken from more advanced probability so some of it might be unfamiliar. The terminology is not always helpful for simple probability problems, but it is just the thing for describing stochastic processes and decision problems under incomplete information.

1.1.1 Do an "experiment" or "trial", get an "outcome", ω.

The set of all possible outcomes is Ω. We often call Ω the "probability space". The probability is "discrete" if Ω is finite or countable (able to be listed in a single infinite numbered list). For now, we do only discrete probability.

1.1.2 The probability of a specific outcome is $P(\omega)$.

We always assume that $P(\omega) \geq 0$ for any $\omega \in \Omega$ and that $\sum_{\omega \in \Omega} P(\omega) = 1$. The interpretation of probability is a matter for philosophers, but we might say that $P(\omega)$ is the probability of outcome ω happening, or the fraction of times event ω would happen in a large number of independent trials. The philosophical problem is that it may be impossible to actually perform a large number of independent trials. People also sometimes say that probabilities represent our often subjective (lack of) knowledge of future events. Probability 1 is something that is certain to happen while probability 0 is for something that cannot happen.

1.1.3 "Event": a set of outcomes, a subset of Ω.

The probability of an event is the sum of the probabilities of the outcomes that make up the event $P(A) = \sum_{\omega \in A} P(\omega)$. We do not distinguish between the outcome ω and the event that that outcome occurred $A = \{\omega\}$. That is, we write $P(\omega)$ for $P(\{\omega\})$ or vice versa. This is called "abuse of notation": we use notation in a way that is not absolutely correct but whose meaning is clear. It's the mathematical version of saying "I could care less" to mean the opposite.

1.1.4 Example: Toss a coin 4 times.

Each toss yields either H (heads) or T (tails). There are 16 possible outcomes, TTTT, TTTH, TTHT, TTHH, THTT, …, HHHH. The number of outcomes is $\#(\Omega) = |\Omega| = 16$. Normally each outcome is equally likely, so $P(\omega) = \frac{1}{16}$ for each $\omega \in \Omega$. If A is the event that the first two tosses are H, then

$$A = \{\text{HHHH, HHHT, HHTH, HHTT}\} \ .$$

There are 4 elements (outcomes) in A, each having probability $\frac{1}{16}$ Therefore

$$P(\text{first two H}) = P(A) = \sum_{\omega \in \Omega} P(\omega) = \sum_{\omega \in \Omega} \frac{1}{16} = \frac{4}{16} = \frac{1}{4}$$

1.1.5 Set operations:

Events are actually sets so set operations apply to events. If A and B are events, the event "A and B" is the set of outcomes in both A and B. This is the set intersection $A \cap B$. The union $A \cup B$ is the set of outcomes in A or in B (or in both). The complement of A, A^c, is the event "not A", the set of outcomes not in A. Events A and B are disjoint if they have no elements in common. The empty event is the empty set, the set with no elements, $\emptyset$. The probability of $\emptyset$ should be zero because the sum that defines it has no terms: $P(\emptyset) = 0$. The complement of $\emptyset$ is Ω. Events A and B are disjoint if $A \cup B = \emptyset$. Event A is contained in event B, $A \subseteq B$, if every outcome in A is also in B.

1.1.6 Basic facts:

Each of these facts is a consequence of the representation $P(A) = \sum_{\omega \in A}$. First $P(\omega)P(A) \leq P(B)$ if $A \subseteq B$. Also, $P(A) + P(B) = P(A \cup B)$ if A and B are disjoint: $A \cup B = \emptyset$. From this it follows that $P(A) + P(A^c) = P(\Omega) = 1$.

1.1.7 Conditional probability:

The probability of outcome A given that B has occured is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \ . \tag{1.1}$$

This is the percent of B outcomes that are also A outcomes. The formula is called "Bayes' rule". It is often used to calculate $P(A \cap B)$ once we know $P(B)$ and $P(A \mid B)$. The formula for that is $P(A \cap B) = P(A \mid B)P(B)$.

1.1.8 Independence:

Events A and B are independent if $P(A \mid B) = P(A)$. That is, knowing whether of not B occured does not change he probability of A. In view of Bayes' rule, this is expressed as

$$P(A \cap B) = P(A) \cdot P(B) \ . \tag{1.2}$$

For example, suppose A is the event that two of the four tosses are H and B is the event that the first toss is H. Then A has 6 elements (outcomes), B has 8, and, as you can check by listing them, $A \cap B$ has 3 elements. Since each element has probability $\frac{1}{16}$, this gives $P(A \cap B) = \frac{3}{16}$ while $P(A) = \frac{6}{16}$ and $P(B) = \frac{8}{16} = \frac{1}{2}$. We might say "duh" for the last calculation since we started the example with the hypothesis that H and T were equally likely. Anyway, this shows that (1.2) is indeed satisfied in this case. This example is supposed to show that while some pairs of events, such as the first and second tosses, are "obviously" independent, others are independent as the result of a calculation. Note that if C is the event that 3 of the 4 tosses are H (instead of 2 for A), then $P(C) = \frac{4}{16} = \frac{1}{4}$ and $P(B \cap C) = \frac{3}{16}$, because

$$B \cap C = \{\text{HHHT, HHTH, HTHH}\}$$

has three elements. Bayes' rule (1.1) gives $P(B \mid C) = \frac{1}{16}/\frac{3}{4} = \frac{3}{4}$. Knowing that there are 3 heads in all raises the probability that the first toss is H from $\frac{1}{2}$ to $\frac{3}{4}$.

1.1.9 Working with conditional probability:

Conditional probability is like ordinary (unconditional) probability. Once we know that the event B occured, the probability of outcome ω is given by Bayes' rule

$$P(\omega \mid B) = \begin{cases} \frac{P(\omega)}{P(B)} & \text{for} \omega \in B, \\ 0 & \text{for} \omega \notin B. \end{cases}$$

That is, we shrink the probability space from Ω to B and "renormalize" the probabilities by dividing by $P(B)$ so that they again sum to one:

$$\sum_{\omega \in B} P(\omega \mid B) = 1 \ .$$

We can apply the rules of conditional probability to conditional $P(\omega \mid B)$ probabilities themselves. If $\tilde{P}(\omega) = P(\omega \mid B)$, we can condition on another event, C. What is the probability $\tilde{P}$ of ω given that C occurred? If $\omega \notin C$ it is zero.

If $\omega \in C$, it is, repeated using Bayes' rule,

$$\begin{aligned}
\tilde{P}(\omega \mid C) &= \frac{\tilde{P}(\omega)}{\tilde{P}(C)} \\
&= \frac{P(\omega \mid B)}{P(C \mid B)} \\
&= \frac{P(\omega)}{P(B)\dfrac{P(C \cap B}{P(B)}} \\
&= \frac{P(\omega)}{P(B \cap C)} \\
&= P(\omega \mid B \cap C) \ .
\end{aligned}$$

The conclusion is that conditioning on B and then on C is the same as conditioning on $B \cap C$ (B and C) all at once.

1.1.10 Algebra of sets and incomplete information:

A set of events, $\mathcal{F}$, is an "algebra" if

i: $A \in \mathcal{F}$ implies that $A^c \in \mathcal{F}$.

ii: $A \in \mathcal{F}$ and $B \in \mathcal{F}$ implies that $A \cup B \in \mathcal{F}$ and $A \cap B \in \mathcal{F}$.

iii: $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$.

We interpret $\mathcal{F}$ as representing a state of partial information. We know whether any of the events in $\mathcal{F}$ occurred but we do not have enough information to determine whether an event not in $\mathcal{F}$ occurred. The above axioms are natural in light of this interpretation. If we know whether A happened, we surely know whether "not A" happened. If we know whether A happened and whether B happened, then we can tell whether "A and B" happened. We definitely know whether $\emptyset$ happened (it did not) and whether Ω happened (it did). Events in $\mathcal{F}$ are called "measurable" or "determined in $\mathcal{F}$". You will often see the term σ–algebra, or sigma algebra, instead of just "algebra". The distinction between σ–algebra and algebra is technical and only arises when Ω is infinite, and rarely then.

1.1.11 Example:

Suppose we know only outcomes only of the first two tosses. One event measurable in $\mathcal{F}$ is

$$\{\text{HH}\} = \{\text{HHHH, HHHT, HHTH, HHTT}\} \ .$$

This is something of an abuse of notation; get used to it. An example of an event not determined by this $\mathcal{F}$ is the event of no more than one H:

$$A = \{\text{TTTT, TTTH, TTHT, THTT, HTTT}\} \ .$$

Just knowing the first two tosses does not tell you with certainty whether the total number of heads is less than two.

1.1.12 Another example:

Suppose we know only the results of the tosses but not the order. This might happen if we toss 4 identical coins at the same time. In this case, we know only the number of H coins. Some measurable sets are (with an abuse of notation)

$$\begin{aligned} \{4\} &= \{\text{HHHH}\} \\ \{3\} &= \{\text{HHHT, HHTH, HTHH, THHH}\} \\ &\vdots \\ \{0\} &= \{\text{TTTT}\} \end{aligned}$$

The event $\{2\}$ has 6 outcomes (list them), so its probability is $6 \cdot \frac{1}{16} = \frac{3}{8}$. There are other events measureable in this algebra, such as "less than 3 H", but, in some sense, the events listed "generate" the algebra.

1.1.13 Terminology:

What we call "outcome" is sometimes called "random variable". I don't use this because it can be confusing in that we often think of variables as real or complex numbers. A "real valued function" of the random variable ω is a real number X for each ω, written $X(\omega)$. The most common abuse of notation in probability is to write X instead of $X(\omega)$. We will do this most of the time, but not just yet. We often think of X as a random number whose value is determined by the outcome (random variable) ω. A common convention is to use upper case letters for random numbers and lower case letters for specific values of that variable. For example, the "cumulative distribution function" (CDF), $F(x)$ is the probability that $X \leq x$, that is $F(x) = \sum_{X(\omega) \leq x} P(\omega)$.

1.1.14 Informal event terminology:

We often describe events in words. For example, we might write $P(X \leq x)$ where, strictly, we might be supposed to say $B_x = \{\omega \mid X(\omega) \leq\}$ then $P(X = leqx) = P(B_x)$. If there are two functions, X_1 and X_2, we might try to calculate, for example, $P(X_1 = X_2)$, which is actually the probability of the set of ω so that $X_1(\omega) = X_2(\omega)$.

1.1.15 Measurable:

A function (of a random variable) $X(\omega)$ is measurable with respect to the algebra $\mathcal{F}$ if the value of X is completely determined by the information in $\mathcal{F}$. To give a mathematical definition, for any number, x we can consider the event that $X = x$, which is $A_x = \{\omega \ : \ X(\omega) = x\}$. In discrete probability, A_x will be the empty set for almost all x values and be another set only for those values of x actually taken by $X(\omega)$ for one of the outcomes ω. The function $X(\omega)$ is "measurable with respect to $\mathcal{F}$ if the sets A_x are all measurable. People often write $X \in \mathcal{F}$ (an abuse of notation) to indicate that X is measurable with respect to $\mathcal{F}$. In the second example above, the function $X =$ number of H minus number of T is measurable, while the function $X =$ number of T before the first H is not.

1.1.16 Generating an algebra of sets:

Suppose there are events A_1, ..., A_k that you know. The algebra, $\mathcal{F}$, generated by these sets is the algebra that expresses the information about the outcome you gain by knowing these events. One definition of $\mathcal{F}$ is that an event A is in $\mathcal{F}$ if A can be expressed in terms of the known events A_j using the set operations intersection, union, and complement a number of times. For example, we could define an event A by saying "ω is in A_1 and (A_2 or A_3) but not A_5 or A_5". An equivalent to saying that $\mathcal{F}$ is the smallest algebra of sets that contains the known events A_j. Obviously (think about this!) any algebra that contains the A_j contains any event described by set operations on the A_j, that is the definition of algebra of sets. Also the sets defined by set operations on the A_j form an algebra of sets. For example, if A_1 is the event that the first toss is H and A_2 is the event that the second toss is H, then A_1 and A_2 generate the algebra of events determined by knowing the results of the first two tosses. This is example 1 above.

1.1.17 Generating by a function:

A function $X(\omega)$ defines an algebra of sets generated by the sets A_x. This is the smallest algebra, $\mathcal{F}$, so that X is measurable with respect to $\mathcal{F}$. Example 2 above has this form. We can think of $\mathcal{F}$ as being the algebra of sets defined by statements about the values of $X(\omega)$. For example, one $A \in \mathcal{F}$ would be the set of ω with X either between 4 and 5 or greater than 11.

We write $\mathcal{F}_X$ for the algebra of sets generated by X and ask, what it means that another function of ω, $Y(\omega)$, is measurable with respect to $\mathcal{F}_X$. The information interpretation of $\mathcal{F}_X$ says that $Y \in \mathcal{F}_X$ if knowing the value of $X(\omega)$ determines the value of $Y(\omega)$. This means that if ω_1 and ω_2 have the same X value ($X(\omega_1) = X(\omega_2)$) then they also have the same Y value. Said another way, if A_x is not empty, then there is some number, $u(x)$, so that $Y(\omega) = u(x)$ for every $\omega \in A_x$. This means that $Y(\omega) = u(X(\omega))$ for all $\omega \in \Omega$). Altogether, saying $Y \in \mathcal{F}_X$ is a fancy way of saying that Y is a function of X. Of course, $u(x)$ only needs to be defined for those values of x actually taken by the random variable X.

For example, if X is the number of H in 4 tosses, and Y is the number of H minus the number of T, then, for any 4 tosses, ω, $Y(\omega) = 2X(\omega) - 4$. That is, $u(x) = 2x - 4$.

1.1.18 Expected value:

A random variable (actually, a function of a random variable) $X(\omega)$ has expected value
$$E[X] = \sum_{\omega \in \Omega} X(\omega) P(\omega) \; .$$
(Note that we do not write ω on the left. We think of X as simply a random number and ω as a story of how X was generated.) This is the "average" value in the sense that if you could perform the "experiment" of sampling X vary many times and average the resulting numbers, you would get roughly $E[X]$. This is because $P(\omega)$ is the fraction of the time you would get ω and $X(\omega)$ is the number you get for ω. If $X_1(\omega)$ and $X_2(\omega)$ are two random variables, then $E[X_1 + X_2] = E[X_1] + E[X_2]$. Also, $E[cX] = cE[X]$ if c is a constant (not random).

1.1.19 Best approximation property:

If we wanted to approximate a random variable, X, (function $X(\omega)$ with ω not written) by a single non random number, x, what value would we pick? That would depend on the sense of "best". One such sense is "least squares", choosing x to minimize the expected value of $(X - x)^2$. A calculation, which uses the above properties of expected value, gives
$$\begin{aligned} E\left[(X-x)^2\right] &= E[X^2 - 2Xx + x^2] \\ &= E[X^2] - 2xE[X] + x^2 \; . \end{aligned}$$
Minimizing this over x gives the optimal value
$$x_{\text{opt}} = E[X] \; . \tag{1.3}$$

1.1.20 Conditional expectation, elementary version:

There are two senses of the term "conditional expectation". We start with the original sense then turn to the related but different sense often used in stochastic processes. Conditional expectation is defined from conditional probability in the obvious way
$$E[X|B] = \sum_{\omega} X(\omega) P(\omega|B) \; .$$
For example, we can caluclate
$$E[\#\text{of H in 4 tosses} \;\mid\; \text{at least one H}] \; .$$
Write B for the event {at least one H}. Since only ω =TTTT does not have at least one H, $|B| = 15$ and $P(\omega \mid B) = \frac{1}{15}$ for any $\omega \in B$. Let X be the number of H. Unconditionally, $E[\text{H}] = 2$ (see below). This means that
$$\frac{1}{16} \sum_{x \in \Omega} X(\omega) = 2 \; .$$

Note that $X(\omega) = 0$ for all $\omega \notin B$ (only TTTT), so that implies that

$$\begin{aligned}
\frac{1}{16}\sum_{\omega\in B} X(\omega)P(\omega) &= 2 \\
\frac{15}{16}\cdot\frac{1}{15}\sum_{\omega\in B} X(\omega)P(\omega) &= 2 \\
\frac{1}{15}\sum_{\omega\in B} X(\omega)P(\omega) &= \frac{2\cdot 16}{15} \\
E[X \mid B] &= \frac{32}{15} \quad = \quad 2 + .133\ldots \quad .
\end{aligned}$$

Knowing that there was at least on H increases the expected number of H by $.133\ldots$.

1.1.21 Conditional expectation, modern version:

The modern conditional expectation starts with an algebra, $\mathcal{F}$, rather than just a set. It defines a (function of a) random variable, $Y(\omega) = E[X \mid \mathcal{F}]$, that is measurable with respect to $\mathcal{F}$ even though X is not. This function represents the best prediction of X given the information in $\mathcal{F}$. In the elementary case (paragraph 1.20), the information is the occurence or non occurence of a single event, B. In this case, the algebra, $\mathcal{F}_B$ consists only of the sets B, B^c, $\emptyset$, and Ω. The modern definition gives a function $Y(\omega)$ so that

$$Y(\omega) = \begin{cases} E[X \mid B] & \text{if } \omega \in B, \\ E[X \mid B^c] & \text{if } \omega \notin B. \end{cases}$$

Make sure you understand the fact that this two valued function Y is measurable with respect to $\mathcal{F}_B$.

Only slightly more complicated is the case where $\mathcal{F}$ is generated by a "partition" of Ω. A partition is a collection of events B_1, $\ldots$, B_n, so that each outcome, ω is in one and only one of the events. The sets $\{4\}$, $\{3\}$, $\ldots$, $\{0\}$ in paragraph 1.12 form a partition, as do the sets A_x in paragraph 1.15 (if you keep only the A_x that are not empty). The algebra of sets generated by the sets in a partition consists of unions of sets in the partition (think this through). The conditional expectation $Y(\omega) = E[X \mid \mathcal{F}]$ is defined to be

$$Y(\omega) = E[X \mid B_j] \;\text{ if } \omega \in B_j \quad ,$$

where $E[X \mid B_j]$ is in the elementary sense of paragraph 1.20. This is well defined because there is exactly one B_j for each ω. A single set B defines a partition: $B_1 = B$, $B_2 = B^c$, so this agrees with the earlier definition in that case.

Finally, as long as the probability space, Ω is finite, any algebra of sets is generated by some partition. The events in the partition are events in $\mathcal{F}$ that cannot be subdivided within $\mathcal{F}$.

1.1.22 Best approximation property:

Suppose we have a random variable, $X(\omega)$, that is not measurable with respect to the algebra of sets $\mathcal{F}$. That is, the information in $\mathcal{F}$ does not completely determine the values of X. The conditional expectation, $Y(\omega) = E[X \mid \mathcal{F}]$, has the property that it is the best approximation to X among functions measurable with respect to Y, in the least squares sense. That is, if $\tilde{Y} \in \mathcal{F}$, then

$$E\left[(\tilde{Y} - X)^2\right] \geq E\left[(Y - X)^2\right] \ .$$

In fact, this later will be the definition of conditional expectation in situations where the partition definition is not directly applicable. Suppose $\mathcal{F}$ is generated by the partition B_1, ..., B_n. Any random variable $\tilde{Y} \in \mathcal{F}$ is determined by it's (constant) values on the sets B_k: $\tilde{Y}(\omega) = \tilde{y}_k$ for $\omega_k \in B_k$. Just as in paragraph 1.19, the best value for $\tilde{y}_k$ is $E[X \mid B_j]$.

1.2 Markov Chains, I

Markov[1] chains form a simple class of stochastic processes. They seem to represent a good level of abstraction and generality: many practical models are Markov chains. Here we discuss Markov chains in "discrete time" (the continuous time version is called a "Markov process) and having a finite "state space" (see below). We also suppose that the "transition probabilities" are stationary, i.e. independent of time.

1.2.1 Time:

The time variable, t, will be an integer representing the number of time units from a starting time. The actual time between t and $t+1$ could be a nanosecond (for modeling computer communication networks) or a month (for modeling bond rating changes), or whatever.

1.2.2 State space:

At time t the system will be in one of a finite list of states. This set of states is the "state space", $\mathcal{S}$. To be a Markov chain, the "state" should be a "complete" description of the actual state of the system at time t. This means that it should contain any information about the system at time t that helps predict the state

[1] The Russian mathematician A. A. Markov was active in the last decades of the 19^{th} century. He is known for his path breaking work on the distribution of prime numbers as well as on probability.

at future times $t+1$, $t+2$, ... This will be more clear soon. The state at time t will be called $X(t)$ or X_t. Eventually, there may be an ω also, so that the state is a function of t and ω: $X(t,\omega)$ or $X_t(\omega)$. The states may be called $s_1, \ldots, s_m$, or simply $1, 2, \ldots, m$. depending on the context.

1.2.3 Path space:

The sequence of states X_1, X_2, ..., X_T, is a “path”. The set of paths is “path space”. This path space is the probability space, Ω, for the Markov chain. An outcome is completely determined by the sequence fo states in the path. That is, in the case of a Markov chain, there might not be a distinction between the path $X = (X_1, \ldots, X_m)$ and the outcome ω. We will soon have a formula for the probablity of any path X. An event is a collection of paths such as the set of all paths that do not contain state s_6 or the set of paths that end in $X_T = s_1$, etc. The number of paths of length T is m^T, where $m = |\mathcal{S}|$ is the number of states. As a practical matter this (albeit finite) number is often too large for computation. For example, for 7 states and 10 steps ($m = 7$, $T = 10$) we have $|\Omega| = 7^{10} = 28,2475,294 \approx 3 \cdot 10^8$. A 1GHz computer would take at least an hour to list and calculate the probability of each path.

1.2.4 Transition probabilities:

The transition probability, P_{jk}, is the probability of going from state j to state k in one step. That is:

$$P_{jk} = P\left(X_{t+1} = k \mid X_t = j\right) \ .$$

The Markov chain is “stationary” if the transition probabilities P_{jk} are independent of t. Each transition probability P_{jk} is between 0 and 1, with values 0 and 1 allowed, though 0 is more common than one. Also, with j fixed, the P_{jk} must sum to 1 (summing over k) because $k = 1, 2, \ldots, m$ is a complete list of the possible states at time $t+1$.

1.2.5 Transition matrix:

These transition probabilities form an $m \times m$ matrix, P (an unfortunate conflict of notation). The (j,k) entry of P being the transition probability P_{jk}. The sum of the entries of the transition matrix P in row j is $\sum_k P_{jk} = 1$. A matrix with these properties: no negative entries, all row sums equal to 1, is a “stochastic matrix”. Any stochastic matrix can be the transition matrix for a Markov chain. Methods from linear algebra often enter into the analysis of Markov chains. For example, the time s transition probability

$$P^s_{jk} = P(X_{t+s} = k \mid X_t = j)$$

is the (j,k) entry of P^s, the s^{th} power of the transition matrix (explanation below). The “steady state” probabilities form an eigenvector of P.

1.2.6 Path probabilities:

The Markov property allows us to compute the probability of any path or portion of a path by multiplying transition probabilities. For example, suppose we want the probability of the successive transitions $i \to j \to k$. This is $P(X_{t+1} = j \text{ and } X_{t+2} = k \mid X_t = i)$. Using the conditional Bayes' rule, this is

$$P(X_{t+2} = k \mid X_{t+1} = j \text{ and } X_t = i) \cdot P(X_{t+1} = j \mid X_t = i) \;.$$

Here the Markov property comes in. It states that if we know X_{t+1}, the value of X_t is irrelevant in predicting X_{t+2}. That is

$$P(X_{t+2} = k \mid X_{t+1} = j \text{ and } X_t = i) = P(X_{t+2} = k \mid X_{t+1} = j) = P_{jk} \;.$$

Combining the above two facts, we get

$$\begin{aligned} P(i \to j \to k) &= P(X_{t+1} = j \text{ and } X_{t+2} = k \mid X_t = i) \\ &= P_{ij} \cdot P_{jk} \;. \end{aligned}$$

To give the probability of a whole path, $X = (X_1, \ldots, X_T)$, we have to give the "initial distribution" probabilities for X_1 and the transition probabilities. The transition probabilities take care of the rest. We will call the probabilities for X_1 f^1 or $f(1)$. That is, $P(X_1 = j) = f^1_j$. The latter may also be written $f(j, 1)$. In general we use notation $f^t_j = P(X_t = j)$. Using f^1 and the P_{jk}, we can calculate the probabilities of paths:

$$P(X_1 = j \text{ and } X_2 = k) = f^1_j \cdot P_{jk} \;,$$

$$P(X_1 = j \text{ and } X_2 = k \text{ and } X_3 = l) = f^1_j \cdot P_{jk} \cdot P_{jk} \;,$$

and so on. Expressed slightly differently, we have

$$P(X) = f^1_{X_1} \cdot P_{X_1,X_2} \cdot \;\cdots\; \cdot P_{X_{T-1},X_T} \;. \tag{1.4}$$

1.2.7 Example 3, coin flips:

The state space has $m = 2$ states, called U (up) and D (down). H and T would conflict with T being the length of the chain. Let us consider paths of length $T = 50$. Example 1 has paths of length 4. Let us suppose that a coin starts in the U position. At every time step, the coin turns over with 20% probability. The transition probabilities are $P_{UU} = .8$, $P_{UD} = .2$, $P_{DU} = .2$, $P_{DD} = .8$. The transition matrix is (taking U for 1 and D for 2):

$$P = \begin{pmatrix} .8 & .2 \\ .2 & .8 \end{pmatrix}$$

For example, we can calculate

$$P^2 = P \cdot P = \begin{pmatrix} .68 & .32 \\ .32 & .88 \end{pmatrix} \quad \text{and} \quad P^4 = P^2 \cdot P^2 = \begin{pmatrix} .5648 & .4352 \\ .4352 & .5648 \end{pmatrix} \;.$$

This implies that $P(X_5 = U) = P(X_1 = U \to X_5 = U) = P^5_{UU} = .5648$

1.2.8 Example 4, hidden Markov model:

There are two coins, F (fast) and S (slow). Either coin will be either U or D at any given time. Only one coin is present at any given time but sometimes the coin might be replaced (F for S or vice versa) without changing its U–D status. The F coin has the same U–D transition probabilities as example 3. The S coin has U–D transition probabilities:

$$\begin{pmatrix} .9 & .1 \\ .05 & .95 \end{pmatrix}$$

The probability of coin replacement at any given time is 30%. The replacement (if it happens) is done after the (possible) coin flip without changing the U–D status of the coin after that flip. The Markov chain has 4 states, which can be numbered (somewhat arbitrarily) 1: UF, 2: DF, 3: US, 4: DS. States 1 and 3 are U states while states 1 and 2 are F states, etc. The transition matrix is 4×4. We can calculate, for example, the (non) transition probability for UF → UF. We first have a U → U (non) transition then an F → (non) transition. The probability is then $P(U \to U \mid F) \cdot P(F \to F) = .8 \cdot .7 = .56$. The other entries can be found in a similar way. The transitions are:

$$\begin{pmatrix} UF \to UF & UF \to DF & UF \to US & UF \to DS \\ DF \to UF & DF \to DF & DF \to US & DF \to DS \\ US \to UF & US \to DF & US \to US & US \to DS \\ DS \to UF & DS \to DF & DS \to US & DS \to DS \end{pmatrix} .$$

The resulting transition matrix is

$$P = \begin{pmatrix} .8 \cdot .7 & .2 \cdot .7 & .8 \cdot .3 & .2 \cdot .3 \\ .2 \cdot .7 & .8 \cdot .7 & .2 \cdot .3 & .8 \cdot .3 \\ .9 \cdot .7 & .1 \cdot .7 & .9 \cdot .3 & .1 \cdot .3 \\ .05 \cdot .7 & .95 \cdot .7 & .05 \cdot .3 & .95 \cdot .3 \end{pmatrix} .$$

If we start with UF and want to know the probability of being D after 4 time periods, the answer is $P^4_{12} + P^4_{14}$ because states $2 = DF$ and $4 = DS$ are the two D states.

1.2.9 Example 5, incomplete state information:

In the model of example 4 we might be able to observe the U–D status but not F–S. Suppose $Y_y = U$ if $X_t = UF$ or $X_t = UD$, and $Y_t = D$ if $X_t = DF$ or $X_t = DD$. Then the sequence Y_t is a stochastic process but it is not a Markov chain. We can better predict $U \leftrightarrow D$ transitions if we know whether the coin is F or S, or even if we have a basis for guessing. For example, suppose $Y_8 = U$ and we want to guess whether Y_9 will again be U. If Y_7 is D then we are more likely to have the F coin so a $Y_8 = U \to Y_9 = D$ transition is more likely. That is, with Y_8 fixed, $Y_7 = D$ makes it less likely to have $Y_9 = U$. This is a violation of the Markov property brought about by incomplete state information. Models of this kind are called “hidden markov” models. We suppose that there is a Markov chain but that we have incomplete information about it. Statistical estimation of the unobserved variable is a topic for another day.

Chapter 2

Lecture 2

2.1 Simple Random Walk

Simple random walk is a great example of Markov chains. We will see lots of the big topics: martingales, forward and backward equations, change of measure, hitting times, etc. First for random walk, then for more general cases (Markov chains), then for harder cases (stochastic processes, diffusions).

2.1.1 Definition:

The state space is the set of all integers, 0, ±1, etc. The path starts with $X_0 = 0$. At each time, the path moves at most one unit to the left or right. I will use the notation p for the probability of moving to the right, q for the probability of moving to the left, and $1-(p+q)$ for the probability of not moving. Warning: this is not standard notation. Formally, this can be written as:

$$\begin{aligned} p &= P(x \to x+1) = P(X_{t+1} = x+1 \mid X_t = x) \ , \\ q &= P(x \to x-1) = P(X_{t+1} = x-1 \mid X_t = x) \ , \\ 1-(p+q) &= P(x \to x) = P(X_{t+1} = x \mid X_t = x) \ . \end{aligned}$$

The walk is called “symmetric” or “unbiased” if $p = q$. Otherwise, the walk is unsymmetric or biased. Some of the analysis is simpler if $p+q = 1$ so that there are only two possible states at time $t+1$ instead of 3.

2.1.2 Quantities of interest:

Our work will focus on calculating various quantities. The most basic is $E[V(X_t)]$ for some “payout” function $V(x)$. Another involves hitting times, the first time X_t reaches some set. For example, we could define τ to be the first time $|X_t| = 5$. Other quantities of interest might be average values such as

$$E\left[\frac{1}{T}\sum_{t=1}^{T} V(X_t)\right] \ ,$$

$$E\left[\max_{1\leq t\leq T} X_t^2\right] \ ,$$

and so on.

2.1.3 $\mathcal{F}_t$:

The algebra of sets determined by knowing the path up to and including time t is called $\mathcal{F}_t$. This algebra is generated by the random variables X_1, …, X_t. A partition (see Lecture 1, paragraph 1.21) of sets generating $\mathcal{F}_t$ is given as follows: for every fixed path, $x = (x_1, \ldots, x_t)$, of length t, associate B_x, the set of all paths that have the same values up to time t: $X_1 = x_1$, …, $X_t = x_t$. We can calculate $P(B_x)$ by multiplying the transition probabilities for the given transitions $x_s \to x_{s+1}$ as in Lecture 1, paragraph 2.6.

2.1.4 Conditioning on $\mathcal{F}_t$:

Many of the techniques for computing quantities such as in paragraph 1.2 are based on computing the conditional expectations with respect to $\mathcal{F}_t$. The Markov property makes this possible. Suppose that f_t is some function of the path that is determined by the path up to time t, $f_t(X) \in \mathcal{F}_t$. This means that the value of f_t is determined by the random variables X_1, …, X_t (see lecture 1, paragraph 1.17). Another way to say this is that f_t must be constant on the basic partition sets B_x, that is, have a single value for each sequence x_1, …, x_t. Maybe the best way to say it is that f_t is determined by the information up to time t and that this information is precisely the values of X_s for $1 \leq s \leq t$. Suppose that $F(X)$ is some function of the path, X, up to time T (examples in paragraph 1.2). Then define

$$f_t = E[F \mid \mathcal{F}_t] \ .$$

The unconditional expectations are f_0, which is a constant (since it doesn't depend on anything).

2.1.5 Backwards equations:

The technique for computing f_0 is to compute first $f_T = F$, then compute f_{T-1} from f_T, and in general to have a way to compute f_t from f_{t+1}. An equation for f_t in terms of f_{t+1} is a "backward equation" because time runs backward. A fair part of the course is finding and solving backward equations. Backward equations are simple because going from $\mathcal{F}_{t+1}$ to $\mathcal{F}_t$ removes only one variable, X_{t+1}. By the principle of repeated expectations (lecture 1, paragraph 1.?), because there is more information in $\mathcal{F}_{t+1}$ than in $\mathcal{F}_t$ ($\mathcal{F}_t \subset \mathcal{F}_{t+1}$),

$$f_t = E[f_{t+1} \mid \mathcal{F}_t] \ .$$

Now, in the case of random walk, if we have the $\mathcal{F}_t$ information, we know, in particular, the value of X_t. This leaves only 3 possible values of X_{t+1}. Therefore, the expectation on the left above is just a sum of 3 terms (see paragraph 1.1):

$$\begin{aligned} f_t(X_1, \ldots, X_t) &= p \cdot f_{t+1}(X_1, \ldots, X_t, X_t + 1) \\ &+ q \cdot f_{t+1}(X_1, \ldots, X_t, X_t - 1) \\ &+ (1 - (p+q)) \cdot f_{t+1}(X_1, \ldots, X_t, X_t) \ . \end{aligned}$$

In most cases where this backward equation is practical, the functions f_t and the backward equation are simpler than the general case given here, usually because of the special form of the original F and the Markov property.

2.1.6 Expected payouts:

A simple and useful illustration of these ideas is the “final payout” case $F(X) = V(X_T)$. In that case, the functions f_t depend on X_t only. If we call that variable just n, we have

$$f_t(n) = p \cdot f_{t+1}(n+1) + q \cdot f_{t+1}(n-1) + (1 - (p+q)) \cdot f_{t+1}(n) \ .$$

This is truly practical, computing one function of one variable in terms

Chapter 3

Lecture 3

3.1 Recurrence relations for Markov Chains

3.1.1 Recapitulation and notation:

To summarize terminology for Markov Chains (lecture 1, paragraph 2.??)

$\mathcal{F}_t$: the algebra generated by X_1, …, X_t. The partition generating $\mathcal{F}_t$ consists of sets B_x where $x = (x_1, \ldots, x_t)$, is an initial segment of a path of lenght t. The sets are $B_x = \{X \mid X_1 = x_1, \ldots, X_t = x_t\}$. To check your understanding, show that the number of paths in B_x is s^{T-t}, where s is the number of states: $s = |\mathcal{S}|$. This algebra represents knowing the path X up to and including time t. Being measurable with respect to $\mathcal{F}_t$ means being constant on each of the sets B_x, i.e. a function of X_1, …, X_t.

$\mathcal{G}_t$: the algebra generated by X_t alone. The patrition generating $\mathcal{G}_t$ consists of one set, B_j, for each state $j \in \mathcal{S}$. Then $B_j = \{X \mid X_t = j\}$. There are s such B_j, each with s^{T-1} paths. This algebra represents knowing only the present state but not past or future states. Being measurable with respect to $\mathcal{G}_t$ means being constant on each of the B_j, i.e. a function of j.

$\mathcal{H}_t$: the algebra generated by X_t, …, X_T. This represents knowledge of the present and all future states.

The Markov property is that

$$E\left[F(X) \mid \mathcal{G}_t\right] = E\left[F(X) \mid \mathcal{F}_t\right] \ ,$$

for any $F \in \mathcal{H}_t$ (i.e. F depending only on present and future states). This is the modern version. The classical expression for the same property is that if F depends only on X_t, …, X_T, then

$$E\left[F(X) \mid X_t = j\right] = E\left[F(X) \mid X_t = j, X_{t-1} = k, \ldots\right] \ .$$

3.1.2 The "law of total probability":

The classical theorem is that if B_k, $k = 1, \ldots, n$ is any partition of Ω, then, for any function,

$$E\left[F\right] = \sum_{k=1}^{n} P(B_k) E\left[F \mid B_k\right] \ .$$

It is easy to verify this using the (classical) definition of conditional expectation. This is a special case of a relation about modern style conditional expectation: if $\mathcal{F}$ and $\mathcal{G}$ are two algebras with $\mathcal{G} \subset \mathcal{F}$, ($\mathcal{G}$ has less information) then

$$E[F \mid \mathcal{G}] = E\left[E[F \mid \mathcal{F}] \mid \mathcal{G}\right] \ .$$

The classical statement corresponds to the modern one with $\mathcal{G}$ being the "trivial" algebra consisting of only $\emptyset$ and Ω. A classical statement of the more general modern version might start with any event, A. The relation is

$$E[F \mid A] = \sum_{k=1}^{n} P(B_k \mid A) \cdot E[F \mid B_k \text{ and } A] \ .$$

3.1.3 Backward equation, classical version:

The simplest case is when we want the expected value of a "payout", V, that depends only on the final state: $F(X) = V(X_T)$. It is possible to compute $E[V(X_T)]$ as the byproduct of a system collection of calculations of related quantities:

$$f_t(j) = E\left[V(X_T) \mid X_t = j\right] \ .$$

We apply the law of total probability to the right side with A being the event $X_t = j$ and B_k defined respectively by $X_{t+1} = k$. The Markov property implies that

$$E\left[V(X_T) \mid X_{t+1} = k \text{ and } X_t = j\right] = E\left[V(X_T) \mid X_{t+1} = k\right] = f_{t+1}(k).$$

This gives:

$$\begin{aligned} f_t(j) &= \sum_{k=1}^{s} P(X_{t+1} = k \mid X_t = j) \cdot E[V(X_T) \mid X_{t+1} = k \text{ and } X_t = j] \\ f_t(j) &= \sum_{k=1}^{s} P_{jk} f_{t+1}(k) \ . \qquad (3.1) \end{aligned}$$

This gives us a way to calculate all the $f_t(j)$ working backwards in time. The final values $f_T(j)$ are clearly given by

$$F_T(j) = E[V(X_T) \mid X_T = j] = V(j) \ .$$

Then we can use (3.1) to compute all the values f_{T-1}, then all the values f_{T-2}, and so on. It is a major shortcoming of the backward equation method that you must compute the values of $f_t(j)$ for each state $j \in \mathcal{S}$. In many cases $\mathcal{S}$, though finite, is too large for such computations to be practical. Backward equation, classical version:

3.1.4 Backward equation, matrix version:

The equation (3.1) may be expressed in matrix terms. For each t, define a vector, f_t, with s components given by

$$f_t = (f_t(1), \ldots, f_t(s))^* \; .$$

The notation $(f_t(1), \ldots, f_t(s))^*$ refers the column vector that is the transpose of the row vector $(f_t(1), \ldots, f_t(s))$. We will make good use of the distinction between row vectors, which may be thought of an $1 \times s$ matrices, and column vectors, which may be thought of as $s \times 1$ matrices. The recurrence relation (3.1) is equivalent to

$$f_t = P \cdot f_{t+1} \; . \tag{3.2}$$

Here, P is the transition matrix defined in lecture 1, paragraph 2.??, and the right side is interpreted as matrix multiplication. If f_t were a row vector, the expression $P \cdot f_{t+1}$ would not make sense as matrix multiplication. The recurrence relation (3.2) may be iterated to give

$$f_{t-k} = P^k f_t \; .$$

3.1.5 Forward equation, classical version:

The backward equation describes the evolution of expectation values while the forward equation describes the evolution of probabilities. We use the notation

$$u_t(j) = P(X_t = j) \; .$$

We can compute the time $t+1$ probabilities in terms of the time t probabilities using the law or total probability above. We wish to compute $u_{t_1}(k) = P(X_{t+1} = k)$ and the partition is the s events $B_j = \{X_t = j\}$. This gives

$$\begin{aligned} u_{t+1}(k) &= P(X_{t+1} = k) \\ &= \sum_{j=1}^{s} P(X_t = j) P(X_{t+1} = k \mid X_t = j) \\ u_{t+1}(k) &= \sum_{j=1}^{s} u_t(j) P_{jk} \end{aligned} \tag{3.3}$$

This is a forward moving evolution equation that allows us to compute the probability distribution at later times from the distribution at earlier times.

3.1.6 Initial data and path probabilities:

A point I've ignored until now is that the transition matrix alone does not determine the probabilities. We also need the initial probabilities $P(X_1 = j) = u_1(j)$. Right now, that means that the "initial values" or "initial data" we need to compute all the $u_t(j)$ is actually new information. With this we can complete the path probability computation. If $X = (X_1, X_2, \ldots, X_T)$, it's probability is

$$P(X) = u_1(X_1) \prod_{t=1}^{T-1} P_{X_t, X_{t+1}} \ .$$

3.1.7 Forward equation, matrix version:

In contrast to the matrix version of the backward equation, we let u_t be the row vector $u_t = (u_t(1), \ldots, u_t(s))$. Then the forward equation (3.3) may be expresses as

$$u_{t+1} = u_t P \ , \tag{3.4}$$

where P again is the transition matrix. It may seem odd to express matrix vector multiplication with the vector on the left of the matrix, but it is natural if we think of u_t as a $1 \times s$ matrix. The expression Pu_t is not even compatible for matrix multiplication. As with the backward equation, we can iterate (3.2) to get, for example, $u_t = u_1 P^{t-1}$.

3.1.8 Expectation value:

We combine the conditional expectations $f_t(j)$ defined in paragraph 1.3 with the probabilities $u_t(j)$ above and the law of total probability to get, for any given t,

$$\begin{aligned} E[V(X_T)] &= \sum_{j=1}^{s} P(X_t = j) E[V(X_T) \mid X_t = j] \\ &= \sum_{j=1}^{s} u_t(j) f_t(j) \\ &= u_t f_t \ . \end{aligned}$$

The last line is the matrix product of the row vector u_t, thought of as a $1 \times s$ matrix, with the column vector f_t, thought of as an $s \times 1$ matrix. By the rules of matrix multiplication, the result should be a 1×1 matrix, that is, a number. We will be using this formula and generalizations of it often throughout the course. For now, note the curious fact that although u_t and f_t are different for different t values, the product $u_t f_t$ is not; it is invariant. For this invariance to be possible, the forward evolution for u_t and the backward evolution for f_t must be related.

3.1.9 Relationship between the forward and backward equations:

In fact, if we know that $u_t f_t$ is independent of t, then the backward evolution (3.2) implies the forward evolution (3.4) and vice versa. For example,

$u_{t+1}f_{t+1} = u_t f_t$, together with the backward evolution implies that $u_{t+1}f_{t+1} = u_t P f_{t+1}$. This implies that

$$(u_{t+1} - u_t P)\, f_{t+1} = 0 \; .$$

(Note that we used the associativity, $(AB)C = A(BC)$, of matrix multiplication ($u(Pf) = (uP)f$) and the distributive property. This is why we were eager to express the evolution equations as matrix multiplication and, in particular, to distinguish between row and column vectors.) If this is true for a set of s linearly independent vectors f_{t+1}, then the vector $(u_{t+1} - u_t P)$ must be zero, which is (3.4). A theoretically minded reader can verify that enough f vectors are available if the transition matrix is nonsingular. In the same way, the backward evolution of f is a consequence of invariance and the forward evolution of u.

3.1.10 Duality:

Duality refers to a collection of ideas useful in linear algrbra and its generalizations. In it's simplest form, it is the relationship between a matrix and it's transpose. The set of column vectors with s components is a vector space. The set of s component row vectors is the dual space. We can combine an element of a vector space with an element of its dual to get a number. This is the product of the $1 \times s$ matrix u with the $s \times 1$ matrix f. Any linear transformation on the vector space of column vectors is represented by an $s \times s$ matrix, P. This matrix then defines a linear transformation, the dual transformation, on the dual space of row vectors, given by $u \to uP$. In this sense, the forward and backward equations are dual to each other.

3.1.11 Duality, adjoint, and transpose:

Duality may be related to the matrix transpose operation. If you want to keep all vectors as colums, then the row vector we called u would be called u^* for the column vector u. We denote the transpose of a real matrix by a $*$ so that T or t is not over used. If we think of u as a column vector, then the forward evolution equation (3.4) would be written $u_{t+1} = P^* u_t$. For this reason, the transpose of a matrix is sometimes called its dual. The invatiant quantity would be written $u_t^* f_t$, etc. If we ever meet a matrix, A, with complex entries, A^* will denote the conjugate transpose matrix: flip the matrix and take the complex conjugate of the entries. That matrix is often called the adjoint matrix to A. Warning, the term "adjoint" is often used for the matrix $\det(A)A^{-1}$, whose entries are the determinants of principal minors of A. I will not use adjoint in this sense. Later in the course, the matrix P will be replaced by a "differential operator" that is he "generator" of a kind of Markov process. The adjoint of the generator is another differential operator. Duality will be with us until the end.

3.2 Martingales and stopping times

3.2.1 Stochstic process:

We have a probability space, Ω. The information available at time t is represented by the algebra of events $\mathcal{F}_t$. We assume that for each t, $\mathcal{F}_t \subset \mathcal{F}_{t+1}$; since we are supposed to gain information every known event in $\mathcal{F}_t$ is also known at time $t+1$. A stochastic process is a family of random variables, $X_t(\omega)$, with $X_t \in \mathcal{F}_t$ (reminder, this in an abuse of notation that represents the hypothesis that X_t is measureable with respect to $\mathcal{F}_t$). Sometimes it happens that the random variables X_t contain all the information in the $\mathcal{F}_t$ in the sense that $\mathcal{F}_t$ is generated by X_1, …, X_t. This the "minimal algebra" in which the X_t form a stochastic process. In other cases $\mathcal{F}_t$ contains more information. Economists use these possibilities when they distinguish between the "weak efficient market hypothesis" (the $\mathcal{F}_t$ are minimal), and the "strong hypothesis" ($\mathcal{F}_t$ contains all the information in the world, literally). In the case of minimal $\mathcal{F}_t$, it may be possible to identify the outcome, ω, with the path $X = X_1, \ldots, X_T$. This is not possible when the $\mathcal{F}_t$ are not minimal. For the definition of stochastic process, the actual probabilities are not important, just the algebras of sets and "random" variables X_t.

3.2.2 Example 1, Markov chains:

In this example, the $\mathcal{F}_t$ are minimal and Ω is the path space of sequences of length T from the state space, $\mathcal{S}$. The variables X_t are may be called "coordinate functions" because X_t is coordinate t (or entry t) in the sequence X. In principle, we could express this with the notation $X_t(X)$, but that would drive people crazy. Although we distinguish between Markov chains (discrete time) and Markov processes (continuous time), the term "stochastic process" can refer to either continuous or discrete time.

3.2.3 Example 2, diadic sets:

This is a set of definitions for discussing averages over a range of length scales. The "time" variable, t, represents the amount of averaging that has been done. At the "first" time, $t = 1$, we have only the overall average. At "later" times, we have averages over smaller and smaller sets. Only at the final time, T, is the original random variable completely known. To go from time $t+1$ to time t, we combine two level $t+1$ averages to produce a coarser level t average. The actual averaging process is discussed below. Here we only define the sets being averaged over. The coming definitions would be simpler if time and "space" variables were to start with 0 rather than 1. I've chosen to start always with 1 to be consistent with notations used above and below. The whole space, Ω, consists of 2^{T-1} objects, which we call 1, …, 2^{T-1} (It would be 2^t if we were to start with $t = 0$ rather than $t = 1$.). The partition defining $\mathcal{F}_t$ is given by "diadic" sets with 2^{T-t} consecutive elements each, called $B_{t,k}$ for $k =$

$1, \ldots, 2^{t-1}$. At time $t = 1$ there is just one B, which is the whole of Ω. At time $t = 1$, there are two, $B_{2,1} = \{1, \ldots, 2^{T-2}\}$, and $B_{2,2} = \{2^{T-2}+1, \ldots, 2^{T-1}\}$. At time $T-1$ there are $|\Omega|/2 = 2^{T-2}$ diadic sets with two elements each: $B_{T-1,1} = \{1,2\}$, $B_{T-1,2} = \{3,4\}$, $\ldots$, $B_{T-1,2^{T-2}} = \{2^{T-1}-1, 2^{T-1}\}$. At level $T-2$, the partition sets $B_{T-2,k}$ contain 4 consecutive elements each. In general, $B_{t,k} = \{(k-1)2^{T-t}+1, \ldots, k2^{T-t}\}$. The reader should check in detail that the general definition agrees with the cases $t = 1, 2, T-2, T-1$, and T. The diadic property is that each level t set is the uninion of two consecutive level $t+1$ sets: $B_{t,k} = B_{t+1,2k-1} \cap B_{t+1,2k}$.

For now, we will take the define the X_t by $X_t(\omega) = k$ if $\omega \in Bt, k$. For example, this gives $X_T(\omega) = \omega$, $X_1(\omega) = 1$ for all $\omega \in \Omega$, and, in general, $X_t(\omega) = \text{int}(\omega/2^{??})$, where $\text{int}(a)$ is the largest integer not exceeding a.

3.2.4 Martingales:

A real valued stochastic process, X_t, is a martingale if

$$E[X_{t+1} \mid \mathcal{F}_t] = X_t \ .$$

If we take the overall expectation of both sides we see that the expectation value does not depend on t, $E[X_{t+1}] = E[X_t]$. The martingale property says more. Whatever information you might have at time t notwithstanding, still the expectation of future values is the present value. There is a gambling interpretation: X_t is the amount of money you have at time t. No matter what has happened, your expected winnings at between t and $t+1$, the "martingale difference" $Y_{t+1} = X_{t+1} - X_t$, has zero expected value. You can also think of martingale differences as a generalization of independent random variables. If the random variables Y_t were actually independent, then the sums $X_t = \sum_{k=1}^{t} Y_t$ would form a martingale (using the $\mathcal{F}_{\sqcup}$, generated by the Y_1, $\ldots$, Y_t). The reader should check this.

3.2.5 A lemma on conditional expectation:

In working with martingales we often make use of a basic lemma about conditional expectation. Suppose $U(\omega)$ and $V(\omega)$ are real valued random variables and that $V \in \mathcal{F}$. Then

$$E[VU \mid \mathcal{F}] = VE[U \mid \mathcal{F}] \ .$$

This is easy to see in the classical definition of conditional expectation. Suppose B is one of the sets in the partition defining $\mathcal{F}$ and that $W = E[U(\omega) \mid \omega \in B]$. We know that $V(\omega)$ is constant in B because $V \in \mathcal{F}$. Call this value v. Then $E[VU \mid B] = vE[U \mid B] = vW$. This shows that no matter which partition set ω falls in, $E[VU \mid B] = VE[U \mid B]$, which is exactly the (classical version of) the lemma.

3.2.6 More martingales:

This lemma leads to lots of martingales. Suppose the "multipliers" M_t are functions of Y_1, …, Y_{t-1} (leaving out Y_t), then the sums $X_t = \sum_{k=1}^{t} M_t Y_t$ also form a martingale if the Y_t have mean value zero. Let us check this. In the algebra $\mathcal{F}_t$ we know the values of all the Y_k for $1 \leq k \leq t$. Therefore, we know the value of M_{t+1}, which is to say that $M_{t+1} \in \mathcal{F}_t$. This shows that

$$E[X_{t+1} \mid \mathcal{F}_t] = X_t + M_{t+1} E[Y_{t+1} \mid \mathcal{F}_t] = X_t \ .$$

At the end we used the fact that $E[Y_{t+1}] = 0$, and that Y_{t+1} is independent of all the earlier Y_k which generarate $\mathcal{F}_t$. This is a simple generalization of summing independent mean zero random variables. Even though the martingale differences $X_{t+1} - X_t = M_{t+1} Y_{t+1}$ are not independent, they still have mean value zero, conditioned on $\mathcal{F}_t$.

3.2.7 Weak and strong efficient markets:

It is possible that the family of random variables X_t might or might not form a martingale depending on what increasing family of algebras you use. For example, suppose X_t is a stochastic process with respect to the algebras $\mathcal{F}_t$ and form a martingale with respect to them. Now suppose $\mathcal{G}_t$ is the algebra generated by X_1, …, X_{t+1}. Clearly, $E[X_{t+1} \mid \mathcal{G}_t] = X_{t+1} \neq X_t$. The X_t form a martingale with respect to the $\mathcal{F}_t$ but not with respect to the additional information in $\mathcal{G}_t$.

3.2.8 Doob's principle:

Notice what happened here. We started with a simple martingale that was built of the sum of independent mean zero random variables. Then we built a more complex stochastic process, $X_{t+1} = X_t + M_{t+1} Y_{t+1}$, where the value of M_{t+1} is known at time t. One can think of this as building an investment strategy; collecting information by watching the market up to time t then placing a "bet" of size M on the still unknown random variable Y_{t+1}. No matter how this is done, the result is still a martingale. This is a general feature of martingales: any betting strategy that at time t uses only $\mathcal{F}_t$ information produces another martingale. Other instances of this principle are formulated below. This "Doob's principle", named for the probabilist who formulated it, is one of the things that makes martingales handy.

3.2.9 Example, conditional expectations:

Suppose $\mathcal{F}_t$ is any expanding family of algebras and V is any random variable. (We are allowed to say "any", with no technical hypotheses, because Ω is finite. This luxury does not last forever.) The conditional expectations $F_t = E[V \mid \mathcal{F}_t]$ form a martingale. This is a consequence of the rules of iterated conditional expectation, lecture 1, paragraph 1.??. In particular, if X_t are the states of a Markov chain, then the random variables

$$F_t = f_t(X_t) = E[V(X_T) \mid \mathcal{F}_t]$$

form a martingale.

3.2.10 Example 2, continued:

Suppose we have a function $V(\omega)$ defined for integers ω in the range $1 \leq \omega \leq 2^{T-1}$. Suppose that we specify uniform probabilities, $P(\omega) = 2^{-T+1}$, for all ω. Then the conditional expectations that are the values of F_t are averages of V over dyadic blocks of size 2^{T-t}. The random variable F_1 is just the average of V. Next, $F_2(\omega)$ equals the average over the first half if ω is in the first half and over the second half if ω is in the second half. The graph of F_1 is just a constant while the graph of F_2 is two constants separated by a step at the midpoint. The graph of F_2 is 4 constants with 3 steps, and so on. If we plot all these graphs together, we get a better and better picture of the graph of the original function, V. You could do the same with a two dimensional function given by an image. What this looks like can be seen on the class bboard.

3.2.11 Doob's principle continued:

Suppose F_t is any martingale with martingale differences $Y_t = F_t - F_{t-1}$, and that $M_t \in \mathcal{F}_t$. Then the modified stochastic process G_t defined by

$$G_{t+1} = G_t + M_t Y_{t+1}$$

is also a martingale. This follows as before: $E[G_{t+1} - G_t \mid \mathcal{F}_t]$ is just $E[M_t Y_{t+1} \mid \mathcal{F}_t]$ which vanishes because $M_t \in \mathcal{F}_f$ and $E[Y_{t+1} \mid \mathcal{F}_t] = 0$.

3.2.12 Investing with Doob:

Economists sometimes use this to make a point about active trading in the stock market. Suppose that F_t, the price of a stock at time t is a martingale. Suppose that at time t we look at the entire history of F from time 1 to t an decide an amount M_t to invest at time t. The change in our "portfolio" (shares in 1 stock and cash) value by time $t+1$ will be $M_t(F_{t+1} - F_t) = M_t Y_{t+1}$. The portfolio value at time t will be G_t. The fact that the values G_t also form a martingale is said to show that active investing is no better than a "buy and hold" strategy that just produces the value F_t, or a multiple of it depending on how much you invest. The well known book **A Random Walk on Wall Street** is mostly an exposition of this point of view. The fallacy is that investors are not only interested in the expected value, but also in the risk.

3.2.13 Stopping times:

We have Ω and the expanding family $\mathcal{F}_t$. A stopping time is a function $\tau(\omega)$ that is one of the times 1, …, T, so that the event $\{\tau \leq t\}$ is in $\mathcal{F}_t$. Stopping times might be thought of as possible strategies. Whatever your criterion for stopping is, you have enough information at time t to know whether you should stop at time t. Many stopping times are expressed as the first time something happens, such as the first time $X_t > a$. We cannot ask to stop, for example, at the last t with $X_t > a$ because we might not know at time t whether $X_{t'} > a$ for some $t' > t$.

3.2.14 Doob's stopping time theorem for one stopping time:

Because stopping times are nonanticipating strategies, they also cannot make money from a martingale. One version of this statement is that $E[X_\tau] = E[X_1]$. The proof of this makes use of the events B_t, that $\tau = t$. The stopping time hypothesis is that $B_t \in \mathcal{F}_t$. Since τ has some value $1 \leq \tau \leq T$, the B_t form a partition of Ω. Also, if $\omega \in B_t$, $\tau(\omega) = t$, so $X_\tau = X_t$. Therefore,

$$\begin{aligned}
E[X_1] &= E[X_T] \\
&= \sum_{t=1}^{T} E[X_T \mid B_t] P(B_t) \\
&= \sum_{t=1}^{T} E[X_\tau] P(\tau = t) \\
&= E[X_\tau] \; .
\end{aligned}$$

In this derivation we made use of the classical statement of the martingale property, if $B \in \mathcal{F}_\sqcup$ then $E[X_T \mid B] = E[X_t \mid B]$. In our $B = B_t$, $X_t = X_\tau$.

This simple idea, using the martingale property applied to the partition B_t, is crucial for much of the theory of martingales. The idea itself was first used Kolmogorov in the context of random walk or Brownian motion. Doob realized that Kolmogorov's was even simpler and more beautiful when applied to martingales.

3.2.15 Stopping time paradox:

The technical hypotheses above, finite state space, bounded stopping times, may be too strong, but they cannont be completely ignored, as this famous example shows. Let X_t be a symmetric random walk starting at zero. This forms a martingale, so $E[X_\tau] = 0$ for any stopping time, τ. On the other hand, suppose we take $\tau = \min(t \mid X_t = 1)$. Then $X_\tau = 1$ always, so $E[X_\tau] = 1$. The catch is that there is no T with $\tau(\omega) \leq T$ for all ω. Even though $\tau < \infty$ "almost surely" (more to come on that expression), $E[\tau] = \infty$ (explination later). Even that would be OK if the possible values of X_t were bounded. Suppose you choose T and set $\tau' = min(\tau, T)$. That is, you wait until $X_t = 1$ or $t = T$, whichever comes first, to stop. For large T, it is very likely that you stopped for $X_t = 1$. Sill, those paths that never reached 1 probably drifted just far enough in the negative direction so that their contribution to the overall expected value cancels the 1 to yield $E[X_{\tau'}] = 0$.

3.2.16 More stopping times theorem:

Suppose we have an increasing family of stopping times, $1 \leq \tau_1 \leq \tau_2 \cdots$. In a natural way the random variables $Y_1 = X_{\tau_1}$, $Y_2 = X_{\tau_2}$, etc. also form a martingale. This is a final elaborate way of saying that strategizing on a martingale is a no win game.

Chapter 4

Lecture 4

4.1 Continuous probability

This section is a quick and sketchy introduction to the modern terminology of probability following Kolmogorov in what we call continuous spaces. Although the modern approach has lots of baggage, it ultimately makes things easier, as we will begin to see here.

4.1.1 Continuous spaces

I use this to mean probability spaces that are not countable (discrete). In discrete probability, we first defined $P(\omega)$, the probability of any particular outcome. Then the probability of an event, A was the sum of the probabilities of the outcomes that make up that event:

$$P(A) = \sum_{\omega \in A} P(\omega) \; . \tag{4.1}$$

In continuous probability, the rule (though there are exceptions), is that the probability of any particular outcome is zero. Also, there are uncountably many outcomes in a typical event. Both of these make (4.1) inapplicable. We do not know how to sum uncountable many numbers, and, we might expect such a sum rule to give the answer zero if all the terms in the sum were zero.

Examples of continuous probability spaces:

R, the real numbers. If ω is a real number and $u(x)$ is a probability density, then the probability of a small interval $(\omega - \epsilon, \omega + \epsilon)$ containing ω is (with an abuse of notation)

$$P(\omega - \epsilon, \omega + \epsilon) = \int_{\omega - \epsilon}^{\omega + \epsilon} u(x) dx \to 0 \;\text{ as } \epsilon \to 0.$$

Thus the probability of ω itself should naturally be zero.

R^n, sequences of n numbers (possibly viewed as a row or column vector depending on the context): $X = (x_1 \ldots , X_n)$.

$\mathcal{S}^{\mathcal{N}}$. Here $\mathcal{S}$ is the state space for a Markov chain (might be finite or countable) and $\mathcal{N}$ is the "natural" numbers, $1, 2, 3, \ldots$. An element is an infinite sequence of elements of $\mathcal{S}$: $X = (X_1, X_2, \ldots)$. Generally, the probability of any particular infinite sequence is zero. For example, if we have a two state Markov chain with transition matrix $\begin{pmatrix} .6 & .4 \\ .3 & .7 \end{pmatrix}$. If we call the states U and D, then the probability of the infinite string $UUU \cdots$ should be $u(U) \cdot .6 \cdot .6 \cdot \cdots = 0$: multiplying together infinitely many .6 numbers converges to zero.

$C([0, T] \to R)$, the path space for Brownian motion. The C stands for "continuous". The $[0, T]$ is the time interval $0 \leq t \leq T$; the square brackets tell us to include the endpoints (0 and T in this case). Round parentheses $(0, T)$ would mean to leave out 0 and T. The final R is the "target" space, the real numbers in this case. An element of Ω is a continuous function from the interval $[0, T]$ to R. If we call this function X_t for $0 \leq t \leq T$, X_t is a real number for each $t \in [0, T]$ and X is a continuous function of t.

4.1.2 Probability measures:

We want to define the probabilities of events $A \subset \Omega$. Since we cannot base these on the probabilities of the individual outcomes in A, we just assume the probabilities are defined for events. For this we first define $\sigma-$algebra. An algebra of events is a $\sigma-$algebra if, for any sequence of events $A_n \in \Omega$, the union union $\cup_{n=1}^{\infty} A_n$ is also an event in $\mathcal{F}$. Suppose $\mathcal{F}$ is a $\sigma-$algebra of events in Ω. The numbers $P(A)$ for $A \in \mathcal{F}$ are a "probability measure" if

i. If $A \in \mathcal{F}$ and $B \in \mathcal{F}$ are disjoint events, then $P(A \cup B) = P(A) + P(B)$.

ii. $P(A) \geq 0$ for any event $A \in \mathcal{F}$.

iii. $P(\Omega) = 1$.

iv. If $A_n \in \mathcal{F}$ is a sequence of events each disjoint from all the others and $\cup_{n=1}^{\infty} A_n = A$, then $\sum_{n=1}^{\infty} P(A_n) = P(A)$.

The last property is called "countable additivity". All the probability measures we deal with in this course are countably additive.

4.1.3 R^n:

A "ball" in n dimensional space is any of the sets $B_r(x) = \{y \mid |x - y| < r$. This might be called an interval in one dimension and a disk in two, but the term ball applies to any dimension, including 1 and 2. With $|x - y| \leq r$, we would have a "closed" ball, as opposed to the "open" ball above. This makes no difference here. In fact, a $\sigma-$algebra that contains all open balls also contains all closed balls, and any set in R^n you can describe without advanced mathematical analysis. The $\sigma-$algebra generated by open balls is called the Borel algebra, and events measurable in this algebra are called Borel sets. A function $u(x)$ is a probability density if it is never negative and $\int_{R^n} u(x)dx = 1$. Such a probability density defines a probability measure on the Borel algebra by

$$P(A) = \int_A u(x)dx \ .$$

It is can be shown that if u is measurable with respect to the Borel sets then this probability measure is countable additive.

4.1.4 Integration with respect to a measure:

The definition of integration with respect to a general probability measure is easier than the definition of the Riemann integral. Let Ω be a probability space, $\mathcal{F}$ a $\sigma-$algebra of events, and P a probability measure. A function $f(\omega)$ is measurable with respect to $\mathcal{F}$ if all of the events $A_{ab} = \{a \leq f \leq b\}$ $= \{\omega \mid a \leq f(\omega) \leq b\}$ are in $\mathcal{F}$. Because $\mathcal{F}$ is an algebra, the condition $a \leq f$ can be replaced by $a < f$, etc. Any function on R^n (i.e. any function of n real variables), no matter how many weird discontinuities you try to throw in, will be measurable with respect to the Borel algebra, unless you know serious advanced analysis. It happens in general that a function may fail to be measurable with respect to some $\mathcal{F}$, but this will always (in this course) be due to a lack of information (small $\mathcal{F}$) rather than discontinuities in u.

The integral is written

$$E[f] = \int_{\omega \in \Omega} f(\omega)dP(\omega) \ .$$

In R^n with a density u, this agrees with teh classical definition

$$E[f] = \int_{R^n} f(x)u(x)dx \ .$$

Note that the abstract variable ω is replaced by the concrete variable, x, in this more concrete situation. The general definition is forced on us once we make the natural requirements

i. If $A \in \mathcal{F}$ is any event, then $E[1_A] = P(A)$. The integral of the indicator function if an event is the probability of that event.

ii. If f_1 and f_2 have $f_1(\omega) \leq f_1(\omega)$ for all $\omega \in \Omega$, then $E[f_1] \leq E[f_2]$. "Integration is monotone".

iii. For any reasonable functions f_1 and f_2 (e.g. bounded), we have $E[af_1 + bf_2] = aE[f_1] + bE[f_2]$. "Integration is linear".

Now suppose f is a nonnegative bounded function: $0 \leq f(\omega) \leq M$ for all $\omega \in \Omega$. The integral of f is determined by the three properties above. Choose a small number ϵ and define the "ring sets" $A_n = \{(n-1)\epsilon \leq f < n\epsilon$. The A_n depend on ϵ but we do not indicate that. Although the events A_n might be complicated, fractal, or whatever, Each of them is measurable. The "step function" $g(\omega) = \sum_n (n-1)\epsilon 1_{A_n}$ takes the value $(n-1)\epsilon$ on each of the sets A_n (each ω is in only one A_n. For any ω, only one of the terms in the sum is different from zero.). The sum defining g is finite because f is bounded, though the number of terms is M/ϵ. Also, $g(\omega) \leq f(\omega)$ for each $\omega \in \Omega$ (though by at most ϵ). Therefore, the three properties of integration imply that

$$E[f] \geq E[g] = \sum_n (n-1)\epsilon E[A_n] = \sum_n (n-1)\epsilon P((n-1)\epsilon \leq f < n\epsilon) \; .$$

In the same way, we can consider the upper function $h = \sum_n n\epsilon 1_{A-n}$ and have

$$E[f] \leq E[h] = \sum_n n\epsilon E[A_n] = \sum_n n\epsilon P((n-1)\epsilon \leq f < n\epsilon) \; .$$

If you draw a picture of this situation for $\Omega = R$, you will see the lower (g) and upper (h) step functions bracketing f. When you replace ϵ by $\epsilon/2$, the lower step goes up and the upper step goes down. This gives a sequence of approximations $G(\epsilon) \leq E[f] \leq H(\epsilon)$ with $G(\epsilon)$ increasing and $H(\epsilon)$ decreasing as $\epsilon \to 0$. Finally, note that $H(\epsilon) - G(\epsilon) \leq \epsilon$, because that is how close the upper and lower step approximations h and g are. Thus, as $\epsilon \to 0$, the upper and lower approximations converge to the same number, which must be $E[f]$. It is sometimes said that the difference between classical (Riemann) integration and modern integration (here) is that we used to cut the x axis into little pieces, but it is simpler to cut the y axis instead.

If the function f is positive but not bounded, it might happen that $E[f] = \infty$. The "cut off" functions, $f_M(\omega) = \min(f(\omega), M)$, might have $E[f_M] \to \infty$ as $M \to \infty$. If f is both positive and negative (for different ω), we integrate the positive part, $f_+(\omega) = \max(f(\omega), 0)$, and the negative part $f_-(\omega) = \min(f(\omega), 0$ separately and subtract the results. We do not attempt a definition if $E[f_+] = \infty$ and $E[f_-] = -\infty$.

4.1.5 Markov chains with $T = \infty$:

The probability space, Ω, is the set of all infinite sequences $X = (X_1, X_2, \ldots)$, where each X_t is one of the states in the state space $\mathcal{S}$.

Just as the Borel algebra of sets can be generated by balls, the algebra of sets here can be generated by "cylinder" sets (don't ask me how they got that name). For each sequence of length L, $x = (x_1, \ldots, x_l)$, there is a cylinder set $B_x = \{X \mid X_1 = x_1, \ldots, X_L = x_L\}$. Other sets can be made from countable set operations starting with these. For example, the event containing the single sequence $UUU \cdots$ is the intersection of the events having the first L entries U. In a slightly more complicated way, it is possible to express the event "the first $UUDDU$ occurs before the first $DDUD$" in terms of cylinder sets. The probabilities $P(B_x) = u_1(X_1)\prod_{t=1}^{L-1} P_{x_t,x_{t+1}}$ give rise to a probability measure that is countably additive on this $\sigma-$algebra, another theorem of Kolmogorov.

4.1.6 Conditional expectation:

We have a random variable $X(\omega)$ that is measurable with respect to the $\sigma-$algebra, $\mathcal{F}$, and a subalgebra $\mathcal{G} \subset \mathcal{F}$. We want to define the conditional expectation $Y = E[X \mid \mathcal{G}]$. When Ω is finite we can define $Y(\omega)$ be knowing which partition block ω is in. In continuous probability, a subalgebra might or might not be generated by a partition (I don't know), but even if it were, the sets in the partition would usually have probability zero so Bayes' rule would not be applicable. For example, suppose we have a two dimensional random variable $X = (X_1, X_2)$ with a density $u(x_1, x_2)$ and we want $P(X_1 > 3 \mid X_2 = 0)$. The event $B = \{X_2 = 0\}$ has probability $P(B) = 0$. There is a "classical" definition of conditional expectation for this case, but the one "modern" definition works for all cases. The definition is that $Y(\omega)$ is the random variable measurable with respect to $\mathcal{G}$ that best approximates X in the least squares sense

$$E[(Y-X)^2] = \min Z \in \mathcal{G} E[(Z-X)^2] \; .$$

This is one of the definitions we gave before, the one that works for continuous and discrete probability. In the theory, it is possible to show that there is a minimizer and that it is unique.

4.1.7 Generating a $\sigma-$algebra:

When the probability space, Ω, is finite, we can understand an algebra of sets by using the partition of Ω that generates the algebra. This is not possible for continuous probability spaces. Another way to specify an algebra for finite Ω was to give a function $X(\omega$, or a collection of functions $X_k(\omega)$ that are supposed to be measurable with respect to $\mathcal{F}$. We noted that any function measurable with respect to the algebra generated by functions X_k is actually a function of the X_k. That is, if $F \in \mathcal{F}$ (abuse of notation), then there is some function $u(x_1, \ldots, x_n)$ so that

$$F(\omega) = u(X_1(\omega), \ldots, X_n(\omega)) \; . \tag{4.2}$$

The intuition was that $\mathcal{F}$ contains the information you get by knowing the values of the functions X_k. Any function measurable with respect to this algebra is determined by knowing the values of these functions, which is precisely what (4.2) says. This approach using functions is often convenient in continuous probability.

If Ω is a continuous probability space, we may again specify functions X_k that we want to be measurable. Again, these functions generate an algebra, a $\sigma-$algebra, $\mathcal{F}$. If F is measurable with respect to this algebra then there is a (Borel measurable) function $u(x_1,\ldots)$ so that $F(\omega) = u(X_1,\ldots)$, as before. In fact, it is possible to define $\mathcal{F}$ in this way. Saying that $A \in \mathcal{F}$ is the same as saying that $\mathbf{1}_A$ is measurable with respect to $\mathcal{F}$. If $u(x_1,\ldots)$ is a Borel measurable function that takes values only 0 or 1, then the function F defined by (4.2) defines a function that also takes only 0 or 1. The event $A = \{\omega \mid F(\omega) = 1$ has (obviously) $F = \mathbf{1}_A$. The $\sigma-$algebra generated by the X_k is the set of events that may be defined in this way. A complete proof of this would take a few pages.

4.1.8 Example in two dimensions:

Suppose Ω is the unit square in two dimensions: $(x,y) \in \Omega$ if $0 \leq x \leq 1$ and $0 \leq y \leq 1$. The "x coordinate function" is $X(x,y) = x$. The information in this is the value of the x coordinate, but not the y coordinate. An event measurable with respect to this $\mathcal{F}$ will be any event determined by the x coordinate alone. I call such sets "bar code" sets. You can see why by drawing some.

4.1.9 Marginal density and total probability:

The abstract situation is that we have a probability space, Ω with generic outcome $\omega \in \Omega$. We have some functions $(X_1(\omega),\ldots,X_n(\omega)) = X(\omega)$. With Ω in the background, we can ask for the joint PDF of $(X_1,\ldots,X_n)$, written $u(x_1,\ldots,x_n)$. A formal definition of u would be that if $A \subseteq R^n$, then

$$P(X(\omega) \in A) = \int_{x \in A} u(x)dx \ . \tag{4.3}$$

Suppose we neglect the last variable, X_n, and consider the reduced vector $\tilde{X}(\omega) = (X_1,\ldots,X_{n-1})$ with probability density $\tilde{u}(x_1,\ldots,x_{n-1})$. This $\tilde{u}$ is the "marginal density" and is given by integrating u over the forgotten variable:

$$\tilde{u}(x_1,\ldots,x_{n_1}) = \int_{-\infty}^{\infty} u(x_1,\ldots,x_n)dx_n \ . \tag{4.4}$$

This is a continuous probability analogue of the law of total probability: integrate (or sum) over a complete set of possibilities, all values of x_n in this case.

We can prove (4.4) from (4.3) by considering a set $B \subseteq R^{n-1}$ and the corresponding set $A \subseteq R^n$ given by $A = B \times R$ (i.e. A is the set of all pairs $\tilde{x}, x_n)$ with $\tilde{x} = (x_1,\ldots,x_{n-1}) \in B$). The definition of A from B is designed so that

$P(X \in A) = P(\tilde{X} \in B)$. With this notation,

$$\begin{aligned} P(\tilde{X} \in B) &= P(X \in A) \\ &= \int_A u(x)dx \\ &= \int_{\tilde{x} \in B} \int_{x_n=-\infty}^{\infty} u(\tilde{x}, x_n)dx_n d\tilde{x} \\ P(\tilde{X} \in B) &= \int_B \tilde{u}(\tilde{x})d\tilde{x} \ . \end{aligned}$$

This is exactly what it means for $\tilde{u}$ to be the PDF for $\tilde{X}$.

4.1.10 Classical conditional expectation:

Again in the abstract setting $\omega \in \Omega$, suppose we have random variables $(X_1(\omega), \ldots, X_n(\omega))$. Now consider a function $f(x_1, \ldots, x_n)$, its expected value $E[f(X)]$, and the conditional expectations

$$v(x_n) = E[f(X) \mid X_n = x_n] \ .$$

The Bayes' rule definition of $v(x_n)$ has some trouble because both the denominator, $P(X_n = x_n)$, and the numerator,

$$E[f(X) \cdot \mathbf{1}_{X_n = x_n}] \ ,$$

are zero.

The classical solution to this problem is to replace the exact condition $X_n = x_n$ with an approximate condition having positive (though small) probability: $x_n \leq X_n \leq x_n + \epsilon$. We use the approximation

$$\int_{x_n}^{x_n+\epsilon} g(\tilde{x}, \xi_n)d\xi_n \approx \epsilon g(\tilde{x}, x_n) \ .$$

The error is roughly proportional to ϵ^2 and much smaller than either the terms above. With this approximation the numerator in Bayes' rule is

$$\begin{aligned} E[f(X) \cdot \mathbf{1}_{x_n \leq X_n \leq x_n+\epsilon}] &= \int_{\tilde{x} \in R^{n-1}} \int_{\xi_n = x_n}^{\xi_n = x_n+\epsilon} f(\tilde{x}, \xi_n)u(\tilde{x}, x_n)d\xi_n d\tilde{x} \\ &\approx \epsilon \int_{\tilde{x}} f(\tilde{x}, x_n)u(\tilde{x}, x_n)d\tilde{x} \ . \end{aligned}$$

Similarly, the denominator is

$$P(x_n \leq X_n \leq x_n + \epsilon) \approx \epsilon \int_{\tilde{x}} u(\tilde{x}, x_n)d\tilde{x} \ .$$

If we take the Bayes' rule quotient and let $\epsilon \to 0$, we get the classical formula

$$E[f(X) \mid X_n = x_n] = \frac{\int_{\tilde{x}} f(\tilde{x}, x_n) u(\tilde{x}, x_n) d\tilde{x}}{\int_{\tilde{x}} u(\tilde{x}, x_n) d\tilde{x}} \quad . \tag{4.5}$$

By taking f to be the characteristic function of an event (all possible events) we get a formula for the probability density of $\tilde{X}$ given that $X_n = x_n$, namely

$$\tilde{u}(\tilde{x} \mid X_n = x_n) = \frac{u(\tilde{x}, x_n)}{\int_{\tilde{x}} u(\tilde{x}, x_n) d\tilde{x}} \quad . \tag{4.6}$$

This is the classical formula for conditional probability density. The integral in the denominator insures that, for each x_n, $\tilde{u}$ is a probability density as a function of $\tilde{x}$, that is

$$\int \tilde{u}(\tilde{x} \mid X_n = x_n) d\tilde{x} = 1 \; ,$$

for any value of x_n. It is very useful to notice that as a function of $\tilde{x}$, u and $\tilde{u}$ almost the same. They differ only by a constant normalization. For example, this is why conditioning Gaussian's gives Gaussians.

4.1.11 Modern conditional expectation:

The classical conditional expectation (4.5) and conditional probability (4.6) formulas are the same as what comes from the "modern" definition from paragraph 1.6. Suppose $X = (X_1, \ldots, X_n)$ has density $u(x)$, $\mathcal{F}$ is the $\sigma-$algebra of Borel sets, and $\mathcal{G}$ is the $\sigma-$algebra generated by X_n (which might be written $X_n(X)$, thinking of X as ω in the abstract notation). For any $f(x)$, we have $\tilde{f}(x_n) = E[f \mid \mathcal{G}]$. Since $\mathcal{G}$ is generated by X_n, the function $\tilde{f}$ being measurable with respect to $\mathcal{G}$ is the same as it's being a function of x_n. The modern definition of $\tilde{f}(x_n)$ is that it minimizes

$$\int_{R^n} \left(f(x) - \tilde{f}(x_n) \right)^2 u(x) dx \; , \tag{4.7}$$

over all functions that depend only on x_n (measurable in $\mathcal{G}$).

To see the formula (4.5) emerge, again write $x = (\tilde{x}, x_n)$, so that $f(x) = f(\tilde{x}, x_n)$, and $u(x) = u(\tilde{x}, x_n)$. The integral (4.7) is then

$$\int_{x_n=-\infty}^{\infty} \int_{\tilde{x} \in R^{n-1}} \left(f(\tilde{x}, x_n) - \tilde{f}(x_n) \right)^2 u(\tilde{x}, x_n) d\tilde{x} dx_n \; .$$

In the inner integral:

$$R(x_n) = \int_{\tilde{x} \in R^{n-1}} \left(f(\tilde{x}, x_n) - \tilde{f}(x_n) \right)^2 u(\tilde{x}, x_n) d\tilde{x} \; ,$$

$\tilde{f}(x_n)$ is just a constant. We find the value of $\tilde{f}(x_n)$ that minimizes $R(x_n)$ by minimizing the quantity

$$\int_{\tilde{x}\in R^{n-1}} \left(f(\tilde{x},x_n)-g\right)^2 u(\tilde{x},x_n)d\tilde{x} =$$
$$\int f(\tilde{x})^2 u(\tilde{x},x_n)d\tilde{x} + 2g\int f(\tilde{x})u(\tilde{x},x_n)d\tilde{x} + g^2\int u(\tilde{x},x_n)d\tilde{x} \ .$$

The optimal g is given by the classical formula (4.5).

4.1.12 Modern conditional probability:

We already saw that the modern approach to conditional probability for $\mathcal{G} \subset \mathcal{F}$ is through conditional expectation. In its most general form, for every (or almost every) $\omega \in \Omega$, there should be a probability measure P_ω on Ω so that the mapping $\omega \to P_\omega$ is measurable with respect to $\mathcal{G}$. The measurability condition probably means that for every event $A \in \mathcal{F}$ the function $p_A(\omega) = P_\omega(A)$ is a $\mathcal{G}$ measurable function of ω. In terms of these measures, the conditional expectation $\tilde{f} = E[f \mid \mathcal{G}]$ would be $\tilde{f}(\omega) = E_\omega[f]$. Here E_ω means the expected value using the probability measure P_ω. There are many such subscripted expectations coming.

A subtle point here is that the conditional probability measures are defined on the original probability space, Ω. This forces the measures to "live" on tiny (generally measure zero) subsets of Ω. For example, if $\Omega = R^n$ and $\mathcal{G}$ is generated by x_n, then the conditional expectation value $\tilde{f}(x_n)$ is an average of f (using density u) only over the hyperplane $X_n = x_n$. Thus, the conditional probability measures P_X depend only on x_n, leading us to write P_{x_n}. Since $\tilde{f}(x_n) = \int f(x)dP_{x_n}(x)$, and $\tilde{f}(x_n)$ depends only on values of $f(\tilde{x},x_n)$ with the last coordinate fixed, the measure dP_{x_n} is some kind of δ measure on that hyperplane. This point of view is useful in many advanced problems, but we will not need it in this course (I sincerely hope).

4.1.13 Semimodern conditional probability:

Here is an intermediate "semimodern" version of conditional probability density. We have $\Omega = R^n$, and $\tilde{\Omega} = R^{n-1}$ with elements $\tilde{x} = (x_1,\ldots,x_{n-1})$. For each x_n, there will be a (conditional) probability density function $\tilde{u}_{x_n}$. Saying that $\tilde{u}$ depends only on x_n is the same as saying that the function $x \to \tilde{u}_{x_n}$ is measurable with respect to $\mathcal{G}$. The conditional expectation formula (4.5) may be written

$$E[f \mid \mathcal{G}](x_n) = \int_{R^{n-1}} f(\tilde{x},x_n)\tilde{u}_{x_n}(\tilde{x})d\tilde{x} \ .$$

In other words, the classical $u(\tilde{x} \mid X_n = x_n)$ of (4.6) is the same as the semimodern $\tilde{u}_{x_n}(\tilde{x})$.

4.2 Gaussian Random Variables

The central limit theorem (CLT) makes Gaussian random variables important. A generalization of the CLT is Donsker's "invariance principle" that gives Brownian motion as a limit of random walk. In many ways Brownian motion is a multivariate Gaussian random variable. We review multivariate normal random variables and the corresponding linear algebra as a prelude to Brownian motion.

4.2.1 Gaussian random variables, scalar:

The one dimensional "standard normal", or Gaussian, random variable is a scalar with probability density

$$u(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \; .$$

The normalization factor $\frac{1}{\sqrt{2\pi}}$ makes $\int_{-\infty}^{\infty} u(x)dx = 0$ (a famous fact). The mean value is $E[X] = 0$ (the integrand $xe^{-x^2/2}$ is antisymmetric about $x = 0$). The variance is (using integration by parts)

$$\begin{aligned}
E[X^2] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \left(xe^{-x^2/2} \right) dx \\
&= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \left(\frac{d}{dx} e^{-x^2/2} \right) dx \\
&= -\frac{1}{\sqrt{2\pi}} \left(xe^{-x^2/2} \right) \Big|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx \\
&= 0 + 1
\end{aligned}$$

Similar calculations give $E[X^4] = 3$, $E[X^6] = 15$, and so on. I will often write Z for a standard normal random variable. A one dimensional Gaussian random variable with mean $E[X] = \mu$ and variance $\text{var}(X) = E[(X-\mu)^2] = \sigma^2$ has density

$$u(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \; .$$

It is often more convenient to think of Z as the random variable (like ω) and write $X = \mu + \sigma Z$. We write $X \sim \mathcal{N}(\mu, \sigma^2)$ to express the fact that X is normal (Gaussian) with mean μ and variance σ^2. The standard normal random variable is $Z \sim \mathcal{N}(0, 1)$

4.2.2 Multivariate normal random variables:

The $n \times n$ matrix, H, is positive definite if $x^* H x > 0$ for any n component column vector $x \neq 0$. It is symmetric if $H^* = H$. A symmetric matrix is positive definite if and only if all its eigenvalues are positive. Since the inverse of a symmetric matrix is symmetric, the inverse of a symmetric positive definite (SPD) matrix is also SPD. An n component random variable is a mean zero multivariate normal if it has a probability density of the form

$$u(x) = \frac{1}{z} e^{-\frac{1}{2}x^* Hx} \; ,$$

for some SPD matrix, H. We can get mean $\mu = (\mu_1, \ldots, \mu_n)^*$ either by taking $X + \mu$ where X has mean zero, or by using the density with x^*Hx replaced by $(x-\mu)^*H(x-\mu)$.

If $X \in R^n$ is multivariate normal and if A is an $m \times n$ matrix with rank m, then $Y \in R^m$ given by $Y = AX$ is also multivariate normal. Both the cases $m = n$ (same number of X and Y variables) and $m < n$ occur.

4.2.3 Diagonalizing H:

Suppose the eigenvalues and eigenvectors of H are $Hv_j = \lambda_j v_j$. We can express $x \in R^n$ as a linear combination of the v_j either in vector form, $x = \sum_{j=1}^n y_j v_j$, or in matrix form, $x = Vy$, where V is the $n \times n$ matrix whose columns are the v_j and $y = (y_1, \ldots, y_n)^*$. Since the eigenvectors of a symmetric matrix are orthogonal to each other, we may normalize them so that $v_j^* v_k = \delta_{jk}$, which is the same as saying that V is an orthogonal matrix, $V^*V = I$. In the y variables, the "quadratic form" x^*Hx is diagonal, as we can see using the vector or the matrix notation. With vectors, the trick is to use the two expressions $x = \sum_{j=1}^n y_j v_j$ and $x = \sum_{k=1}^n y_k v_k$, which are the same since j and k are just summation variables. Then we can write

$$\begin{aligned}
x^*Hx &= \left(\sum_{j=1}^n y_j v_j \right)^* H \left(\sum_{k=1}^n y_k v_k \right) \\
&= \sum_{jk} \left(v_j^* H v_k \right) y_j y_k \\
&= \sum_{jk} \lambda_k v_j^* v_k y_j y_k \\
x^*Hx &= \sum_k \lambda_k y_k^2 \; . \qquad (4.8)
\end{aligned}$$

The matrix version of the eigenvector/eigenvalue relations is $V^*HV = \Lambda$ (Λ being the diagonal matrix of eigenvalues). With this we have $x^*Hx = (Vy)^*HVy = y^*(V^*HV)y = y^*\Lambda y$. A diagonal matrix in the quadratic form is equivalent to having a sum involving only squares $\lambda_k y_k^2$. All the λ_k will be positive if H is positive definite. For future reference, also remember that $\det(H) = \prod_{k=1}^n \lambda_k$.

4.2.4 Calculations using the multivariate normal density:

We use the y variables as new integration variables. The point is that if the quadratic form is diagonal the multiple integral becomes a product of one dimensional gaussian integrals that we can do. For example,

$$\begin{aligned}
\int_{R^2} e^{-\frac{1}{2}(\lambda_1 y_1^2+\lambda_2 y_2^2)} dy_1 dy_2 &= \int_{y_1=-\infty}^{\infty}\int_{y_2=-\infty}^{\infty} e^{-\frac{1}{2}(\lambda_1 y_1^2+\lambda_2 y_2^2)} dy_1 dy_2 \\
&= \int_{y_1=-\infty}^{\infty} e^{-\lambda_1 y_1^2/2} dy_1 \cdot \int_{y_2=-\infty}^{\infty} e^{-\lambda_2 y_2^2/2} dy_2 \\
&= \sqrt{2\pi/\lambda_1}\cdot\sqrt{2\pi/\lambda_2}\ .
\end{aligned}$$

Ordinarily we would need a Jacobian determinant representing $\left|\frac{dx}{dy}\right|$, but here the determinant is $\det(V) = 1$, for an orthogonal matrix. With this we can find the normalization constant, z, by

$$\begin{aligned}
1 &= \int u(x)dx \\
&= \frac{1}{z}\int e^{-\frac{1}{2}x^* Hx} dx \\
&= \frac{1}{z}\int e^{-\frac{1}{2}y^*\Lambda y} dy \\
&= \frac{1}{z}\int \exp(-\frac{1}{2}\sum_{k=1}^{n}\lambda_k y_k^2))dy \\
&= \frac{1}{z}\int \left(\prod_{k=1}^{n} e^{-\lambda_k y_k^2}\right) dy \\
&= \frac{1}{z}\prod_{k=1}^{n}\left(\int_{y_k=-\infty}^{\infty} e^{-\lambda_k y_k^2} dy_k\right) \\
&= \frac{1}{z}\prod_{k=1}^{n}\sqrt{2\pi/\lambda_k} \\
1 &= \frac{1}{z}\cdot\frac{(2\pi)^{n/2}}{\sqrt{\det(H)}}\ .
\end{aligned}$$

This gives a formula for z, and the final formula for the multivariate normal density

$$u(x) = \frac{\sqrt{\det H}}{(2\pi)^{n/2}} e^{-\frac{1}{2}x^* Hx}\ . \tag{4.9}$$

4.2.5 The covariance, by direct integration:

We can calculate the covariance matrix of the X_j. The jk element of $E[XX^*]$ is $E[X_j X_k] = \mathrm{cov}(X_j, X_k)$. The covariance matrix consisting of all these elements is $C = E[XX^*]$. Note the conflict of notation with the constant C above. A direct way to evaluate C is to use the density (4.9):

$$\begin{aligned} C &= \int_{R^n} xx^* u(x)dx \\ &= \frac{\sqrt{\det H}}{(2\pi)^{n/2}} \int_{R^n} xx^* e^{-\frac{1}{2}x^* Hx} dx \ . \end{aligned}$$

Note that the integrand is an $n \times n$ matrix. Although each particular xx^* has rank one, the average of all of them will be a nonsingular positive definite matrix, as we will see. To work the integral, we use the $x = Vy$ change of variables above. This gives

$$C = \frac{\sqrt{\det H}}{(2\pi)^{n/2}} \int_{R^n} (Vy)(Vy)^* e^{-\frac{1}{2}y^* \Lambda y} dy \ .$$

We use $(Vy)(Vy)^* = V(yy^*)V^*$ and take the constant matrices V outside the integral. This gives C as the product of three matrices, first V, then an integral involving yy^*, then V^*. So, to calculate C, we can calculate all the matrix elements

$$B_{jk} = \frac{\sqrt{\det H}}{(2\pi)^{n/2}} \int_{R^n} y_j y_k^* e^{-\frac{1}{2}y^* \Lambda y} dy \ .$$

Clearly, if $j \neq k$, $B_{jk} = 0$, because the integrand is an odd (antisymmetric) function, say, of y_j. The diagonal elements B_{kk} may be found using the fact that the integrand is a product:

$$B_{kk} = \frac{\sqrt{\det H}}{(2\pi)^{n/2}} \prod_{j \neq k} \left(\int_{y_j} e^{-\lambda_j y_j^2/2} dy_j \right) \cdot \int_{y_k} y_k^2 e^{-\lambda_k y_k^2/2} dy_k \ .$$

As before, λ_j factors (for $j \neq k$) integrate to $\sqrt{2\pi/\lambda_j}$. The λ_k factor integrates to $\sqrt{2\pi/(\lambda_k)^{3/2}}$. The λ_k factor differs from the others only by a factor $1/\lambda_k$. Most of these factors combine to cancel the normalization. All that is left is

$$B_{kk} = \frac{1}{\lambda_k} \ .$$

This shows that $B = \Lambda^{-1}$, so

$$C = V\Lambda^{-1}V^* \ .$$

Finally, since $H = V\Lambda V^*$, we see that

$$C = H^{-1} \ . \tag{4.10}$$

The covariance matrix is the inverse of the matrix defining the multivariate normal.

4.2.6 Linear functions of multivariate normals:

A fundamental fact about multivariate normals is that a linear transformation of a multivariate normal is also multivariate normal, provided that the transformation is onto. Let A be an $m \times n$ matrix with $m \leq n$. This A defines a linear transformation $y = Ax$. The transformation is "onto" if, for every $y \in R^m$, there is at least one $x \in R^n$ with $Ax = y$. If $n = m$, the transformation is onto if and only if A is invertible ($\det(A) \neq 0$), and the only x is $A^{-1}y$. If $m < n$, A is onto if its m rows are linearly independent. In this case, the set of solutions is a "hyperplane" of dimension $n - m$. Either way, the fact is that if X is an n dimensional multivariate normal and $Y = AX$, then Y is an m dimensional multivariate normal. Given this, we can completely determine the probability density of Y by calculating its mean and covariance matrix. Writing μ_X and μ_Y for the means of X and Y respectively, we have

$$\mu_Y = E[Y] = E[AX] = AE[X] = A\mu_X \ .$$

Similarly, if $E[Y] = 0$, we have

$$C_Y = E[YY^*] = E[(AX)(AX)^*] = E[AXX^*A^*] = AE[XX^*]A^* = AC_XA^* \ .$$

The reader should verify that if C_X is $n \times n$, then this formula gives a C_Y that is $m \times m$. The reader should also be able to derive the formula for C_Y in terms of C_X without assuming that $\mu_Y = 0$. We will soon give the proof that linear functions of Gaussians are Gaussian.

4.2.7 Uncorrelation and independence:

The inverse of a symmetric matrix is another symmetric matrix. Therefore, C_X is diagonal if and only if H is diagonal. If H is diagonal, the probability density function given by (4.9) is a product of densities for the components. We have already used that fact and will use it more below. For now, just note that C_X is diagonal if and only if the components of X are uncorrelated. Then C_X being diagonal implies that H is diagonal and the components of X are independent. The fact that uncorrelated components of a multivariate normal are actually independent firstly is a property only of Gaussians, and secondly has curious consequences. For example, suppose Z_1 and Z_2 are independent standard normals and $X_1 = Z_1 + Z_2$ and $X_2 = Z_1 - Z_2$, then X_1 and X_2, being uncorrelated, are independent of each other. This may seem surprising in view of that fact that increasing Z_1 by $1/2$ increases both X_1 and X_2 by the same $1/2$. If Z_1 and Z_2 were independent uniform random variables (PDF $= u(z) = 1$ if $0 \leq z \leq 1$, $u(z) = 0$ otherwise), then again X_1 and X_2 would again be uncorrelated, but this time not independent (for example, the only way to get $X_1 = 2$ is to have both $Z_1 = 1$ and $Z_2 = 1$, which implies that $X_2 = 0$.).

4.2.8 Application, generating correlated normals:

There are simple techniques for generating (more or less) independent standard normal random variables. The Box Muller method being the most famous.

Suppose we have a positive definite symmetric matrix, C_X, and we want to generate a multivariate normal with this covariance. One way to do this is to use the Choleski factorization $C_X = LL^*$, where L is an $n \times n$ lower triangular matrix. Now define $Z = (Z_1, \ldots, Z_n)$ where the Z_k are independent standard normals. This Z has covariance $C_Z = I$. Now define $X = LZ$. This X has covariance $C_X = LIL^* = LL^*$, as desired. Actually, we do not necessarily need the Choleski factorization; L does not have to be lower triangular. Another possibility is to use the "symmetric square root" of C_X. Let $C_X = V\Sigma V^*$, where Σ is the diagonal symmetric matrix with eigenvalues of C_X ($\Sigma = \Lambda^{-1}$ where Λ is given above), and V is the orthogonal matrix if eigenvectors. We can take $A = V\sqrt{\Sigma}V^*$, where $\sqrt{\Sigma}$ is the diagonal matrix. Usually the Choleski factorization is easier to get than the symmetric square root.

4.2.9 Central Limit Theorem:

Let X be an n dimensional random variable with probability density $u(x)$. Let $X^{(1)}$, $X^{(2)}$, $\ldots$, be a sequence of independent samples of X, that is, independent random variables with the same density u. Statisticians call this iid (independent, identically distributed). If we need to talk about the individual components of $X^{(k)}$, we write $X_j^{(k)}$ for component j of $X^{(k)}$. For example, suppose we have a population of people. If we choose a person "at random" and record his or her height (X_1) and weight (X_2), we get a two dimensional random variable. If we measure 100 people, we get 100 samples, $X^{(1)}$, $\ldots$, $X^{(100)}$, each consisting of a height and weight pair. The weight of person 27 is $X_2^{(27)}$. Let $\mu = E[X]$ be the mean and $C = E[(X-\mu)(X-\mu)^*]$ the covariance matrix. The Central Limit Theorem (CLT) states that for large n, the random variable

$$R^{(n)} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (X^{(k)} - \mu)$$

has a probability distribution close to the multivariate normal with mean zero and covariance C. One interesting consequence is that if X_1 and X_2 are uncorrelated then an average of many independent samples will have $R_1^(n)$ and $R_2^{(n)}$ nearly independent.

4.2.10 What the CLT says about Gaussians:

The Central Limit Theorem tells us that if we average a large number of independent samples from the same distribution, the distribution of the average depends only on the mean and covariance of the starting distribution. It may be surprising that many of the properties that we deduced from the formula (4.9) may be found with almost no algebra simply knowing that the multivariate normal is the limit of averages. For example, we showed (or didn't show) that if X is multivariate normal and $Y = AX$ where the rows of A are linearly independent, then Y is multivariate normal. This is a consequence of the averaging property. If X is (approximately) the average of iid random variables U_k, then Y is the average of random variables $V_k = AU_k$. Applying the CLT to the averaging of the V_k shows that Y is also multivariate normal.

Now suppose U is a univariate random variable with iid samples U_k, and $E[U_k] = 0$, $E[U_k^2 = \sigma^2]$, and $E[U_k^4] = a_4 < \infty$ Define $X_n = \frac{1}{\sqrt{n}} \sum_{k=n}^{n} U_k$. A calculation shows that $E[X_n^4] = 3\sigma^4 + \frac{1}{n} a_4$. For large n, the fourth moment of the average depends only on the second moment of the underlying distribution. A multivariate and slightly more general version of this calculation gives “Wick’s theorem”, an expression for the expected value of a product of components of a multivariate normal in terms of covariances.

Chapter 5

Lecture 5

5.1 Brownian Motion

Brownian motion is the simplest of the stochastic processes called diffusion processes. It is helpful to see many of the properties of general diffusions appear explicitly in Brownian motion. In fact, all the other diffusion processes may be described in terms of Brownian motion. Furthermore, Brownian motion arises as a limit or many discrete stochastic processes in much the same way that Gaussian random variables appear as a limit of other random variables through the central limit theorem. Finally, the solutions to many other mathematical problems, particularly various common partial differential equations, may be expressed in terms of Brownian motion. For all these reasons, Brownian motion is a central object to study.

5.1.1 Path space:

I will call brownian motion paths $W(t)$ or W_t. In other places people might use B_t, b_t, $Z(t)$, Z_t, etc. The probability space ω will be the space of continuous functions of t for $t \geq 0$ so that $W_0 = 0$. Later, we might consider other starting positions, but that will be explicitly stated when we get there. We might consider finite time or infinite time. That is, we might consider functions W_t, for $0 \leq t \leq T$ or for all $t \geq 0$. The $sigma-$algebra will be the algebra generated by all the "coordinate" functions $X(W) = X_t$ for various t values. Since this is an infinite collection of functions, what we really mean is to consider first finite collections, $t_1 < \cdots < t_n$, and take the $\sigma-$ algebra generated by all these. This complex definition of $\mathcal{F}$ leads to lots of technicality in complete rigorous discussions of Brownian motion. Also important are the $\sigma-$algebras, $\mathcal{F}_t$ with information up to time t. These are generated by the coordinate functions for $t_1 < \cdots < t_n \leq t$.

5.1.2 Increment probabilities:

The probability measure for Brownian motion, called Wiener measure, is specified by giving the probabilities of generating events. These generating events are events generated by finitely many coordinate functions. Let $t_0 < t_1 < \cdots < t_n$. The Brownian motion increments (sometimes called "shocks" by finance people) are $X_k = W_{t_{k+1}} - W_{t_k}$. W_t is a Brownian motion if the increments form a multivariate Gaussian, distinct increments are independent, $E[X_k] = 0$, and

$$\text{var}[X_k] = E[X_k^2] = E[(W_{t_{k+1}} - W_{t_k})^2] = t_{k+1} - t_k \ . \tag{5.1}$$

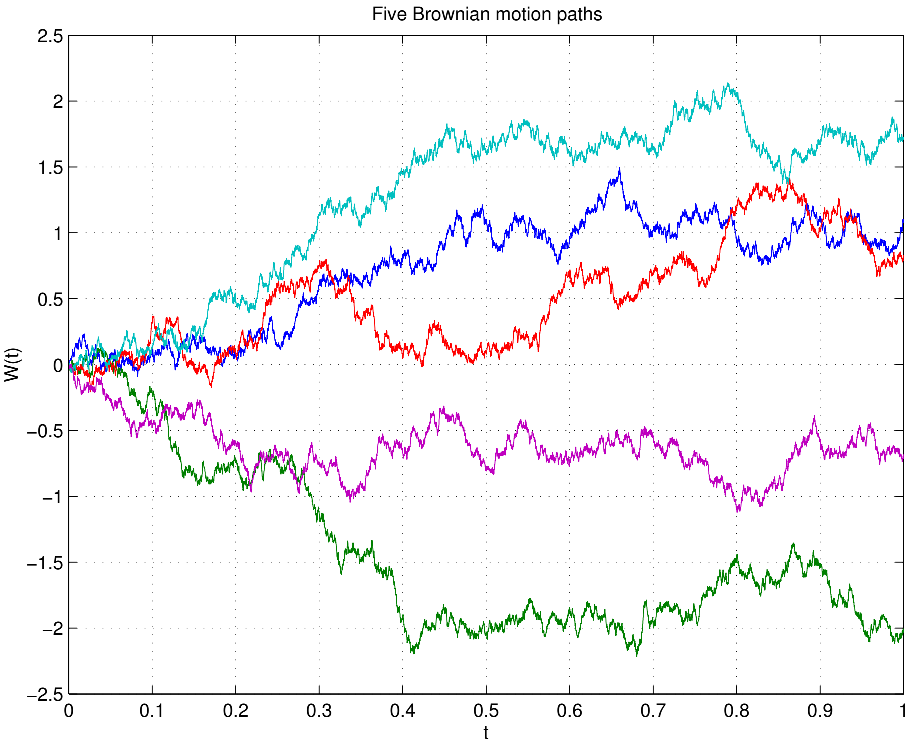
If (5.1) holds for every $n \geq 2$ and every set of times (increasing, of course), then the probability measure is Wiener measure.

5.1.3 Consistency:

There is some technical mathematics between the claim of the above paragraph and it's proof. A first step might be to see that all the different probabilities for different n and t_k are consistent with each other. There is something real here; if $\text{var}(X_k) = (t_{k+1} - t_k)^2$, the probabilities are inconsistent (see below).

Suppose $m < n$ and we have two increasing sequences of times $t_1 < t_2 < \cdots < t_n$ and $\tilde{t}_n < \cdots \tilde{t}_m$. Suppose that the $\tilde{t}_k$ are a subset of the t_j. This means that the random variables $W_{\tilde{t}_k}$ are a subset of the random variables W_{t_j}. Call the joint probability density for the W_{t_j} $u(w_{t_1}, \ldots, w_{t_n})$, and let $\tilde{u}(w_{\tilde{t}_1}, \ldots, w_{\tilde{t}_m})$ be the density for the $W_{\tilde{t}_k}$. We should be able to get the probability density $\tilde{u}$ for the subset of the variables from the larger density u by integrating over the variables not present. That is, $\tilde{u}$ should be the marginal density for the $\tilde{t}_k$ derived from the the density u.

For example, suppose $n = 3$, $m = 2$, $t_1 < t_2 < t_3$ and $\tilde{t}_1 = t_1$ and $\tilde{t}_2 = t_3$. That is, the $\tilde{t}_k$ leave out the middle t. From (5.1), we find that the increments $X_1 = W_{t_2} - W_{t_1}$ and $X_2 = W_{t_3} - W_{t_2}$ are jointly gaussian with zero mean, correlation zero, and variance $\sigma_1^2 = t_2 - t_1$ and $\sigma_2^2 = t_3 - t_2$ respectively. For the $\tilde{t}_k$ we get that the increment $\tilde{X}_1 = W_{t_3} - W_{t_1}$ is gaussian with zero mean and variance $\tilde{\sigma}_1^2 = t_3 - t_1$. On the other hand, $\tilde{X}_1 = X_1 + X_2$ (check this from their definitions), so the distribution of the random variable $\tilde{X}_1$ is determined by those of X_1 and X_2. Are these two definitions of $\tilde{X}_1$ consistent? Yes. The sum of independent normals X_1 and X_2 is normal with variance $\sigma_1^2 + \sigma_2^2$. This shows that leaving out a single intermediate time gives consistent probability distributions. If we leave out times one at a time, we get the overall consistency statement. You should check that if the variance of X_k is not a linear function of $t_{k+1} - t_k$, the distributions are not consistent.



5.1.4 Rough paths, total variation:

The above picture shows 5 Brownian motion paths. They are random and differ in gross features (some go up, others go down), but the fine scale structure of the paths is the same. For one thing, each of the paths, and any part of any path, has infinite "variation" (more technically, "total variation"). Consider times $T_1 < T_2$, choose a large number, n, and divide the time interval $[T_1, T_2]$ into $n-1$ equal side small subintervals t_k, t_{k+1}, where $t_k = T_1 + (k-1)\Delta t$, with $\Delta t = (T_2 - T_1)/(n-1)$. The quantity

$$V = \sum_{k=1}^{n-1} \left| W_{t_{k+1}} - W_{t_k} \right| \tag{5.2}$$

is the Δt variation of W between T_1 and T_2. By the independent increments property, the terms on the right side of (5.2) are independent. By (5.1), they have the same distribution. We estimate the sum of $n-1$ iid random variables using the Central Limit Theorem. The expected value is

$$E[V] = (n-1) \cdot E[|X_1|]$$

where $X_1 \sim \mathcal{N}(0, \Delta t)$. Therefore

$$\begin{aligned} E[|X_1|] &= \frac{1}{2\pi\Delta t}\int_{x=-\infty}^{\infty} |x|\, e^{-x^2/(2\Delta t)}dx \\ &= 2\cdot\frac{1}{2\pi\Delta t}\int_{x=0}^{\infty} xe^{-x^2/(2\Delta t)}dx \\ &= C\sqrt{\Delta t}\ , \end{aligned}$$

where $C = \sqrt{2/\pi}$. Substituting the definition of Δt, this shows that $E[V] = const\sqrt{n-1}$, with $const = (2(T_2 - T_1)/\pi)^{1/2}$. As you take more and more intervals ($n \to \infty$), the total movement of W between T_1 and T_2 goes to infinity.

By contrast, suppose U_t is a differentiable function of time. Then

$$|Ut_{k+1} - U_{t_k}| \approx \left|\frac{dU_t}{dt}\right| (t_{k+1} - t_k)\ ,$$

so

$$\sum_k |Ut_{k+1} - U_{t_k}| \to \int_{T_1}^{T_2} \left|\frac{dU}{dt}\right| dt < \infty \quad \text{as } n \to \infty.$$

The variation of a differentiable function has a limit, the "total variation" as the partition t_k gets finer. For Brownian motion, the finer you look, the more variation you see. Brownian motion paths are not differentiable in the ordinary sense of calculus. The Ito calculus is called for instead.

5.1.5 Dynamic trading:

The infinite total variation of Brownian motion has a consequence for dynamic trading strategies. Some of the simplest dynamic trading strategies, Black-Scholes hedging, and Merton half stock/half cash trading, call for trades that are proportional to the change in the stock price. If the stock price is a diffusion process and there are transaction costs proportional to the size of the trade, then the total transaction costs will either be infinite (in the idealized continuous trading limit) or very large (if we trade as often as possible). It turns out that dynamic trading strategies that take trading costs into account can approach the idealized zero cost strategies when trading costs are small. Next term you will learn how this is done.

5.1.6 Quadratic variation:

The quadratic variation for the partition t_k as above is

$$Q(T_1, T_2, n) = \sum_{k=1}^{n-1} \left(W_{t_{k+1}} - W_{t_k}\right)^2\ . \tag{5.3}$$

This sum takes the squares of the increments, $X_k = W_{t_{k+1}} - W_{t_k}$ rather than the absolute values. For continuous paths, small $\Delta t = (T_2 - T_1)/(n-1)$ should imply small X_k. Therefore the quadratic variation should be smaller than the total variation. In fact, for a differentiable function, $Q \to 0$ as $n \to \infty$. For Brownian motion, the quadratic variation terms are just small enough for the sum not to go to zero or infinity as $n \to \infty$. In fact, the basic formula (5.1) implies that

$$E[Q(T_2, T_1, n)] = \sum_k (t_{k+1} - t_k) = T_2 - T_1 \ , \tag{5.4}$$

for any partition. Since the sum in (5.3) has a large number of iid terms for large n, the Central Limit Theorem suggests that the sum should be close to its expected value. Thus, we have the quadratic variation as the limit

$$Q(T_1, T_2) = \lim_{n\to\infty} Q(T_1, T_2, n) = T_2 - T_1 \ .$$

For other diffusion processes, the quadratic variation limit exists but it's value depends on the path. The quadratic variation is an important ingredient in the Ito calculus.

5.1.7 Trading volatility:

The quadratic variation of a stock price (or a similar quantity) is called it's "realized volatility". The fact that it is possible to buy and sell realized volatility says that the (geometric) Brownian motion model of stock price movement is not completely realistic. That model predicts that realized volatility is a constant, which is nothing to bet on.

5.1.8 Almost sure convergence:

An event, A, is called "almost sure" if $P(A) = 1$. For example, a probabilist would say that the quadratic variation formula (5.4) is true almost surely and might write

$$Q_n \to Q \quad \text{as } n \to \infty \quad a.s. \ .$$

It might seem that this should be called "sure" because we have no doubt that it will happen. The "almost" refers to the fact that (5.4) is might not be true for every $W \in \Omega$. There are paths, continuous functions W_t, so that the limit is infinite and others so that the limit is zero (e.g. differentiable paths). In continuous probability, there are many events that are impossible because they have probability zero, not because the do not exist.

5.1.9 Markov property:

Brownian motion has the Markov property. This is a consequence of the independent increments property. For any t, we have the $\sigma-$ algebras $\mathcal{F}_t$ generated by the W_s for $0 < s \leq t$ (representing past and present), $\mathcal{G}_t$ generated by W_t (representing the present), and $\mathcal{H}_\sqcup$ (representing the future). The Markov property is that for any function $F \in \mathcal{H}_t$, $E[F \mid \mathcal{G}_\sqcup] = E[F \mid \mathcal{F}_\sqcup]$. A function measurable with respect to $\mathcal{H}_t$ depends on the values W_s for $s \geq t$. But W_s for $s \geq t$ is determined by W_t and increments X for intervals (t_k, t_{k+1}) that are measurable in $\mathcal{H}_t$ and independent of all increments that are $\mathcal{F}_t$ measurable. blabla.

5.1.10 Conditional probabilities for intermediate times:

5.1.11 Brownian bridge construction:

5.1.12 Continuous time stochastic process:

The general abstract definition of a continuous time stochastic process is just a probability space, Ω, and, for each $t > 0$, a $\sigma-$algebra $\mathcal{F}_t$. These algebras should be nested (corresponding to increase of information) $\mathcal{F}_{t_1} \subseteq \mathcal{F}_{t_2}$ if $t_1 \leq t_2$. There should also be a family of random variables $Y_t(\omega)$, with Y_t measurable in $\mathcal{F}_t$ (i.e. having a value known at time t). This explains why probabilists often write W_t instead of $W(t)$. For each t, we think of W_t as a function of ω with t simply being a parameter. The Brownian motion has the property that, for every ω (not almost every), the map $t \to W_t(\omega)$ is a continuous function of t. Other stochastic processes, such as the Poisson jump process, do not have continuous sample paths.

5.1.13 Continuous time martingales:

A stochastic process M_t (with Ω and the $\mathcal{F}_t$) is a martingale if $E[M_s \mid \mathcal{F}_t] = M_t$ for $s > t$. Brownian motion forms the first example of a continuous time martingale. Another famous martingale related to Brownian motion is $M_t = W_t^2 - t$ (the reader should check this). For any random variable, Y, the conditional expectations $Y_t = E[Y \mid \mathcal{F}_t]$ form a martingale. The Ito calculus is based on the idea that a stochastic integral with respect to W should produce a martingale.

5.2 Brownian motion and the heat equation

We saw for Markov chains that actual calculations of probabilities and expectation values often make use of forward and backward equations, which we call evolution equations, for probabilities (here, probability densities) and conditional probabilities. For Brownian motion, both the forward and backward equations are “the” heat equation, though the backward equation is often called the “backward heat equation”. We will also find heat equations with boundary conditions that allow us to compute hitting time probability densities and expectations that involve hitting times.

5.2.1 Forward equation for the probability density:

For now we will write X_t for Brownian motion. A Brownian motion starting at $X_0 = 0$ will have probability density at time t that is $\mathcal{N}(0, t)$. We denote this density by

$$g(x,t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2t} \ . \tag{5.5}$$

Directly calculating partial derivatives, we can verify that

$$\partial_t g = \frac{1}{2} \partial_x^2 g. \tag{5.6}$$

This g will play a role below as the "transition density" for Brownian motion, which is more general than just the density for X_t. For example, we could also consider a more general initial density $X_0 \sim u_0(x)$, and independent Gaussian increments as before. (We write $Y \sim v(y)$ to indicate that v is the probability density for the random variable Y, and sometimes also $Y_1 \sim Y_2$ to mean that Y_1 and Y_2 have the same density.) Then the increment $X_t - X_0$ will be $\mathcal{N}(0,t)$ and independent of X_0. That is, X_t is the sum of independent random variables X_0, with density u_0, and $X_t - X_0$, with density $g(\cdot, t)$. Therefore, the density for X_t is

$$u(x,t) = \int_{y=-\infty}^{\infty} g(x-y,t) u_0(y) dy \ , \tag{5.7}$$

Again, direct calculation using (5.5) shows taht u satisfies

$$\partial_t u = \frac{1}{2} \partial_x^2 u \ . \tag{5.8}$$

This is the “heat equation”, also called “diffusion equation”. The equation is used in two ways. First, we can compute probabilities by finding the solution to the partial differential equation. Also, we may be able to find solutions to the partial differential equation if there is an independent way to calculate the probability density.

5.2.2 Heat equation via Taylor series:

There is another way to see that the X_t probability density u satisfies the heat equation (5.8) that proceeds directly from (5.5). This technique has the advantage that we do not have to know the equation in advance. We suppose only that u is a smooth function of x and t and derive the equation by Taylor series calculations. The idea applies in more general situations. It is one approach to the Ito calculus.

The Brownian motion $X_t \sim u(x,t)$ has the property that its increment in a small time interval Δt is $Y = X_{t+\Delta t} - X_t \sim \mathcal{N}(0, \Delta t)$, independent of X_t. As above, this means that $X_{t+\Delta t} = X_t + Y$ has probability density $u(x, t+\Delta t)$ that satisfies

$$u(x, t+\Delta t) = \int g(x-y, \Delta t) u(y,t) dy \ , \tag{5.9}$$

where g is still given by (5.5). Now, for small Δt, the integrand on the right side of (5.9) is significantly different from zero only when $x - y$ is small (not much larger than the order of $\sqrt{\Delta t}$). If u is a smooth function of x, most of the integral will be determined by values of u for y near x. This motivates us to approximate $u(y)$ as a Taylor series about x:

$$u(y) = u(x) + \partial_x u(x) \cdot (y-x) + \frac{1}{2}\partial_x^2 u(x) \cdot (y-x)^2 + O(|x-y|^3);.$$

We integrate the right side of (5.9) with this expansion, remembering that $\int g(x-y,\Delta t)(x-y)^2 dy = \Delta t$, that being the variance of the Δt increment in X. The result is (You can verify that $\int g(y,\Delta t)\,|y|^3\,dy = O(\Delta t^{3/2}.)$:

$$\int g(x-y,\Delta t)u(y,t)dy = u(x,t) + 0 + \Delta t\frac{1}{2}\partial_x^2 u(x,t) + O(\Delta t^{3/2})\ .$$

Of course, we also have

$$u(x,t_\Delta t) = u(x,t) + \Delta t\partial_t u(x,t) + O(\Delta t^2)\ .$$

Using these series for the left and right sides of (5.9) gives

$$u(x,t) + \Delta t\partial_t u(x,t) + O(\Delta t^2) = u(x,t) + \Delta t\frac{1}{2}\partial_x^2 u(x,t) + O(\Delta t^{3/2})\ .$$

We cancel the $u(x,t)$ then divide by Δt and let $\Delta t \to 0$, and we are left with (5.8).

5.2.3 The initial value problem:

The heat equation (5.8) is the Brownian motion analogue of the forward equation for Markov chains. It is often called the forward equation, often to distinguish it from the backward equation discussed below. If we know the time 0 density $u(x,0) = u_0(x)$ and the evolution equation (5.8), the values of $u(x,t)$ are completely and uniquely determined (ignoring mathematical technicalities that would be unlikely to trouble an applied person). The task of finding $u(x,t)$ for $t > 0$ from $u_0(x)$ and (5.8) is called the "initial value problem", with $u_0(x)$ being the "initial value" (or "values"??). This initial value problem is "well posed", which means that the solution, $u(x,t)$, exists and depends continuously on the initial data, u_0. If you want a proof that the solution exists, just use the integral formula for the solution (5.7). Given u_0, the integral (5.7) exists, satisfies the heat equation, and is a continuous function of u_0. The proof that u is unique is more technical (partly because it rests on more technical assumptions).

5.2.4 Ill posed problems:

In some situations, the problem of finding a function u from a partial differential equation and other data may be "ill posed", useless for practical purposes. A problem is ill posed if it is not well posed. This means either that the solution does not exist, or that it does not depend continuously on the data, or that it is not unique. For example, if I try to find $u(x,t)$ for positive t knowing only $u_0(x)$ for $x > 0$, I must fail. A mathematician would say that the solution, while it exists, is not unique, there being many different ways to give $u_0(x)$ for $x > 0$, each leading to a different u. A more subtle situation arises, for example, if we give $u(x,T)$ for all x and wish to determine $u(x,t)$ for $0 \leq t < T$. For example, if $u(x,T) = \mathbf{1}_{[0,1]}(x)$, there is no solution (trust me). Even if there is a solution, for example given by (5.7), is does not depend continuously on the values of $u(x,T)$ for $T > t$ (trust me).

The heat equation (5.8) relates values of u at one time to values at another time. However, it is "well posed" only for determining u at future times from u at earlier times. This "forward equation" is well posed only for moving forward in time.

5.2.5 Conditional expectations:

We saw already for Markov chains that certain conditional expected values can be calculated by working backwards in time with the backward equation. The Brownian motion version of this uses the conditional expectation

$$f(x,t) = E[V(X_T) \mid X_t = x] \ . \qquad (5.10)$$

The "modern" formulation of this gives $f_t = E[V(X_t) \mid \mathcal{F}_t]$, which is, as has been repeated, a function of $X_t = x$ only. Of course, these definitions mean the same thing. The definition is also sometimes written as $f(x,t) = E_{x,t}[X_t]$. This is in the spirit of writing $E_\alpha[]$ for expectation with respect to the given probability measure P_α. Here, the probability measure $P_{x,t}$ is Brownian motion starting from x at time t, which is defined by the densities of increments for times larger than t as before.

5.2.6 Backward equation by direct verification:

The expectation (5.10) depends on the increment $X_T - X_t$, which is $\mathcal{N}(0, T-t)$ and independent of X_t. Thus, the conditional density of X_T given that $X_t = x$ is (as a function of y) $g(y-x, T-t)$. Writing the expectation $f(x,t)$ as an integral, we get

$$f(x,t) = \int_{-\infty}^{\infty} g(x-y, T-t) V(y) dy \ . \qquad (5.11)$$

Since this depends on x and t only through g, we can again verify through explicit calculation that

$$\partial_t f + \frac{1}{2}\partial_x^2 f = 0 \ . \qquad (5.12)$$

Note that the sign of ∂_t here is not what it was in (5.8), which is because we are calculating $\partial_t g(T-t)$ rather than $\partial_t g(t)$. This (5.12) is the "backward equation".

5.2.7 Backward equation by Taylor series:

As with the forward equation (5.8), we can find the backward equation by Taylor series expansions. Indeed, since $\mathcal{F}_t \subset \mathcal{F}_{t+\Delta t}$, we have, in “modern” notation,

$$f_t = E[V(X_T) \mid \mathcal{F}_t] = E[E[V(X_T) \mid \mathcal{F}_{t+\Delta t}] \mid \mathcal{F}_t] = E[f(X_{t_\Delta t} \mid \mathcal{F}_t] \; .$$

Using the probability density for the increment $X_{t+\Delta t} - X_t$, this gives the integral relation

$$f_t(x,t) = \int_{y=-\infty}^{\infty} g(x-y,\Delta t) f_{t+\Delta t}(y) dy \; . \tag{5.13}$$

Using Taylor series on the right and left (in different ways as above) again leads to (5.12).

5.2.8 The final value problem:

We get a well posed problem by giving the partial differential equation (5.12) together with the “final values” $f(x,T) = V(x)$ (The definition (5.10) makes this obvious.). The “backwards heat equation enables us to find values of f at early times from given values at later times. The initial value problem, finding f_t with $t > 0$ from f_0 is not well posed. Although there may be occasional solutions, it is not a useful way to find the general solution, either because the general solution does not exist or because the solution that happens to exist does not depend in a continuous way on the values f_0.

5.2.9 Duality:

You can check directly the duality property that $\int f(x,t)u(x,t)dx$ is independent of t. As for the Markov chain case, this is a consistency relation between the forward and backward evolution equations that makes one “dual” to the other. Also as for Markov chains, the integral is an expression of the law of total probability, integrating the expected payout starting at x at time t multiplied by the probability density for being at x at time t. This is $E[V(X_T)]$, and is thus independent of t.

5.2.10 The smoothing property, regularity:

Solutions of the forward or backward heat equation become smooth functions of x and t even if the initial data (for the forward equation) or final data (for the backward equation) are not smooth. For u, this is clear from the integral formula (5.7). If we differentiate with respect to x, this derivative passes under the integral and onto the g factor. This applies also to x or t derivatives of any order, since the corresponding derivatives of g are still smooth integrable functions of x. The same can be said for f using (5.11); as long as $t < T$, any derivatives of f with respect to x and/or t are bounded. A function that has all partial derivatives of any order bounded is called “smooth”. (Warning, this term is not used consistently. Some people say “smooth” to mean, for example, merely having derivatives up to second order bounded.) Solutions of more general forward and backward equations often, but not always, have the smoothing property.

5.2.11 Rate of smoothing:

Suppose the payout (and final value) function, $V(x)$, is a discontinuous function such as $V(x) = \mathbf{1}_{x>0}(x)$ (a "digital" option in finance). For t close to T, $f(x,t)$ will be a differentiable function of x, but the derivative will be very large in some places. In fact,

$$\max_x |\partial_x f(x,t)| \sim \frac{1}{\sqrt{T-t}} \; .$$

Higher derivatives of f "explode" faster as t approaches T. If $V(x) = x_+$ (x_+ being the "positive part" of x, either x or 0 depending on which is larger), then the $\partial_x f$ is bounded as t approaches T, but the curvature "blows up". The fact that derivatives of f blow up at t approaches T makes numerical solution of the backward equation difficult and inaccurate.

5.2.12 Diffusion:

It sometimes helps the intuition to think of particles diffusing through some medium, ink particles diffusing through still water, for example. Then $u(x,t)$ can represent the density of particles about x at time t. If ink has been diffusing through water for some time, there might be dark regions with a high density of particles (large u) and lighter regions with smaller u. This helps us interpret, for example, solutions of the heat equation (5.8) without the requirement that $\int u(x,t)dx = 1$. For ink in water, it is a reasonable approximation to think of each particle performing it's own Brownian motion independent of all the others. If the density of particles were too high (e.g. all particles and no water), we would have to adjust the model. A physical argument that tiny particles in water should undergo Brownian motion, and that their density should satisfy the heat equation, was given by the German physicist Albert Einstein, and was the basis of his Nobel Prize (relativity and quantum mechanics seeming too uncertain at the time).

5.2.13 Heat:

Heat also can diffuse through a medium, as happens when we put a thick metal pan over a flame and wait for the other side to heat up. We can think of $u(x,t)$ as representing the temperature in a metal at location x at time t. This helps us interpret solutions of the heat equation (5.8) when u is not necessarily positive. In particular, it helps us imagine the "cancellation" that can occur when regions of positive and negative u are close to each other. Heat flows from the high temperature regions to low or negative temperature regions to create a more uniform equilibrium temperature. A physical argument that heat (temperature) flowing through a metal should satisfy the heat equation was given by the French mathematical physicist, friend of Napoleon, and founder of Ecole Polytechnique, Joseph Fourier.

5.2.14 Hitting times:

A stopping time, τ, is any time that depends on the Brownian motion path X so that the event $\tau \leq t$ is measurable with respect to $\mathcal{F}_t$. This is the same as saying that for each t there is some process that has as input the values X_s for $0 \leq s \leq t$ and as output a decision $\tau \leq t$ or $\tau > t$. One kind of stopping time is a hitting time:

$$\tau_a = \min\left(t \mid X_t = a\right) \ .$$

More generally (particularly for Brownian motion in more than one dimension) if A is a closed set, we may consider $\tau_A = \min(t \mid X_t \in A)$. It is useful to define a Brownian motion that stops at time τ: $\tilde{X}_t = X_t$ if $t \leq \tau$, $\tilde{X}_t = X_\tau$ if $t \geq \tau$.

5.2.15 Probabilities for stopped Brownian motion:

Suppose X_t is Brownian motion starting at $X_0 = 1$ and $\tilde{X}$ is the Brownian motion stopped at time τ_0, the first time $X_t = 0$. The probability measure, P_t, for $\tilde{X}_t$ may be written as the sum of two terms, $P_t = P_t^s + P_t^{ac}$. (Since $\tilde{X}_t$ is a single number, the probability space is $\Omega = R$, and the $\sigma-$algebra is the Borel algebra.) The "singular" part, P_t^s, corresponds to the paths that have been stopped. If $p(t)$ is the probability that $\tau \leq t$, then $P_t^s = p(t)\delta(x)$, which means that for any Borel set, $A \subseteq R$, $P_t^s(A) = p(t)$ if $0 \in A$ and $P_t^s(A) = 0$ if $0 \notin A$. This δ is called the "delta function" or "delta mass"; it puts weight one on the point zero and no weight anywhere else. Probabilists sometimes write δ_{x_0} for the measure that puts weight one on the point x_0. Physicists write $\delta_{x_0}(x) = \text{‘}delta(x = x_0)$. The "absolutely continuous" part, P_t^{ac}, is given by a density, $u(x,t)$. This means that $P_t^{ac}(A) = \int_A u(x,t)dx$. Because $\int_R u(x,t)dx = 1 - p(t) < 1$, u, while being a density, is not a probability density.

This decomposition of a measure (P) as a sum of a singular part and absolutely continuous part is a special case of the Radon Nikodym theorem. We will see the same idea in other contexts later.

5.2.16 Forward equation for u:

The density for the absolutely continuous part, $u(x,t)$, is the density for paths that have not touched $X = a$. In the diffusion interpretation, think of a tiny ink particle diffusing as before but being absorbed if it ever touches a. It is natural to expect that when $x \neq a$, the density satisfies the heat equation (5.8). u "knows about" the boundary condition because of the "boundary condition" $u(a,t) = 0$. This says that the density of particles approaches zero near the absorbing boundary. By the end of the course, we will have several ways to prove this. For now, think of a diffusing particle, a Brownian motion path, as being hyperactive; it moves so fast that it has already visited a neighborhood of its current location. In particular, if X_t is close to a, then very likely $X_s = a$ for some $s < t$. Only a small minority of the particles at x near a, with small density $u(x,t) \to 0$ as $x \to a$ have not touched a.

5.2.17 Probability flux:

Suppose a Brownian motion starts at a random point $X_0 > 0$ with probability density $u_0(x)$ and we take the absorbing boundary at $a = 0$. Clearly, $u(x,t) = 0$ for $x < 0$ because a particle cannot cross from positive to negative without crossing zero, the Brownian motion paths being continuous. The probability of not being absorbed before time t is given by

$$1 - p(t) = \int_{x>0} u(x,t)dx \ . \tag{5.14}$$

The rate of absorbtion of particles, the rate of decrease of probability, may be calculated by using the heat equation and the boundary condition. Differentiating (5.14) with respect to t and using the heat equation for the right side then integrating gives

$$\begin{aligned} -\dot{p}(t) &= \int_{x>0} \partial_t u(x,t)dx \\ &= \int_{x>0} \frac{1}{2}\partial_x^2 u(x,t)dx \\ \dot{p}(t) &= \frac{1}{2}\partial_x u(x,0) \ . \end{aligned} \tag{5.15}$$

Note that both sides of (5.15) are positive. The left side because $P(\tau \leq t)$ is an increasing function of t, the right side because $u(0,t) = 0$ and $u(x,t) > 0$ for $x > 0$. The identity (5.15) leads us to interpret the left side as the probability "flux" (or "density flux if we are thinking of diffusing particles). The rate at which probability flows (or particles flow) across a fixed point ($x = 0$) is proportional to the derivative (the gradient) at that point. In the heat flow interpretation this says that the rate of heat flow across a point is proportional to the temperature gradient. This natural idea is called Fick's law (or possibly "Fourier's law").

5.2.18 Images and Reflections:

We want a function $u(x,t)$ that satisfies the heat equation when $x > 0$, the boundary condition $u(0,t) = 0$, and goes to δ_{x_0} as $t \downarrow 0$. The "method of images" is a trick for doing this. We think of δ_{x_0} as a unit "charge" (in the electrical, not financial sense) at x_0 and $g(x - x_0, t) = \frac{1}{\sqrt{2\pi}} e^{-(x-x_0)^2/2t}$ as the response to this charge, if there is no absorbing boundary. For example, think of putting a unit drop of ink at x_0 and watching it spread along the x axis in a "bell shaped" (i.e. gaussian) density distribution. Now think of adding a negative “image charge” at $-x_0$ so that $u_0(x) = \delta_{x_0} - \delta_{-x_0}$ and correspondingly

$$u(x,t) = \frac{1}{\sqrt{2\pi t}} \left(e^{-(x-x_0)/2t} - e^{-(x+x_0)/2t} \right) . \tag{5.16}$$

This function satisfies the heat equation everywhere, and in particular for $x > 0$. It also satisfies the boundary condition $u(0,t) = 0$. Also, it has the same initial data as g, as long as $x > 0$. Therefore, as long as $x > 0$, the u given by (5.16) represents the density of unabsorbed particles in a Brownian motion with absorption at $x = 0$. You might want to consider the image charge contribution in (5.16), $\frac{1}{\sqrt{2\pi}} e^{-(x-x_0)^2/2t}$, as “red ink” (the ink that represents negative quantities) that also diffuses along the x axis. To get the total density, we subtract the red ink density from the black ink density. For $x = 0$, the red and black densities are the same because the distance to the sources at $\pm x_0$ are the same. When $x > 0$ the black density is higher so we get a positive u. We can think of the image point, $-x_0$, as the reflection of the original source point through the barrier $x = 0$.

5.2.19 The reflection principle:

The explicit formula (5.16) allows us to evaluate $p(t)$, the probability of touching $x = 0$ by time t starting at $X_0 = x_0$. This is

$$p(t) = 1 - \int_{x>0} u(x,t)dx = \int_{x>0} \frac{1}{\sqrt{2\pi t}} \left(e^{-(x-x_0)/2t} - e^{-(x+x_0)/2t} \right) dx .$$

Because $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi t}} e^{-(x-x_0)/2t} dx = 1$, we may write

$$p(t) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi t}} e^{-(x-x_0)/2t} dx + \int_{0}^{\infty} \frac{1}{\sqrt{2\pi t}} e^{-(x+x_0)/2t} dx .$$

Of course, the two terms on the right are the same! Therefore

$$p(t) = 2 \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi t}} e^{-(x-x_0)/2t} dx .$$

This formula is a particular case the Kolmogorov reflection principle. It says that the probability that $X_s < 0$ for some $s \leq t$ is (the left side) is exactly twice the probability that $X_t < 0$ (the integral on the right). Clearly some of the particles that cross to the negative side at times $s < t$ will cross back, while others will not. This formula says that exactly half the particles that touch for some $s \leq t$ $x = 0$ have $X_t > 0$. Kolmogorov gave a proof of this based on the Markov property and the symmetry of Brownian motion. Since $X_\tau = 0$ and the increments of X for $s > \tau$ are independent of the increments for $s < \tau$, and since the increments are symmetric Gaussian random variables, they have the same chance to be positive $X_t > 0$ as negative $X_t < 0$.

Chapter 6

Lecture 6

6.1 Integration with respect to Brownian Motion

While integrals of functions of Brownian motion paths are not hard to define, integrals with respect to Brownian motion do give trouble. In fact, there is some ambiguity about what the integral should be. The Ito integral is really just a convention to choose one of the several possibilities. The Ito convention is that the "stochastic integral" with respect to Brownian motion should be a martingale.

Many financial models take the form of stochastic differential equations (SDE). The definition of the solution of an SDE has the same ambiguity as the stochastic integral with respect to Brownian motion. We again choose the Ito convention that the solution as far as possible should be a martingale. In fact, the solution of an Ito SDE is defined in terms of the Ito integral.

6.1.1 Integrals involving a function of t only:

The stochastic integral with respect to Brownian motion is an integral in which dX_t (whatever that means) plays the role of dt in the Riemann integral. The simplest case involves just a function of t:

$$Y_g = \int_0^T g(t) dX_t \ . \tag{6.1}$$

This integral is defined in somewhat the same way the Riemann integral is defined. We choose n and $\Delta t = T/n$ and take

$$Y_g^{(n)} = \sum_{k=0}^{n-1} g(t_k) \Delta X_k \ , \tag{6.2}$$

where $\Delta X_k = X_{t_{k+1}} - X_{t_k}$, and $t_k = k\Delta t$. Since $Y_g^{(n)}$ is a sum of gaussian random variables, it is also gaussian. Clearly $E[Y_g^{(n)}] = 0$. We will understand the limit as $\Delta t \to 0$ (including whether it exists) if we calculate the limit of $\mathrm{var}(Y_g^{(n)} = E[Y_g^{(n)2}]$. Since the ΔX_k are independent normals with mean zero and variance Δt, the variance of the sum is

$$\mathrm{var}(Y_g^{(n)}) = \sum_{k=0}^{n-1} g(t_k)^2 \Delta t \ .$$

The right side is the standard Riemann approximation to the integral $\int_0^T g(t)^2 dt$, so letting $\Delta t \to 0$ gives

$$E[Y_g^2] = \mathrm{var}(Y_g) = \int_0^T g(t)^2 dt \ . \tag{6.3}$$

This may not have seemed so subtle, and it was not. Every reasonable definition of (6.1) gives the same answer.

6.1.2 Different kinds of convergence:

In the abstract setting we have a probability space, Ω, and a family of random variables $Y_n(\omega)$. We want to take the limit as $n \to \infty$. The limit above is the limit "in distribution". The probability density for Y_n converges to the probability density of a random variable Y. The cental limit theorem is of this kind: the probability density converges to a gaussian. Another kind of convergence as "pointwise", asking that, for each ω (or almost every ω) the limit $\lim_{n\to\infty} Y_n(\omega) = Y(\omega)$ should exist. The difference between these notions is that one gives an actual (function of a) random variable, $Y(\omega)$, while the other just gives a probability density without necessarily saying which Y goes with a particular ω. Proving convergence in distribution for gaussian random variables is easy, just calculate the mean and variance. Note that this does not depend on the joint distribution of Y_n and Y_{n+1}. Proving pointwise convergence requires you to understand the differences $Y_{n+1} - Y_n$, which do depend on the joint distributions.

6.1.3 Proving pointwise convergence:

The abstract setting has a probability space, Ω, with a probability measure, and a sequence of random variables, $X_n(\omega)$. The X_n could be just numbers (what we usually call random variables), or vectors (vector values random variables), or even functions of another variable (say, t). In any of these cases, we have a norm, $\|X\|$. For the case of a number, we just use the absolute value, $|X|$. For vectors, we can use any vector norm. For functions, we can also use any norm, such as the "sup" norm, $\|X(\omega)\| = \max_{0\leq t\leq T} |X(\omega, t)|$, or the L^2 norm, $\|X\|^2 = \int_{0\leq t\leq T} X(\omega,t)^2 dt$. A theorem in analysis says that the limit $\lim_{n\to\infty} X_n(\omega)$

exists if

$$S(\omega) = \sum_{n=1}^{\infty} \|X_{n+1}(\omega) - X_n(\omega)\| < \infty \ . \tag{6.4}$$

This is easy to understand. The limit exists if and only if the sum, $X_1(\omega) + \sum_n X_{n+1}(\omega) - X_n(\omega)$, converges. The condition (6.4) just says that this sum converges absolutely. It is possible that the limit exists even though S is infinite. For example (forgetting ω) if $X_n = (-1)^n/n$.

The limit will exist for (almost) every ω if $S(\omega) < \infty$ for (almost) every ω. We know that $S < \infty$ almost surely if $E[S] = \int S(\omega)dP(\omega) < \infty$. The expected value criterion is useful because we might be able to calculate the expected value, particularly in a Stochastic Calculus class that is devoted mostly to such calculations. Of course, it is possible that $S < \infty$ almost surely even though $E[S] = \infty$. For example, suppose $S = 1/Z^2$ where Z is a standard normal $S \sim \mathcal{N}(0,1)$ (OK, not likely for (6.4), but that does not change this point). In jargon, we would say these criteria are not "sharp"; it is possible to fail these tests and still converge. As far as I can tell, a sharp criterion would be much more complicated, and unnecessary here. From (6.4), the criterion $E[S] < \infty$ may be stated:

$$\sum_{n=1}^{\infty} E\Big[\|X_{n+1} - X_n\|\Big] < \infty \ . \tag{6.5}$$

We continue our succession of convenient but not sharp criteria. It is often easier to calculate $E[Y^2]$ than $E[\|Y\|]$. Fortunately, there is the "Cauchy Schwartz" inequality: $E[|Y|] < E[Y^2]^{1/2}$ (proof left to the reader). If we define (and hope to calculate)

$$s_n^2 = E\Big[\|X_{n+1} - X_n\|^2\Big] \ ,$$

then $E\Big[\|X_{n+1} - X_n\|\Big] < s_n$, so (6.5) holds if

$$\sum_{n=1}^{\infty} s_n < \infty \ . \tag{6.6}$$

6.1.4 The integral as a function of X:

We apply the above criteria to showing that the limit (6.2) exists for (almost) any Brownian motion path, X. Pointwise convergence does two things for us. First, it shows that Y_g is a function of X, i.e., a random variable defined on the probability space of Brownian motion paths. Second, it shows that shows that if we use the approximation (6.2) on the computer, we will get an approximation to the right $Y_g(X)$, not just a random variable with (approximately) the right distribution. Whether that is important is a subject of heated debate, with me heatedly on one of the sides. We will see that it is much easier to compare $Y_g^{(n)}$ with $Y_g^{(2n)}$ than with $Y_g^{(n+1)}$. To translate from our situation to the abstract, the abstract X_n will be our $Y_g^{(2^L)}$, the abstract n our L, and the abstract ω our X. That is, we seek to show that the limit

$$\lim_{L\to\infty} Y_g^{(2^L)}(X) = Y_g(X)$$

exists for (almost) every Brownian motion path, X. We will do this by calculating (bounding would be a more apt term) $E[(Y_g^{(2n)} - Y_g^{(n)})^2]$ with $n = 2^L$.

6.1.5 Comparing the Δt and $\Delta t/2$ approximations:

We will fix g and stop writing it. We have $Y^{(n)}$ based on $\Delta t = T/n$ and $Y^{(2n)}$ based on $\Delta t/2 = T/(2n)$. We take $t_k = k\Delta t$, which is appropriate for $Y^{(n)}$. The contribution to $Y^{(n)}$ from the interval (t_k, t_{k+1}) is $g(t_k)(X_{t_{k+1}} - X_{t_k})$. For $Y^{(2n)}$, the interval (t_k, t_{k+1}) is divided into two subintervals $(t_k, t_{k+1/2})$ and $(t_{k+1/2}, t_{k+1})$, using the notation $t_{k+1/2} = t_k + \Delta t/2 = (k+1/2)\Delta t$. The the contribution to $Y^{(2n)}$ from these two intervals added is

$$g(t_k)(X_{t_{k+1/2}} - X_{t_k}) + g(t_{k+1/2})(X_{t_{k+1}} - X_{t_{k+1/2}}) \ .$$

Define ΔY_k to be the difference between the single $Y^{(n)}$ contribution and the two $Y^{(2n)}$ contributions from the interval (t_k, t_{k+1}), so that $Y^{(2n)} - Y^{(n)} = \sum_{k=0}^{n-1} \Delta Y_k$. A calculation gives

$$\Delta Y_k = (g(t_{k+1}) - g(t_{k+1/2}))(X_{t_{k+1}} - X_{t_{k+1/2}}) \ .$$

Only the X values are random, and increments $X_{t_{k+1}} - X_{t_{k+1/2}}$ from distinct intervals are independent. Therefore

$$E\left[(Y^{(2n)} - Y^{(n)})^2\right] = \sum_{k=0}^{n-1} E[\Delta Y_k^2] = \sum_{k=0}^{n-1} \Delta g_k^2 \Delta t/2 \ ,$$

where we have used the notation $\Delta g_k = g(t_{k+1}) - g(t_{k+1/2})$ and the fact that $E[(X_{t_{k+1}} - X_{t_{k+1/2}})^2] = \Delta t/2$.

Now suppose that $|g'(t)| \leq r$ for all t. Then $\Delta g_k \leq \frac{r\Delta t}{2}$ (an interval of length $\frac{\Delta t}{2}$) so

$$E\left[(Y^{(2n)} - Y^{(n)})^2\right] \leq n\frac{r^2\Delta t^2}{4}\frac{\Delta t}{2} \ .$$

Simplifying using the relationship $n\Delta t = T$ gives

$$E\left[(Y^{(2n)} - Y^{(n)})^2\right] \leq T\frac{r^2\Delta t^2}{8} \ .$$

Finally, take n to be of the form 2^L, write $Y_L = Y^{(2^L)}$, and see that we have shown $s_L^2 \leq \mathit{Const}\cdot\Delta t^2$, so $s_L \leq \mathit{Const}\cdot 2^-L$, and the criterion (6.6) is easily satisfied.

6.1.6 Unanswered theoretical questions:

Here are some questions that would be taken up in a more theoretical course and their answers, without proof. *Q1*: This defines Y_g only for functions $g(t)$ that are differentiable. What about other functions? *A1*: Because $E[Y_g^2] = \int_0^T g(t)^2 dt$, we can "extend" the mapping $g \mapsto Y_g$ to any g with $\int g^2 < \infty$, as we do for the Fourier transform. *Q2*: What happens if we let $n \to \infty$ but not by powers of 2? *A2*: This can be done in at least two ways, either using a more sophisticated argument and higher than second order moments, or by using a uniqueness theorem for the limit. Even without this, we met our primary goal of showing that Y_g is a well defined function of X.

6.1.7 White noise:

White noise is something of an idealization, like the $\delta-$function. Imagine a function, $W(t)$ that is gaussian with mean zero and has $W(t)$ independent of $W(s)$ for $t \neq s$. Also imagine that the strength of the noise is independent of time. This is a common model for fluctuations. For example, in modeling phone calls, we may think that the rate of new calls being initiated fluctuates from its mean but that fluctuations at different times are independent. Suppose we try to integrate white noise over intervals of time: $Y_{[a,b]} = \int_a^b W(t)dt$. We can determine how the variance $\sigma^2_{[a,b]} = E[Y^2_{[a,b]}]$ depends on the interval by noting that Y variables for disjoint intervals should be independent. In particular, if $a < b < c$ we have $Y_{[a,c]} = Y_{[a,b]} + Y_{[b,c]}$, so $\sigma^2_{[a,c]} = \sigma^2_{[a,b]} + \sigma^2_{[b,c]}$. The only this can happen, and have, for any offset, d, $\sigma^2_{[a,b]} = \sigma^2_{[a+d,b+d]}$ (homogeneous in time) is for $\sigma^2_{[a,b]} = Const\cdot(b-a)$. The "standard" white noise has $\sigma^2_{[a,b]} = b-a$.

6.1.8 White noise is not a function:

White noise is too rough to be a function, even a random function, in the usual sense. To see this, consider an interval $(0,\epsilon)$. Since $\int_0^\epsilon W(t)dt$ has variance ϵ, it's standard deviation, which is the order of magnitude of a typical $Y_{[0,\epsilon]}$, is $\sqrt{\epsilon}$. In order to have $\int_0^\epsilon W(t)dt \sim \sqrt{\epsilon}$, we must have $W(t) \sim 1/\sqrt{\epsilon}$ in at least over a reasonable fraction of the interval. Letting $\epsilon \to 0$, we see that $W(t)$ should have infinite values almost everywhere, not much of a function. Just as the $\delta-$function is defined in an abstract way as a measure, there are abstract definitions that allow us to make sense of white noise.

Another way to see this is to try to define $Y_T = \int_0^T W(t)^2 dt$. Since we already think white noise paths are discontinuous, it is natural to try to define the Riemann sum using averages over small intervals rather than values $W(t_k)$. We call the averages

$$W_{k,n} = \frac{1}{\Delta t} \int_{t_k}^{t_{k+1}} W(t)dt \; .$$

The approximation to Y_Y is

$$Y_T^{(n)} = \Delta t \sum_{k=0}^{n-1} W_{k,n}^2 \ .$$

The random variables $W_{k,n}$ are independent gaussians with mean zero and variance $\frac{1}{\Delta t^2}\Delta t = \frac{1}{\Delta t}$. Therefore the $W_{k,n}^2$ are independent with mean $\frac{1}{\Delta t}$ and variance $\frac{2}{\Delta t^2}$ (as the reader should verify). Therefore, $Y_T^{(n)}$ has mean $T/\Delta t$ and standard deviation $\sqrt{2T/\Delta t}$. Clearly, as $n \to \infty$, $Y_T^{(n)} \to \infty$. In other words, $\int_0^T W(t)^2 dt = \infty$ by the most reasonable definition.

6.1.9 White noise and Brownian motion:

The integrals $Y_{[a,b]}$ of white noise have the same statistical properties as the increments of Brownian motion. The joint distribution of $Y_{[a_1,b_1]}$, $\ldots$, $Y_{[a_n,b_n]}$ is the same as the joint distribution of the increments $X_{b_1} - X_{a_1}$, $\ldots$, $X_{b_n} - X_{a_n}$ (assuming, though this is not necessary, that $a_1 \leq b_1 \leq a_2 \cdots \leq b_n$): both are multivariate normal with zero covariances and variances $b_k - a_k$. If $W(t)$ were a function, this would lead us to write the three relationships

$$X_t = \int_0^t W(s)ds \ , \quad \frac{dX_t}{dt} = W(t) \ , \quad dX_t = W(t)dt \ . \tag{6.7}$$

Any of these may be taken as the definition of white noise. This is probably the main reason most people (who are interested) are interested in Brownian motion, that it gives a mathematically rigorous and systematic way to make sense of white noise.

6.1.10 Correlations of integrals with respect to Brownian motion:

It seems clear that two integrals with respect to Brownian motion should be jointly gaussian with some covariance we can calculate. In fact, if $Y_f = \int_0^T f(t)dX_t$ and $Y_g = \int_0^T g(t)dX_t$, then the approximations $Y_f^{(n)}$ and $Y_g^{(n)}$ are jointly normal and have covariance $\sum_{k=0}^{n-1} f(t_k)g(t_k)\Delta t$. Taking the limit $\Delta t \to 0$ gives

$$\operatorname{cov}(Y_f, Y_g) = \int_{t=0}^T f(t)g(t)dt \ . \tag{6.8}$$

6.1.11 δ correlated white noise:

The correlation formula (6.8) has an interpretation used by 90% of the interested world, not including most mathematicians. If we write $Y_g = \int_{s=0}^T g(s)dX_s$ and formally interchange the order of integration, we get, since dX_t and dX_s are the only random variables,

$$\begin{aligned} E[Y_f Y_g] &= E\left[\int_{t=0}^{T} f(t)dX_t \int_{s=0}^{T} g(s)dX_s\right] \\ &= \int_{t\in[0,T]}\int_{s\in[0,T]} f(t)g(s)E[dX_t dX_s]\ . \end{aligned}$$

We get (6.8) with the rule

$$E[dX_t dX_s] = \delta(t-s)dt\ . \tag{6.9}$$

This is the first instance of the informal Ito rule $dX^2 = dt$. It is equivalent to (6.7) with the rule $E[W(t)W(s)] = \delta(t-s)$, which is another indication that white noise is not a normal function. If we write $W(t)dt = dX_t$ to write $Y_f = \int f(t)W(t)dt$, the formula (6.8) follows.

A useful approximation to white noise with a time step Δt is

$$W^{(\Delta t)}(t) = \sum_k Z_k \mathbf{1}_{I_k}(t) \tag{6.10}$$

where I_k is the interval $[t_k, t_{k+1}]$ and the Z_k are independent gaussians with the proper variance $\text{var}(Z_k) = \frac{1}{\Delta t}$. For example, this gives

$$\int_0^T f(t)W^{(\Delta t)}dt = \sum_k \int_{I_k} f(t)dtZ_k\ ,$$

which is a random variable practically identical to the approximation (6.2). The difference is that $\Delta t f(t_k)$ is replaced by $\int_{I_k} f(t)dt = f(t_k)\Delta t + o(\Delta t)$. We identify the random variables ΔX_k and $\Delta t Z_k$ because they are both multivariate normal and have the same mean ($E[] = 0$), variance, and covariances.

6.2 Ito Integration

6.2.1 Forward dX_t:

We want to define stochastic integrals such as

$$Y_T = \int_0^T V(X_t)dX_t\ . \tag{6.11}$$

The Ito convention is that $E[dX_t \mid \mathcal{F}_t] = 0$. When we make $\Delta t = T/n$ approximations to (6.11), we always do it in a way that makes the analogue of dX_t have conditional expectation zero. For example, we might use

$$Y_T^{(n)} = \sum_{k=0}^{n-1} V(X_{t_k})(X_{t_{k+1}} - X_{t_k})\ . \tag{6.12}$$

The specific choice $\Delta X_k = X_{t_{k+1}} - X_{t_k}$ gives $E[\Delta X_k \mid \mathcal{F}_{t_k}] = 0$, which is in keeping with the Ito convention. We will soon show that the limit $Y_T^{(n)}(X) \to Y_T(X)$ exists. This limit is the Ito integral.

6.2.2 Example 1:

This example illustrates the convergence of the approximations, the way in which the Ito integral differs from an ordinary integral, and the fact that other approximations of dX_t lead to different limits. Take

$$Y_T = \int_0^T X_t dX_t \ ,$$

and use the approximation

$$Y_t^{(n)} = \sum_{k=0}^{n-1} X_{t_k}(X_{t_{k+1}} - X_{t_k}) \ .$$

The trick (see any book on this) is to write

$$X_{t_k} = \frac{1}{2}(X_{t_{k+1}} + X_{t_k}) - \frac{1}{2}(X_{t_{k+1}} - X_{t_k}) \ .$$

Now,

$$(X_{t_{k+1}} + X_{t_k})(X_{t_{k+1}} - X_{t_k}) = X^2_{t_{k+1}} + X^2_{t_k} \ ,$$

so, using $t_n = T$ and $X_0 = 0$,

$$\begin{aligned}\sum_{k=0}^{n-1} X_{t_k}(X_{t_{k+1}} - X_{t_k}) &= \frac{1}{2}\sum_{k=0}^{n-1}\left(X^2_{t_{k+1}} - X^2_{t_k}\right) + \frac{1}{2}\sum_{k=0}^{n-1}\left(X_{t_{k+1}} - X_{t_k}\right)^2 \\ &= \frac{1}{2}X_T^2 + \frac{1}{2}\sum_{k=0}^{n-1}\left(X_{t_{k+1}} - X_{t_k}\right)^2\end{aligned}$$

The second term on the right is the sum of a large number of independent terms with the same distribution, and mean $\frac{1}{2}E[\Delta X_k^2] = \frac{\Delta t}{2}$. Thus, the second term is approximately $\frac{n\Delta t}{2} = \frac{T}{2}$. Letting $\Delta t \to 0$, we get

$$\int_0^T X_t dX_t = \frac{1}{2}X_t^2 - \frac{1}{2}T \ .$$

This is one of the martingales we saw earlier. The Ito integral (6.11) always gives a martingale, as we will see.

6.2.3 Other definitions of the stochastic integral give different answers:

A sensible person might suggest other approximations to (6.11). With $I_k = [t_k, t_{k+1}]$, we approximated $\int_{I_k} V(X_t)dX_t$ by $V(X_{t_k})(X_{t_{k+1}} - X_{t_k})$, which seems like the rectangle rule for ordinary integration. What would happen if we try the trapezoid rule,

$$\text{(Wrong!)} \qquad \int_{I_k} V(X_t)dX_t \approx \frac{1}{2}[V(X_{t_k}) + V(X_{t_{k+1}})](X_{t_{k+1}} - X_{t_k}) \ \ ?$$

The reader should check that in the example $V(x) = x$ above this would give

$$\text{(Wrong!)} \qquad \int_0^T X_t dX_t = \frac{1}{2}X_t^2 \ \ .$$

Also, if X_t were a differentiable function of t, with derivative $\frac{dX_t}{dt} = W(t)$, we could write

$$\text{(Wrong!)} \qquad \int_0^T X_t dX_t = \int_0^T X_t \frac{dX_t}{dt} dt = \frac{1}{2}\int_0^T \frac{dX_t^2}{dt} dt = X_T^2/2 \ .$$

From this it seems that the Ito calculus is different from ordinary calculus because the function X_t is not differentiable in the ordinary sense. The derivative, white noise, is not a function in the ordinary sense.

6.2.4 Convergence and existence of the integral (6.11):

We show that the approximation (6.12) converges to something as $\Delta t \to 0$ (really $\Delta t = T/2^k$, $k \to \infty$), assuming that V is "Lipschitz continuous": $|V(x) - V(x')| \leq C\,|x - x'|$. For example, $V(x)$ would be Lipschitz continuous if V' were a bounded function. The convergence is again "pointwise"; the event that the approximations do not converge has probability zero. As in paragraph 1.5, we compare the contributions from interval $I_k = [k\Delta t, (k+1)\Delta t]$ when we have Δt, corresponding to $n = 2^L$ subintervals, and $\Delta t/2$ corresponding to $2n = 2^{L+1}$ intervals. For Δt there is just

$$\int_{t \in I_k} V(X_t)dX_t \approx V(X_k)(X_{k+1} - X_k) \ .$$

We use the shorthand X_k for X_{t_k}, and below, $X_{k+1/2}$ for $X_{(k+1/2}\Delta t$. For $\Delta t/2$ there are two contributions:

$$\int_{t \in I_j} V(X_t)dX_t \approx V(X_k)(X_{k+1/2} - X_k) + V(X_{k+1/2})(X_{k+1} - X_{k+1/2}) \ .$$

The difference between these is

$$D_j = V(X_{k+1/2} - V(X_k))(X_{k+1} - X_{k+1/2}) \ .$$

Therefore, using the old double summation trick,

$$\begin{aligned} s_L^2 &= E\left[\left(Y_T^{(2n)} - Y_T^{(n)}\right)^2\right] \\ &= E\left[\left(\sum_{k=0}^{n-1} D_k\right)^2\right] \\ &= \sum_{j=0}^{n-1}\sum_{k=0}^{n-1} E[D_j D_k] \; . \end{aligned}$$

The terms with $j \neq k$ are zero. Suppose, for example, that $k > j$. Then $E[(X_{k+1} - X_{k+1/2}) \mid \mathcal{F}_{k+1/2}] = 0$, so $E[D_j, D_k] = 0$. When V is Lipschitz continuous,

$$E[D_k^2] \leq C^2 E[(X_{k+1} - X_{k+1/2})^2 (X_{k+1/2} - X_k)^2] = C^2 \Delta t^2/2.$$

since there are $n = 2^L$ terms, this gives $s_L^2 \leq n\Delta t \leq C^2 \Delta t/4 = C^2 T^2 2^{-L}/4$, so $\sum_L s_L < \infty$, which implies pointwise convergence.

6.2.5 How continuous are Brownian motion paths:

We know that Brownian motion paths are continuous but not differentiable. The total variation, the total distance travelled (not the net distance $|X_{t'} - X_t|$), is infinite for any interval. To understand the accuracy and convergence of approximations like (6.12), we would like some positive quantitative measure of continuity of Brownian motion paths. One positive statement is "Hölder continuity". The function $f(t)$ is Hölder continuous with exponent α if there is some C so that

$$|f(t') - f(t)| \leq C \left|t' - t\right|^{\alpha} \; ,$$

for any t and t'. Only exponents between larger than zero and not more than one are relevant. Exponent $\alpha = 1$ is for Lipschitz continuous functions. A larger α means a more regular function. Besides Brownian motion, fractals such as the Koch snowflake and the space filling curve are other examples of natural Hölder continuous functions. The function $f(t) = -1/\log(t)$ is continuous at $t = 0$ but not Hölder continuous there. The exponent $\alpha = 1/2$ seems natural for Brownian motion because (see the discussion of total variation and quadratic variation)

$$E[|X_{t'} - X_t|] \sim \left|t' - t\right|^{1/2} \; .$$

Actually, this is just slightly optimistic. It is possible to prove, using the Brownian bridge construction (upcoming) that Brownian motion paths are Hölder continuous with any positive exponent less than 1/2:

Lemma: For any positive $\alpha < 1/2$, every $T > 0$, and (almost) every Brownian motion path X_t, there is a C_X so that

$$|X_{t'} - X_t| \leq C_X \left|t' - t\right|^{\alpha} \; .$$

for all $t \leq T$ and $t' \leq T$. Furthermore, $E[C_X] < \infty$.
Remark: The proof of this lemma is really a calculation of (an upper bound for) $E[C_X]$.

6.2.6 Ito integration with nonanticipating functions:

Ito wants to integrate more general functions than $V(X_t)$ with respect to Brownian motion. For example, he might want to calculate

$$\int_0^T \left(\max_{s<t} X_s \right) dX_t \ ,$$

or the iterated integral

$$\int_0^T \left(\int_0^t X_s^2 dX_s \right) dX_t \ .$$

Therefore, we consider the more general Ito integral

$$Y_T = \int_0^T V_t dX_t \ , \tag{6.13}$$

where, for each t, V_t is measurable with respect to $\mathcal{F}_t$. Such functions are called "adapted" or "nonanticipating" or "causal" (possible subtle distinctions between these notions go unmentioned here). Nonanticipating functions are important in studying stochastic decision problems; we are supposed to make decisions at time t based on information in $\mathcal{F}_t$. Martha Stewart can explain the consequences of violating this rule, or appearing to do so. The examples above have

$$V_t = \max_{s<t} X_s$$

and

$$V_t = \int_0^t X_s^2 dX_s$$

respectively, both measurable in $\mathcal{F}_t$. Of course, V_t is a function of X also (ω in the abstract description), but as usual we do not indicate that explicitly.

We can show that integrals as general as (6.13) exist by showing that approximations

$$Y_T^{(n)} = \sum_{j=0}^{n-1} V_{t_j} (X_{j+1} - X_j) \tag{6.14}$$

converge as $\Delta t \to 0$. The argument in paragraph 2.4 works fine for this purpose if you assume that V_t is a Hölder continuous function of t (with $E[C^2] < \infty$, C being the Hölder exponent). Because we might want $V_t = X_t$ (the case of paragraph 2.4), we should allow Hölder exponents less than 1/2. As before, the difference between the Δt and $\Delta t/2$ approximations is

$$Y_T^{(2n)} - Y_T^{(n)} = \sum_{k=0}^{n-1} D_k \ ,$$

with (in the same shorthand notation)

$$D_k = \left(V_{k+1/2} - V_k\right)\left(X_{k+1} - X_{k+1/2}\right) \ .$$

Again, because V is nonanticipating, $E[D_iD_j] = 0$ if $i \neq j$. Also,

$$E[D_k^2] \leq E[C^2](\Delta t/2)^{2\alpha}\Delta t/2 \ ,$$

which proves convergence as before.

6.2.7 Further extension, the Ito isometry:

A mapping is an isometry if distances are the same before and after the mapping is applied. For example, rigid rotations of three dimensional space are isometries; the distance between a pair of points is the same before and after the transformation is applied. The formula (6.3) is the first shows that the mapping $g \mapsto Y_g$ is an isometry in the sense that if the distance between g_1 and g_2 is

$$\|g_1 - g_2\|^2 = \int_0^T (g_1(t) - g_2(t))^2 dt \ ,$$

and the distance between random variables (functions of a random variable) $Y_1(\omega)$ and $Y_2(\omega)$ is

$$\|Y_1 - Y_2\|^2 = E[(Y_1 - Y_2)^2] = \int_\Omega (Y_1(\omega) - Y_2(\omega))^2 dP(\omega) \ ,$$

then we have the isometry (which is just a restatement of (6.3))

$$\|Y_{g_1} - Y_{g_2}\|^2 = \|g_1 - g_2\|^2 \ .$$

Since the mapping is linear, this is the same (just take $g = g_1 - g_2$) as showing that

$$\|Y_g\|^2 = \|g\|^2 \ .$$

Ito showed that his stochastic integral is an isometry in the same sense. The left side is the same, and the right side is related to $\int_0^T V_t^2 dt$. The difference is that the latter integral is random. The final Ito isometry is, using $Y_t(V)$ to indicate that Y_T depends on the function V:

$$E\left[\left(Y_t(V)\right)^2\right] = \int_0^T E[V_t^2]dt \ . \qquad (6.15)$$

It is easy to verify this identity using the approximations (6.14) as usual. The approximations (6.14) might converge to something as $\Delta t \to 0$ even when V_t is not nonanticipating (i.e. anticipating?), but it is very unlikely that the limit would satisfy the Ito isometry.

The isometry formula is useful in practical calculations (see assignment 7). It also has several applications in the theory. One theoretical application is in showing that the mapping $V_t \mapsto Y_T(V)$ may be defined for any nonanticipating V so that the right side of (6.15) is finite. For any such V and any ϵ, we must find a $V^{(\epsilon)}$ so that $V^{(\epsilon)}$ is Hölder continuous in the sense we need, and so that $\int_0^T E[(V_t - V_t^{(\epsilon)})^2] \leq \epsilon$. The Ito isometry formula then shows that, if Y_T were to exist, $E[(Y_T - Y^{(\epsilon)}))^2] \leq \epsilon$, where $Y_T^{(\epsilon)}$ is the Ito integral of $V^{(\epsilon)}$. From this, it is possible to show that the $Y_T^{(\epsilon)}$ do have a limit as $\epsilon \to 0$ (in a certain sense), which is the desired Y_T.

6.2.8 Martingale property:

As a function of T, the Ito integral is a martingale. We can see this from the approximations (6.12). If we fix Δt and let T vary, it is clear that $Y_T^{(n)}$ is a martingale, since each of the increments, $V_{t_n}\left(X_{t_{n+1}} - X_{t_n}\right)$, has mean zero when projected onto functions measurable in $\mathcal{F}_{t_n}$. Actually, I'm cheating a bit here since Δt was supposed to depend on T, but hopefully the idea is clear. The Ito isometry formula is an expression of the martingale property. If Z_n is a discrete time martingale with "martingale differences" $W_n = Z_n - Z_{n-1}$, then (with the convention that $W_0 = Z_0$)

$$Z_n = \sum_{k=0}^{n} W_k \ . \tag{6.16}$$

The martingale property is that $E[W_k \mid \mathcal{F}_j] = 0$ if $k > j$. Therefore, $E[W_k W_j] = 0$ for $k \neq j$ (we may as well suppose $k > j$, why?). Thus $E[Z_n^2] = \sum_{k=0}^n E[W_k^2]$. In the Ito integral may be thought of as a continuous time version of (6.16), with $V_t dX_t$ playing the role of W_k, and the integral playing the role of the sum. Corresponding to $E[W_k W_j] = 0$, we have $E[V_s dX_s V_t dX_t] = \delta(t-s) E[V_t^2]$, which leads to the Ito isometry formula.

Chapter 7

Lecture 7

7.1 Ito Stochastic Differential Equations

7.1.1 Notation:

We switch back to the notation W_t for Brownian motion. We use X_t to denote the solution of the stochastic differential equation (SDE). When we write forward and backward equations for X_t, the independent variable will still be x. Often we work in more than one dimension. In this case, W_t may be a vector of independent Brownian motion paths. As far as possible, we will use the same notation for the one dimensional (scalar) and multidimensional cases. The solution of the Ito differential equation will be X_t. We sometimes call these "diffusions".

7.1.2 The SDE:

A stochastic differential equation is written

$$dX_t = a(X_t, t)dt + \sigma(X_t, t)dW_t \; . \tag{7.1}$$

A solution to (7.1) is a process $X_t(W)$ that is an adapted function of W ($X_t \in \mathcal{F}_t$, where $\mathcal{F}_t$ is generated by the values W_s for $s \leq t$), so that

$$X_T = X_0 + \int_0^T a(X_t, t)dt + \int_0^T \sigma(X_t, t)dW_t \; . \tag{7.2}$$

Because X_t is adapted, the Ito integral on the right of (7.2) makes sense. The term $a(X_t, t)dt$ is called the "drift" term. If $a \equiv 0$, X_t will be a martingale; any change in $E[X_t]$ is due to the drift term. The term $\sigma(X_t, t)dW_t$ is the "noise" term. The coefficient σ may be called the "diffusion" coefficient, or the "volatility" coefficient, though both of these are slight misnomers. The volatility coefficient determines the size of the small scale random motions that characterize diffusion processes. The form (7.1) is really just a shorthand for

(7.2). It is traditionally written in differential notation (dX_t, dt, dW_t) as a reminder that Ito differentials are more subtle than ordinary differentials from calculus with differentiable functions.

What separates diffusion processes from simple Brownian motions is that in diffusions the drift and volatility coefficients may depend on X and t. It might be, for example, that when X is large, its fluctuation rate is also large. This would be modeled by having $\sigma(x,t)$ being an increasing function of x.

In the multidimensional case, we might have $X_t \in R^n$. Clearly, this calls for $a(x,t) \in R^n$ also. This might be called the "drift vector" or "velocity field" or "drift field". The volatility coefficient becomes an $n \times m$ matrix, with $W_t \in R^m$ being m independent sources of noise. The case $m < n$ is called "degenerate diffusion" and arises often in applications. The case $n = m$ and σ non singular is called "nondegenerate diffusion". The mathematical character of the forward and backward equations is far more subtle for degenerate diffusions than for nondegenerate diffusions. The case $m > n$ arises in practice only by mistake.

7.1.3 Existence and uniqueness of Ito solutions:

Just as the Ito value of the stochastic integral is one of several possible values depending on details of the definition, we might expect the solution of (7.1) to be ambiguous. We will now see that this is not so as long as we use the Ito definition of the stochastic integral in (7.2). The main technical fact in the existence/uniqueness theory is a "short time contraction estimate": the mapping defined by (7.2) is a contraction for if t is small enough. Both the existence and uniqueness theorems follow quickly from this.

Suppose X_t and Y_t are two adapted stochastic processes with $X_0 = Y_0$. We define $\tilde{X}_t$ from X_t using (7.2) by

$$\tilde{X}_T = \int_{t=0}^{T} a(X_t,t)dt + \int_{t=0}^{T} \sigma(X_t,t)dW_t \ .$$

In the same way, $\tilde{Y}$ is defined from Y. We assume that a and σ are Lipshchitz continuous in the x arguments: $|a(x,t) - a(y,t)| \leq M\,|x-y|$, $|a(x,t) - a(y,t)| \leq M\,|x-y|$. The best possible constants in these inequalities are called the "Lipschitz constants" for a and σ. The mapping $X \mapsto \tilde{X}$ is a" contraction" if

$$\left\| \tilde{X} - \tilde{Y} \right\| \leq \alpha \left\| X - Y \right\| \ ,$$

for some $\alpha < 1$, that is, if the mapping shortens distances between objects by a definite ratio less than one. Of course, whether a mapping is a contraction might depend on the sense of distance, the norm $\|\cdot\|$. Because our tool is the Ito isometry formula, we use

$$\|X - Y\|_T^2 = \max_{0 \leq t \leq T} E\left[\left(X_t - Y_t\right)^2 \right] ; .$$

The contraction lemma is:
Lemma: If a and σ are Lipshcitz with Lipschitz constant M, then

$$\left\| \tilde{X} - \tilde{Y} \right\|_T^2 \leq 4M^2 T \left\| X - Y \right\|_T^2 \ . \tag{7.3}$$

For the proof, we first write

$$\tilde{X}_T - \tilde{Y}_T = \int_{t=0}^{T} \big(a(X_t, t) - a(Y_t, t) \big) dt + \int_{t=0}^{T} \big(\sigma(X_t, t) - \sigma(Y_t, t) \big) dW_t \ .$$

We have $E[(\tilde{X}_T - \tilde{Y}_T)^2] \leq 2A + 2B$ where

$$A = E \left[\left(\int_{t=0}^{T} \big(a(X_t, t) - a(Y_t, t) \big) dt \right)^2 \right] \ .$$

and

$$B = E \left[\left(\int_{t=0}^{T} \big(\sigma(X_t, t) - \sigma(Y_t, t) \big) dW_t \right)^2 \right] \ .$$

Bounding the B term is an application of the Ito isometry formula. Indeed,

$$B \leq \int_{t=0}^{T} E \left[\big(\sigma(X_t, t) - \sigma(Y_t, t) \big)^2 \right] dt \ ,$$

Using the Lipschitz continuity of σ then gives

$$B \leq M^2 T \max_{0 \leq t \leq T} E[(X_t - Y_t)^2] \ ,$$

which is the sort of bound we need.

The A term is an application of the Cauchy Schwartz inequality

$$\left(\int_{t=0}^{T} \big(a(X_t, t) - a(Y_t, t) \big) dt \right)^2 \leq \int_{t=0}^{T} \big(a(X_t, t) - a(Y_t, t) \big)^2 dt \cdot \int_{t=0}^{T} 1 dt$$

If we now use the Lipschitz continuity of a and take expectations of both sides, we get

$$A \leq M^2 T \max_{0 \leq t \leq T} E[(X_t - Y_t)^2] \ ,$$

These two inequalities prove the contraction lemma estimate (7.3).

7.1.4 Uniqueness:

The contraction inequality gives a quick proof of the uniqueness theorem. We will see that if X_0 is a random variable, then the solution up to some time T is unique. Of course, then X_T is a random variable and may be thought of as initial data for the next T time period. This give uniqueness up to time $2T$, and so on. Suppose X_t and Y_t were two solutions of (7.2). We want to argue that $E[(X_t - Y_t)^2] < \alpha E[(X_t - Y_t)^2]$ for $\alpha < 1$. This is impossible unless $[(X_t - Y_t)^2] = 0$, that is, unless $X_t = Y_t$. From (7.3), we will have $\alpha < 1$ if $T < 1/4M^2$.

The contraction lemma does not say $E[(\tilde{X}_t - \tilde{Y}_t)^2] < \alpha E[(X_t - Y_t)^2]$. To work with the information it actually gives, define $m_T = \max_{t<T} E[(X_t - Y_t)^2]$, and $\tilde{m}_T = \max_{t<T} E[(\tilde{X}_t - \tilde{Y}_t)^2]$. From the definitions, it is clear that m_t is an increasing function of t, so that (7.3) implies that $E[(\tilde{X}_t - \tilde{Y}_t)^2] \leq \alpha m_t \leq \alpha m_T$ if $T > t$. That is, (7.3) implies that $\tilde{m}_T \leq \alpha m_T$. This gives a contradiction as before: Since X and Y are solutions, we have $\tilde{m} = m$, so $\tilde{m}_T \leq \alpha m_T$ is impossible unless $m_T = 0$.

7.1.5 Existence of solutions via Picard iteration:

The contraction inequality (7.3) allows us also to show that there is an X_T satisfying (7.2), at least for $T < 1/4M^2$. You might remember this construction, Picard iteration, from a class in ordinary differential equations. The first "iterate" does not come close to satisfying the equations but just gets the ball rolling: $X_t^{(0)} = X_0$ for all $t \leq T$. This $X_t^{(0)}$ does not depend on W_t, but it will still be random if X_0 is random. For $k > 0$, the iterates are defined by

$$X_t^{(k)} = \int_{s=0}^{t} a(X_s^{(k-1)}, t)dt + \int_{s=0}^{t} \sigma(X_s^{(k-1)}, t)dW_t \; . \qquad (7.4)$$

The contraction inequality implies that the Picard iterates, $X^{(k)}$, converge as $k \to \infty$. In (7.3), take X to be $X^{(k-1)}$, and $Y = X^{(k)}$. Then $\tilde{X} = X^{(k)}$ and $\tilde{Y} = X^{(k+1)}$. If we define

$$m_T^{(k)} = \max_{0 \leq t \leq T} E\left[\left(X_t^{(k)} - X_t^{(k)}\right)^2\right] \; ,$$

and use the ideas of the previous paragraph, (7.3) gives

$$m_T^{(k+1)} \leq \alpha m_T^{(k)} \; .$$

This implies that, for any $t \leq T$, the iterates $X_t^{(k)}$ have $E[(X_t^{(k+1)} - X_t^{(k)})^2] \leq m_T^{(0)}$, which (as we saw in the previous lecture) implies that $\lim_k \to \infty X_t^{(k)}$ exists. The contraction inequality also shows (reader: think this through) that this limit, X_t satisfies (7.2) and therefore is what we are looking for.

7.1.6 Diffusions as martingales:

If the drift coefficient in (7.1) vanishes, $a(x,t) \equiv 0$, then the process X_t is a martingale. Indeed, any process $X_t = \int_0^t F_s dW_s$, with a nonanticipating F_s is a martingale. There is a very general converse to this statement. More or less (leaving out the technical details, obviously), any adapted process, X_t, with continuous sample paths,

$$P(\text{“}X_t \text{ is a continuous function of } t'') = 1 \ ,$$

has a representation in the form (7.1), except that in general, we must take σ to be a general adapted function of t, not necessarily a function of X_t only. In discrete time, a martingale, X_k, may be written as a sum of martingale differences, $Y_k = X_k - X_{k-1}$, in that $X_k = X_0 + \sum_{j=0}^{k-1} Y_j$. The Ito integral representation of the continuous time martingale X_t is a continuous time version of the representation of a discrete time martingale as a sum of martingale differences. What makes the continuous time version really different (rather than just technically different) is the unique role of Brownian motion. The proof has to construct the Brownian motion path related to X.

7.1.7 The structure of correlated gaussians:

In the multidimensional case, σ will be a matrix. We think of σdW_t as the source of noise. The several components of σdW_t may be correlated, modeling the fact that the noise terms driving the several components of X_t are correlated. The matrix σ tells us how to make correlated noises of varying strengths from uncorrelated noises of constant strength, the components of W_t. The role of σ is to correlate the noise sources and to modulate their strengths.

One often hears people referring, for example, to tow correlated Brownian motion paths, with correlation coefficient ρ. A simpler special case of this would be standard normal random variables, Z_1 and Z_2, with correlation ρ. If we suppose that (Z_1, Z_2) form a multivariate (bivariate) normal, the covariance matrix has entries $C_{11} = \text{var}(Z_1) = 1$, $C_{22} = \text{var}(Z_2) = 1$, and $C_{12} = \text{cov}(Z_1, Z_2) = \rho$. The correlation coefficient and the covariance are the same here because the variances are both one. We can make such correlated normals from uncorrelated normals in the following way. Let U_1 and U_2 be independent standard normals. Take $Z_1 = U_1$ and $Z_2 = \rho U_1 + \sqrt{1-\rho^2}U_2$. The term ρU_1 in the Z_2 formula gives the correct correlation with U_1, provided the rest of Z_2 is independent of Z_1. The term $\sqrt{1-\rho^2}U_2$ gives enough independent noise so that $\text{var}(Z_2) = 1$. In matrix form, this is

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \cdot \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \ .$$

The point is that you can make correlated standard normals from independent ones, but you need a matrix, σ.

And the σ you need is not unique. Suppose σ is an $n \times M$ matrix, and $\sigma_1 = \sigma Q$, where Q is an $m \times m$ orthogonal matrix. If U is an m vector of independent standard normals, then $Z = \sigma U$, and $Z^{(1)} = \sigma_1 U$ are each multivariate normals with the same probability distribution. That is, Z and $Z^{(1)}$ are indistinguishable if you do now know U. Applied to SDEs, this says that Z and $Z^{(1)}$ produce paths X and $X^{(1)}$ that have are indistinguishable if you do not know W. In particular, the "QR" factorization of σ^*, w(i.e. the "LQ" factorization of σ) says that we may take σ to be lower triangular. If σ is lower triangular, the components of W beyond the n^{th} all have coefficient zero. This is why it is a mistake if you have more sources of noise than components of X.

7.2 Ito's Lemma

We want to work out the first few Picard iterates in an example. This leads to a large number of stochastic integrals. We could calculate any of them in an hour or so, but we would soon long for something like the Fundamental Theorem of calculus to make the calculations mechanical. That result is called "Ito's lemma". Not only is it helpful in working with stochastic integrals and SDE's, it is also a common interview question for young potential quants. Here is the answer.

7.2.1 The Fundamental Theorem of calculus:

The following derivation of the Fundamental Theorem of ordinary calculus provides a template for the derivation of Ito's lemma. Let $V(t)$ be a differentiable function of t with $\partial_t V$ being Lipschitz continuous. The Fundamental Theorem states that (writing $\partial_t V$ for dV/dt although V depends only on t):

$$V(T) - V(0) = \int_0^T dV = \int_0^T \partial_t V(t) dt \; .$$

This exact formula follows from two approximate short time approximations, the first of which is

$$V(t + \Delta t) - V(t) = \partial_t V(t) \Delta t + O(\Delta t^2) \; .$$

The second approximation is (writing $\partial_t V(s)$ for $V'(s)$):

$$\int_t^{t+\Delta t} \partial_t V(s) ds = \partial_t V(t) \Delta t + O(\Delta t^2) \; .$$

Using our habitual notation ($\Delta t = T/n = T/2^L$, $t_k = k\Delta t$, $V_k = V(t_k)$), we have, using both approximations above,

$$\begin{aligned}
V_n - V_0 &= \sum_{k=0}^{n-1} \big(V_{k+1} - V_k \big) \\
&= \sum_{k=0}^{n-1} \big(\partial_t V(t_k) \Delta t + O(\Delta t^2) \big) \\
&= \sum_{k=0}^{n-1} \big(\int_{t_k}^{t_{k+1}} \partial_t V(t_k) dt + O(\Delta t^2) \big) \\
&= \int_0^T \partial_t V(t) dt + nO(\Delta t^2) \; .
\end{aligned}$$

Because $n\Delta t = T$, $n\Delta t^2 = TO(\Delta t) \to 0$ as $n \to \infty$.

7.2.2 The Ito dV_t:

The Fundamental Theorem may be stated $dV = \partial_t V dt$. This definition makes

$$\int_0^T dV_t = V(T) - V(0) \; . \tag{7.5}$$

We want to extend this to functions V_t that depend on W as well as t. For any adapted function, we define dV_t so that (7.5) holds. For example, if U_t is an adapted process and $V_T = \int_0^T U_t dW_t$, then $dV_t = U_t dW_t$ because that makes (7.5) hold. Ito's lemma is a statement of what makes (7.5) hold for specific adapted functions V_t.

7.2.3 First version:

Our first version of Ito's lemma is a calculation of dV_t when $V_t = V(W_t, t)$ and V and W are one dimensional. The result is

$$dV_t = \partial_W V(W_t, t) dW_t + \frac{1}{2} \partial_W^2 V(W_t, t) dt + \partial_t V dt \; . \tag{7.6}$$

What's particular to stochastic calculus is the "Ito term" $\frac{1}{2}\partial_W^2 V(W_t, t)dt$. Even if we can't guess the precise form of the term, we know something has to be there. In the special case $V_t = V(W_t)$, the $\partial_t V dt$ term is missing. The guess $dV = \partial_W V dW_t$ would give (see (7.5)) $V(t) - V(0) = \int_0^T \partial_W V dW_t$. We know this cannot be correct: the right side is a martingale while the left side is not (see assignment 5, question 1). To make the martingale integral into the non martingale answer, we have to add a dt integral, which is why some term like $\frac{1}{2}\partial_W^2 V(W_t, t)dt$ is needed. A motivation for the specific form of the Ito term is the observation that it should vanish when V is a linear function of W.

7.2.4 Derivation, short time approximations:

The derivation of Ito's lemma starts with the stochastic versions of the two short time approximations behind the Fundamental Theorem. For convenience, we drop all t subscripts and write ΔW for $W_{t+\Delta t} - W_t$. We have

$$V(W_{t+\Delta t}, t+\Delta t) - V(W,t) = \\ \partial_W V(W,t)\Delta W + \frac{1}{2}\partial_W^2 V(W,t)\Delta W^2 + \partial_t V(W,t)\Delta t + O(\Delta t^{3/2}) \; .$$

The other short time approximation is provided by assignment 7, question 3, applied to $\partial_W V$:

$$\int_t^{t+\Delta t} \partial_W V(W_s,s)dW_s = \partial_W V(W_t,t)\Delta W + \frac{1}{2}\partial_W^2 V(W,t)\big(\Delta W^2 - \Delta t\big) + O(\Delta t^{3/2}) \; .$$

For dt integrals, the result is simply

$$\int_t^{t+\Delta t} U(W_s,s)ds = U(W,t)\Delta t + O(\Delta t^{3/2} \; .$$

The error term is $O(\Delta t^{3/3})$ rather than $O(\Delta t^2)$ because W_t is not a Lipschitz continuous function of t. We conbine these approximations with a little algebra ($\Delta W^2 = \Delta t + (\Delta W^2 - \Delta t)$, which might be considered the main idea of this section) gives

$$\begin{aligned} \Delta V &= \int_t^{t+\Delta t} \partial_W V(W_s,t)dW_s \\ &\quad + \int_t^{t+\Delta t} \left(\frac{1}{2}\partial_W^2 V(W_s,s) + \partial_t V(W_s,s)\right) ds \\ &\quad + \partial_W^2 V(W,t)\big(\Delta W^2 - \Delta t\big) + O(\Delta t^{3/2}) \; . \end{aligned}$$

As with the Fundamental Theorem, we apply this with $t = t_k$ (in the habitual notation) and sum over k, giving:

$$\begin{aligned} V(T) - V(0) &= \int_0^T \partial_W V(W_t,t)dW_t \\ &\quad + \int_0^T \left(\frac{1}{2}\partial_W^2 V(W_t,t) + \partial_t V(W_t,t)\right) dt \\ &\quad + \sum_{k=0}^{n-1} \partial_W^2 V(W_k,t_k)\big(\Delta W_k^2 - \Delta t\big) + O(T\sqrt{\Delta t}) \; . \end{aligned}$$

7.2.5 The non Newtonian step:

The final step in deriving Ito's lemma has no analogue in the proof of Newton's Fundamental Theorem of calculus. We study the term

$$A = \sum_{k=0}^{n-1} \partial_W^2 V(W_k,t_k)\big(\Delta W_k^2 - \Delta t\big)$$

and show that $A \to 0$ as $\Delta t \to 0$ (actually, as $L \to \infty$ with $\Delta t = T/2^L$) almost surely. Previous experience might lead us to calculate $E[A_L^2]$. This follows a well worn path. We have the double sum expression:

$$E[A_L^2] = \frac{1}{4} \sum_{j,k} E\left[(\cdot)_j (\cdot)_k \right] \; .$$

The $j \neq k$ terms have expected value zero because (if $k > j$) $E[\Delta W_k^2 - \Delta t \mid \mathcal{F}_{t_k}] = 0$. We get a bound for the $j = k$ terms using $E[(\Delta W_k^2 - \Delta t)^2 \mid \mathcal{F}_{t_k}] = 2\Delta t^2$:

$$E\left[\partial_W^2 V(W_k, .t_k)^2 \left(\Delta W_k^2 - \Delta t\right)^2 \mid \mathcal{F}_{t_k} \right] \leq C \cdot \Delta t^2 \; .$$

Altogether, we get $E[A_L^2] \leq C\Delta t = C2^{-L}$, which implies that $A_L \to 0$ as $L \to \infty$, almost surely (see the next paragraph). This completes our proof of the first form of Ito's lemma, (7.6).

7.2.6 A Technical Detail:

Here is a proof that uses the inequalities $E[A_L^2] \leq Ce^{-\beta L}$ for some $\beta > 0$ and proves that $A_L \to 0$ as $L \to \infty$ almost surely. The proof is an easier version of an argument used in the previous lecture. As in that lecture, we start with a observation, this time that $|A_L| \to 0$ as $L \to \infty$ if $\sum_{L=1}^{\infty} |A_L| < \infty$. Also, the sum is finite almost surely if it's expected value is finite. That is, if $\sum_{L=1}^{\infty} E[|A_L|] < \infty$. Finally, the Cauchy Schwartz inequality gives $E[|A_l] \leq Ce^{-\beta L/2}$. Since this has a finite sum (over L), we get almost sure convergence $A_L \to 0$ as $L \to \infty$.

7.2.7 Integration by parts:

In ordinary Newtonian (and Leibnitzian) calculus, the integration by parts identity is a consequence of the Fundamental Theorem and facts about differentiation (the Leibnitz rule). So let it be for Ito. For instance, integration by parts might lead to

$$\int_0^T t dW_t = TW_T - \int_0^T W_t dt \; . \tag{7.7}$$

We can check whether this actually is true by taking the Ito differential of tW_t:

$$\begin{aligned} d(tW_t) &= \partial_W(tW_t)dW_t + \frac{1}{2}\partial_W^2(tW_t)dt + \partial_t(tW_t)dt \\ &= tdW_t + W_t dt \; . \end{aligned}$$

This implies that

$$TW_T = \int_0^T t dW_t + \int_0^T W_t dt \; ,$$

which is a confirmation of (7.7). We can get a more general version of the same thing if we apply the Ito differential to $f(t)g(W_t)$ (reader: do this).

7.2.8 Doing $\int W_t dW_t$ the easy way:

If Ito's lemma is to play the role of the Fundamental Theorem of calculus, it should help us calculate stochastic integrals. In ordinary calculus we calculate integrals by differentiating guesses to see which guess works. After a while, we become more systematic guessers. To compute a stochastic integral, we need to guess a function F_t so that dF_t is the integrand. A first example of this is

$$Y_T = \int_0^T W_t dW_t \; . \tag{7.8}$$

Using ordinary calculus as a clue, we might try $F_t = \frac{1}{2}W_t^2$. We calculate, using (7.6),

$$dF = \partial_W F dW + \frac{1}{2}\partial_W^2 F dt + \partial_t F = WdW + dt + 0 \; .$$

We see that we did not get the desired answer, dF is not the integrand WdW. However, it is almost right, missing by dt. To correct for this, try the more sophisticated guess $F = \frac{1}{2}W_t^2 - t$. Repeating the differentiation, we see that indeed

$$d(\frac{1}{2}W_t^2 - t) = W_t dW_t \; .$$

as desired. Ito's lemma than tells us that

$$\frac{1}{2}W_T^2 - T - (\frac{1}{2}W_T^2 - T) = \int_0^T W_t dW_t \; .$$

7.2.9 $\int W_t^2 dX_t$ the easy way:

To calculate

$$\int_0^T W_t^2 dW_t$$

we again start with the calculus guess, which this time is $F = \frac{1}{3}W_t^3$. The Ito differential of this is

$$d\frac{1}{3}W_t^3 = W_t^2 dW_t + \frac{1}{2}2W_t dt \; .$$

This differs from our integrand ($W_t^2 dW_t$) by the term $W_t dt$. We can get $W_t dt$ by differentiating $\int_0^t W_s ds$. Therefore,

$$\int_0^T W_t^2 dW_t = \frac{1}{3}W_T^3 - \int_0^T W_t dt \; .$$

If you still consider this to be a guess, you can check it by taking the differential of both sides. The left side gives $W_T^2 dW_T$. The right side gives $W_T^2 dW_T + W_T dt - W_T dt$, which is the same thing.

7.2.10 Solving an SDE:

Here is one way to solve the SDE:

$$dX_t = X_t dW_t \ , \quad X_0 = 1 \ . \tag{7.9}$$

The ordinary calculus result would be $X_T = e^{W_T}$. To see whether this satisfies (7.9), we calculate the Ito differential:

$$de^W_T = \partial_W e^W_T dW_T + \frac{1}{2}\partial_W^2 e^{W_T} dt + \partial_t e^{W_T} dt = e^W_T dW_T + \frac{1}{2} e^{W_T} dt \ .$$

The first term on the left is indeed $X_T dW_T$, so we need somehow to get rid of the second term. After some false starts, we hit on the idea to try a solution of the form $X_t = A(t)e^{W_t}$. Now the differential is

$$\begin{aligned} d\big(A(t)e^{W_t}\big) &= A(t)\partial_W e^W_t dW_t + A(t)\frac{1}{2}\partial_W^2 e^{W_t} dt + \partial_t A(t) e^{W_t} dt \\ &= A(t)e^W_t dW_t + A(t)\frac{1}{2} e^{W_t} dt + \dot{A}(t) e^{W_t} dt \ . \end{aligned}$$

The first term on the right is the desired answer $X_t dW_t$. The second and third terms will cancel if $\frac{1}{2}A + \dot{A} = 0$, i.e. if $A(t) = e^{-t/2}$. Our new guess, then, is $X_t = e^{W_t - t/2}$. We can check this with the calculation $de^{W_t - t/2} = e^{W_t - t/2} dW_t$ (because, by design, the dt terms cancel).

A consistency check is that X_t should be a martingale, because $X_T = \int_0^T X_t dW_t$, and the Ito integral always gives a martingale. We can check, for example, that $E[X_t] = 1$ for any t.

7.2.11 Differentials of functions of X_t:

The formal formulation (7.1) of an Ito SDE is in fact a relation among Ito differentials, which is precisely what (7.2) says. We can also compute $dV(X_t)$ (or even $dV(X_t, t)$, which is more complicated but not harder) using the reasoning in paragraphs 2.4 and 2.5 above. I will breeze through the argument, commenting only on the differences. Some of the details are left to assignment 8. We can calculate

$$\Delta V(X) = \partial_X V(x_t)\Delta X_t + \frac{1}{2}\partial_X^2 V(X)\Delta X^2 + O(\Delta t^{3/2}) \ .$$

Also

$$\int_t^{t+\Delta t} \partial_X V(X_s) dX_s = \partial_X V(X_t)\Delta X_t + \frac{1}{2}\partial_X^2 V(X_t)\big(\Delta X^2 - \sigma(X_t)^2 \Delta t\big) + O(\Delta t^{3/2}) \ .$$

The new feature is that $E[\Delta X^2] = \sigma(X_t)^2 \Delta t + O(\Delta t^{3/2})$. After this, the derivation proceeds as before, eventually giving

$$dV(X_t) = \partial_X V(X_t) dX_t + \frac{1}{2}\partial_X^2 V(X_t)\sigma(X_t)^2 dt \ . \tag{7.10}$$

7.2.12 The "Ito rule" $dW^2 = dt$:

The first version of Ito's lemma can be summarized as using Taylor series calculations and neglecting all terms of higher than first order except for dW_t^2, which we replace by dt. You might think this is based on the approximation $\Delta W^2 \approx \Delta t$ for small Δt. The real story is a little more involved. The relative accuracy of the approximation $\Delta W^2 \approx \Delta t$ does not improve as $\Delta t \to 0$. Both sides go to zero, and at the same rate, but they do not get closer to each other in relative terms. In fact, the expected error, $E[|\Delta W^2 - \Delta t|]$, is also of order Δt. If $\Delta t = .1$ then ΔW^2 is just as likely to be .2 as .1, not really a useful approximation. The origin of Ito's rule is that ΔW^2 and Δt have the same expected value. For that reason, if we add up m ΔW^2 values, we are likely to get a number close to $m\Delta t$ if m is large. We might say $\int_a^b dW^2 = \int_a^b dt$, thinking that each side is make up of a large number (an infinite number) of tiny ΔW^2 or Δt values. Remember that for any Q, the Ito dQ is what you have to integrate to get Q. Integrating dW^2 gives the same result as integrating dt.

7.2.13 Quadratic variation:

The informal ideas of the preceding paragraph may be fleshed out using the "quadratic variation" of a process. We already discussed the quadratic variation of Brownian motion. For a general stochastic process, X_t, the quadratic variation is

$$\left\langle X \right\rangle_t = \lim_{\Delta t \to 0} \sum_{k=0}^{n} \bigl(X_{k+1} - X_k\bigr)^2 \; . \tag{7.11}$$

If we apply the approximation from assignment 8, question 2b, we get

$$\sum_{k=0}^{n} \bigl(X_{k+1} - X_k\bigr)^2 = \left(\sum_{k=0}^{n} \sigma(X_k, t_k)^2 \Delta X_k^2 \right) + O(n\Delta t^3/2) \; .$$

Our usual trick is to use $\Delta X_k^2 = \Delta t + (\Delta X^2 - \Delta t)$ to write the last sum as an approximation of a dt integral plus something with mean zero that does not add up to much. The result is

$$\left\langle X \right\rangle_t = \int_0^t \sigma^2(X_s, t) ds \; .$$

In particular,

$$d \left\langle X \right\rangle_t = \sigma^2(X_t) dt \; .$$

Ito's lemma for X_t satisfying the SDE (7.1) may be written

$$dV(X_t) = \partial_X V(X_t) dX_t + \frac{1}{2} \partial_X^2 V(X_t) d \left\langle X \right\rangle_t \; . \tag{7.12}$$

7.2.14 Geometric Brownian motion again:

Here is another way to find the solution of $dX_t = X_t dt$. Since we expect X_t to be an exponential, we calculate the SDE satisfied by $Y_t = \log(X_t)$. Ito's lemma in the form (7.12) allows us to calculate

$$\begin{aligned}
dY_t &= \partial_X \log(X_t) dX_t + \frac{1}{2}\partial_X^2 \log(X_t) X_t^2 dt \\
&= \frac{1}{X_t} X_t dW_t + \frac{1}{2}\left(-\frac{1}{X_t^2}\right) X_t^2 dt \\
dY_t &= dW_t - \frac{1}{2} dt \, .
\end{aligned}$$

This gives $Y_t = Y_0 + W_t - \frac{t}{2}$. Since $X_t = e^{Y_t}$, we get $X_t = X_0 e^{W_t - t/2}$, as before.

7.2.15 Remarks on the solution:

The solution $X_t = X_0 e^{W_t - t/2}$ provides some insight into how martingales can behave and the importance of rare events. We know that Brownian motion paths W_t are on the order of $\sqrt{T}$. Therefore for large t, the exponent is $W_t - \frac{t}{2} \approx -t/2$. That is, nearly all (not almost all) geometric Brownian motion paths are exponentially small for any particular large t. Nevertheless, since X_t is a martingale, $E[X_t] = 1$. Those rare paths with $W_t > t/2$ are just big enough and just likely enough to save $E[X_t]$ from being exponentially small. For the record, $P(W_t > t/2) < e^{-t/8}$, is very small (about 1/1000 for $t = 100$). This means that if you simulate, say, 500 paths, there is a pretty good chance that none of them is as big as the mean. Monte Carlo simulation is very unreliable in such cases.

Chapter 8

Lecture 8

8.1 Girsanov's theorem

Girsanov's theorem relates solutions of Ito differential equations with different drifts. It is also an example of an interesting possibility in probability, computing the expected value of one random variable by using a random variable with a different probability measure. In Monte Carlo, this is called "importance sampling", and is used to in making accurate estimates of very small probabilities.

8.1.1 Probability densities and Lebesgue measure:

For Brownian motion, we gave a probability measure but not a probability density. For a simple gaussian random variable $X \sim \mathcal{N}(0,1)$ we instead give a probability density, $u(x) = \sqrt{1}\sqrt{2\pi}e^{-x^2/2}$. This is possible because there already is a measure on the probability space $\Omega = R$, Lebesgue measure. When we write $E[V(X)] = \int V(x)u(x)dx$, the dx refers to integration with respect to Lebesgue measure, which is "uniform measure", $\lambda((a,b)) = b - a$ (here $\lambda(A)$ is the Lebesgue measure of A, applied to $A = (a,b)$). It is also possible to define the "standard normal probability measure", P. This is $P(A) = \int_A u(x)dx$. We then have $E[V(X)] = \int_R V(x)dP(x)$ In abstract probability we describe this situation by saying that the gaussian measure P (possibly written dP) is "absolutely continuous" with respect to Lebesgue measure, λ (possibly written dx). The function $u(x)$ is the density of dP with respect to dx, sometimes written $u(x) = \frac{dP}{dx}$. The formal ratio $\frac{dP}{dx}$ is also called the "Radon Nikodym derivative" of the gaussian measure dP with respect to Lebesgue measure dx.

8.1.2 The Radon Nikodym derivative:

A more abstract version of this situation is that there is a probability space Ω, a $\sigma-$algebra of sets $\mathcal{F}$, and two measures $dP(\omega)$, and $dQ(\omega)$. We will suppose that both are probability measures, though this is not necessary; dQ was Lebesgue measure in the previous paragraph. We say that $L(\omega)$ is the

Radon Nikodym derivative of dP with respect to dQ and write $L(\omega) = \frac{dP(\omega)}{dQ(\omega)}$ if $P(A) = \int_{\omega \in A} L(\omega) dQ(\omega)$. We use L here and below (instead of u as above), because the Radon Nikodym derivative is closely related to what statisticians call the "likelihood ratio". The definition of L is the same as saying that for any function, $V(\omega)$,

$$E_P[V(\omega)] = \int_\Omega V(\omega) dP(\omega) = \int_\Omega V(\omega) L(\omega) dQ(\omega) = E_Q[V(\omega) u(\omega)] \ . \quad (8.1)$$

Following an earlier custom, we write $E_P[\cdot]$ for expectation with respect to the probability measure P.

8.1.3 Radon Nikodym derivative as likelihood ratio:

If X_0 and X_1 are two random variables with densities $u_0(x)$ and $u_1(x)$, then they have probability measures $dP(x) = u_0(x)dx$ and $dQ(x) = u_1(x)dx$ respectively. Therefore, $L(x) = dP(x)/dQ(x) = u_0(x)dx/u_1(x)dx = u_0(x)/u_1(x)$. The Radon Nikodym derivative is the ratio of the probability densities. Statisticians often call probability densities "likelihoods", particularly when thinking of them as a function of some parameter (the mean, variance, etc.). The ratio of probability densities becomes the "likelihood ratio", L. Though our canceling dx from the numerator and denominator is not rigorous, the formula $L = u_0/u_1$ is easy to check in the integral definition (8.1), as in the following example.

8.1.4 Example of one dimensional gaussians:

Suppose the measure P corresponds to a standard normal, $u_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, and Q to a $\mathcal{N}(\mu, 1)$ random variable, $u_\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}$. The Radon Nikodym derivative dP/dQ is given by $L(x) = u_0(x)/u_\mu(x) = e^{-\mu x + \mu^2/2}$. We can verify this by checking, using the standard gaussian integration formulas for expectation values, that

$$\begin{aligned} E_0[V(X)] &= \frac{1}{\sqrt{2\pi}} \int_R V(x) e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_R V(x) L(x) e^{-(x-\mu)^2/2} dx \\ &= E_\mu[V(X) L(X)] \ . \end{aligned}$$

8.1.5 Absolutely continuous measures:

It might seem that it is easy to calculate the Radon Nikodym derivative, and it generally is, provided it exists. Given probability measures P and Q on the same space, Ω, with the same measurable sets, $\mathcal{F}$, there might be an event, A, that has probability zero in the Q probability but positive P probability. In that case, it is impossible to have

$$P(A) = \int_A L(\omega) dQ(\omega) \ ,$$

because the left side must be zero if $Q(A) = 0$. The Radon Nikodym theorem says that this is the only thing that can go wrong.
Theorem: If $Q(A) = 0$ implies $P(A) = 0$ for every measurable event, A, then there is a Radon Nikodym derivative, $L(\omega)$, that represents P in terms of Q.

8.1.6 Examples in finite dimensions:

These illustrate the possibility that one measure may not be absolutely continuous with respect to another, but they do not give much intuition about the subtlety of absolute continuity applied to diffusions. As one example, consider two one dimensional random variables, the standard exponential and the standard normal. The exponential random variable has probability density $u_1(x) = 0$ if $x < 0$ and $u(x) = e^{-x}$ for $x > 0$. The standard normal has density $u_0(x)$ as above. The event $A = \{$"$x < 0$"$\} = (-\infty, 0)$ has Q probability (standard normal probability) $Q(A) = .5$ but P probability $P(A) = 0$, since the exponential random variable is never negative. In this situation we say that the gaussian measure is not absolutely continuous with respect to the exponential measure. On the other hand, the exponential measure is absolutely continuous with respect to Gaussian measure: an event whose gaussian probability is zero also has exponential probability zero.

As another example, suppose we choose the random variable X by first tossing a "fair" coin. If the toss is H (probability p), we set $X = 0$. If the toss is T (probability $q = 1 - p$), we make X to be a standard normal. The probability density for the resulting random variable is

$$u(x) = p\delta(x) + frac1 - p\sqrt{2\pi}e^{-x^2/2} \ .$$

This density is a "mixture" of the delta function and the standard normal density. It is not absolutely continuous with respect to the "pure" standard normal measure because the event $X = 0$ has probability $p > 0$ for u but probability zero for the standard normal alone. Here, the lack of absolute continuity is caused by a concentration of probability rather than the density being zero in a large region ($x < 0$ above).

8.1.7 Cantor measure:

This shows that it is possible to concentrate probability in a set of Lebesgue measure zero without concentrating it at a point as in the delta measure. The Cantor measure (after Georg Cantor, a late nineteenth century German mathematician) is defined on the interval $0 \leq x \leq 1$ by throwing out all the "middle thirds" and concentrating the measure on what remains, which is called the

Cantor set, C. To determine whether an $x \in [0,1]$ is in C, we give it's representation base 3: $x = 0.a_1a_2a_3\cdots$, where each a_k is one of the numbers 0, 1, or 2, and

$$x = \sum_{k=1}^{\infty} 3^{-k} a_k \ .$$

The ordinary decimal representation is the same thing with 3 replaced by 10. For example,

$$\frac{1}{4} = 0.020202\cdots \quad , \quad \frac{1}{5} = .012012\cdots \ .$$

The Cantor set is the set of numbers that have $a_k \neq 1$ for all k. The condition $a_1 \neq 1$ rules out all numbers $x \in (\frac{1}{3}, 23)$, the middle third of $(0,1)$. The middle thirds of the first third and third third are ruled out by the condition $a_2 \neq 0$; numbers of the form $0.01a_3, a_4\cdots$ form the interval $(\frac{1}{9}, \frac{2}{9})$, which is the middle third of the first third, and numbers of the form $0.21a_3, a_4\cdots$ form the interval $(\frac{7}{9}, \frac{8}{9})$, which is the middle third of the third third.

With respect to uniform measure (Lebesgue measure) in the unit interval, the Cantor set has probability zero. The probability that x will be thrown out because $a_1 = 1$ is $\frac{1}{3}$. If x is spared (probability $\frac{2}{3}$), it is thrown out because $a_2 = 1$ again with probability $\frac{1}{3}$. The probability of being spared k times is $\frac{2}{3}^k \to 0$ as $k \to \infty$. To be in C, x must be spared infinitely many times, an event of Lebesgue measure zero.

To define the Cantor measure, we need to give $P_C(A)$ for any $A \subseteq C$. For each $x \in A$ we define a $y \in (0,1)$ by giving the *base 2* binary expansion $y = 0.b_1, b_2, b_3\cdots$, where $b_k = 0$ if $a_k = 0$ and $b_k = 1$ if $a_k = 2$. That is

$$y = \sum_{k \in T(x)} 2^{-k} \quad , \quad \text{where } T(x) = \{k \mid a_k = 2\} \ .$$

The set of all such y coming from an $x \in A$ is called B. The Cantor measure of A will be the ordinary Lebesgue (uniform) measure of B. If $A \subseteq [0,1]$, we define $P_C(A) = P_C(A \cap C)$. We can think of the Cantor measure as coming from an infinite sequence of cut and squeeze steps. First we cut the unit interval in the middle and squeeze the first half into the first third and the second half to the third third. This gives the first and third thirds probability $\frac{1}{2}$ each and the middle third probability zero. We then cut and squeeze out the middle third of the first and third thirds, giving each of the 4 remaining ninths measure $\frac{1}{4}$, and so on.

The Cantor set and Cantor measure provide illustrate some things that can go wrong in measure theory. . . .

8.1.8 Alternative descriptions of a random variable:

It often happens that we can describe a random $X \in \mathcal{S}$ either as a random variable in its own right by giving a probability measure, Q, on $\mathcal{S}$, or as a function of another random variable $\omega \in \Omega$ with measure P. If we have a function $V(x)$ and we want the expected value, $E[V(X)]$, we may calculate it either as $\int_{\mathcal{S}} V(x)dQ(x)$ or as $\int_{\Omega} V(X(\omega)dP(\omega)$. Of course, the function $X(\omega)$ and the measure P determine the measure Q. Nevertheless, we sometimes can make use of a direct description of Q without reference to P and Ω. Girsanov's theorem is about the measure in path space defined by the solution of a stochastic differential equation. In this case, Ω is the space of Brownian motion paths and $X(W)$ is the solution of the SDE for Brownian motion path W, which plays the role of ω here. To state Girsanov's theorem, we have to be able to understand the X measure without reference to the underlying W.

8.1.9 A one dimensional mapping example:

Suppose $\Omega = (0, 1]$ and P is uniform measure, leaving out the point zero for simplicity. For each $\omega \in (0, 1]$ we define $X(\omega) = -\ln(\omega)$ ($\ln(\omega)$ is the log base e). This is a $1 - 1$ transformation; there is a unique $X \geq 0$ for each $\omega \in (0, 1]$, and vice versa. If $V(x) = x^2$, we could evaluate $E[V(X)]$ as the integral $\int_0^1 \ln(\omega)^2 d\omega$.

The other way is to find the PDF for X directly. Since $X \geq 0$, this density is zero for $x < 0$. We call it $u(x)$ and find it from the relation

$$\begin{aligned}
u(x)dx &= P(x \leq X \leq x + dx) \\
&= P(x < -\ln(\omega) < x + dx) \\
&= P(-x - dx < \ln(\omega) < -x) \\
&= P(e^{-x}e^{-dx} < \omega < e^{-x}) \\
&= P(e^{-x} - dxe^{-x} < \omega < e^{-x}) \\
&= dxe^{-x} \ .
\end{aligned}$$

The last line is because ω is uniformly distributed in $(0, 1]$ so the probability of being in any interval (a, b) is $b - a$, here with $a = e^{-x} - dxe^{-x}$ and $b = e^{-x}$. The conclusion is that $u(x) = e^{-x}$, which is to say that X is a standard exponential random variable. Now we can calculate the same expected value as

$$E[X^2] = \int_{x=0}^{\infty} x^2 e^{-x} dx = \int_{\omega=0}^{1} \ln(\omega)^2 d\omega \ .$$

The P measure is uniform measure on $(0, 1]$. The Q measure is standard exponential measure. The mapping is $X(\omega) = -\ln(\omega)$.

8.1.10 Distinguishing random variables:

Suppose we have probability measures P and Q on the same space, $\mathcal{S}$. A "sample" will be a random variable $X \in \mathcal{S}$ with either the P or Q probability measures. One of the main questions in statistics is finding statistical tests to determine whether X was drawn from the P or Q populations, i.e., which of P or Q describes X. A "hypothesis test" is a decomposition of $\mathcal{S}$ into two sets, here called A_P and A_Q (A_P and A_Q disjoint, $A_P \cup A_Q = \mathcal{S}$). The hypothesis test based on this decomposition reports $X \sim P$ if $X \in A_P$ and $X \sim Q$ if $X \in A_Q$. Generally speaking, in statistics your hypothesis test conclusions are not certain, but hold with a certain (hopefully high) likelihood.

Suppose there is an event $A \subseteq \mathcal{S}$ so that $P(A) > 0$ but $Q(A) = 0$. If we use this set for the hypothesis test, (taking $A_P = A$ and $A_Q = \mathcal{S} - A$), then whenever our hypothesis test reports P, it must be correct. In statisticians' language, there is a hypothesis test with zero type II error. This shows that the possibility of an (in some respects) infallible hypothesis test is equivalent to absolute continuity or lack of absolute continuity of measures. If there is a procedure that sometimes knows with 100% certainty that X was drawn from Q rather than P, then Q is not absolutely continuous with respect to P. For example, suppose $\mathcal{S} = \mathcal{R}$, Q is the standard normal measure, P is the exponential, and $A = (-\infty, 0)$. If $X \in A$ we know X came from the gaussian measure because the exponential probability of A is zero.

If measure Q is not absolutely continuous with respect to measure P, it is common that the two measures are "completely singular" with respect to each other. This means that there is a partition with $P(A_P) = Q(A_Q) = 1$, and therefore $Q(A_P) = 0 = P(A_Q)$. If measures P and Q are completely singular then there is a hypothesis test that is right 100% of the time. For example, if P is the standard normal measure and $Q = \delta$ is a delta measure corresponding to $X = 0$ with probability 1, then we take A_P to be all real numbers except zero and $A_Q = \{0\}$. The hypothesis test is to say "normal" if $X \neq 0$ and "δ" if $X = 0$. If the only choices are standard normal or delta, you can never be wrong doing this.

8.1.11 Absolute continuity of diffusion measures:

It is possible to distinguish diffusions with different σ in this way. The main fact is that $\langle X \rangle_T = \int_0^T \sigma(X_t, t)^2 dt$. If we know everything about the path, we will be able to compute ($\Delta t = T/n$, $t_k = k\Delta t$):

$$\langle X \rangle_T = \lim_{\Delta t \to 0} \sum_{k=0}^{n-1} \left(X_{t_{k+1}} - X_{t_k} \right)^2 \tag{8.2}$$

Suppose we are trying to guess whether X satisfies $dX = a_0(X_t, t)dt + \sigma_0(X_t, t)dW_t$ or $dX = a_1(X_t, t)dt + \sigma_1(X_t, t)dW_t$. We compute the quadratic variation for our path X_t (8.2) and see whether it is equal to $\int_0^T \sigma_0(X_t, t)^2 dt$ or $\int_0^T \sigma_1(X_t, t)^2 dt$. If $\sigma_0^2 \neq \sigma_1^2$, it is impossible for both to be correct (but for the unlikely possibility that $\sigma_0 = -\sigma_1$). This proves the negative part of Girsanov's theorem:

Theorem: If $\sigma_0^2 \neq \sigma_1$, then the measures corresponding to stochastic processes

$dX = a_0 dt + \sigma_0 dW$ and $dX = a_1 dt + \sigma_1 dW$ are completely singular with respect to each other.

8.1.12 Likelihood ratio for SDE solutions:

The positive part of Girsanov's theorem is about the measures for SDE solutions with the same σ but different drift, a. Suppose we have $dX = a_0(X,t) + \sigma(X,t)dW_t$ and $dX = a_1(X,t) + \sigma(X,t)dW_t$, which determine measures in path space $dP_0(X)$ and $dP_1(X)$ respectively. The theorem states that the two measures are absolutely continuous with respect to each other and gives a formula, the Girsanov formula, for the likelihood ratio, or Radon Nikodym derivative, $L(X)$. To be technically more precise, we consider a time T and paths up to time T. Then $L(X)$ is a function of the whole path up to time T.

Our strategy for finding the Girsanov formula is to find the formula for $\overline{L}_{\Delta t}$, the likelihood ratio for the forward Euler approximations to the two processes. The limit $L(X) = \lim_{\Delta t \to 0} \overline{L}_{\Delta t}$ will then be clear. There are probably some technicalities needed for a complete mathematical proof, but I will not dwell on them (an understatement).

8.1.13 Multiplying conditional probability densities:

This is a reminder of rule of conditional probability density density that will be used in the following paragraph. Suppose we first choose a random variable, X, from the probability density $u(x)$, then choose the random variable Y from the conditional density $v_1(y \mid X)$. The resulting pair (X,Y) has joint PDF $U(x,y) = u(x)v(y \mid x)$. If we then choose Z from the density $v_2(z \mid Y)$, the PDF for the triple (X,Y,Z) will be $U(x,y,z) = u(x)v_1(y \mid x)v + 2(z \mid y)$. This is a version of a rule we used for Markov chains: we multiply the conditional probabilities (transition probabilities) to get the joint probability of the path. Here the path is the triple (X,Y,Z), but clearly it could be longer.

8.1.14 Measure for the forward Euler method:

Our standard notation is $\Delta t = T/n$, $t_k = k\Delta t$, $\Delta W_k = W_{k+1} - W_k$, and $\overline{X}_k \approx X_{t_k}$. The approximation is

$$\overline{X}_{k+1} = \overline{X}_k + a(\overline{X}_k, t_k)\Delta t + \sigma(\overline{X}_k, t_k)\Delta W_k \ . \tag{8.3}$$

We want an expression for the joint PDF, $U(x_1, \ldots, x_n)$, of the n random variables $\overline{X}_1$, …, $\overline{X}_n$. The conditional probability density for $\overline{X}_k + 1$ conditioned on $\mathcal{F}_k$ actually depends only on $\overline{X}_k$ and t_k. We call it $u(x_{k+1} \mid x_k; t_k, \Delta t)$. The semicolon separates the conditioning variable, x_k, from the other arguments, t_k and Δt. As in the previous paragraph, this is built up by multiplying the conditional probability densities:

$$U(x_1,\ldots,x_n) = \prod_{k=0}^{n-1} u(x_{k+1} \mid x_k, t_k, \Delta t) \; . \tag{8.4}$$

For simplicity we suppose that $X_0 = x_0$ is specified and not random. If X_0 were random with probability density $u_0(x_0)$, we would multiply U by this factor.

The big PDF, U, is the PDF for the approximate path, $\overline{X}$, up to time T. Because time is discrete, this approximate path consists of the n values, x_k. We follow the convention of using lower case letters, x_k to be the variables corresponding to the random variables X_k. Suppose, for example, we want to know the probability that the approximate path is less than r for all $t_k \leq T$. This is the event $A = \{x_k \leq r, 1 \leq k \leq n\}$, so it's probability is given by

$$\int_{x_1=-\infty}^{x_1=r} \cdots \int_{x_n=-\infty}^{x_n=r} U(x_1,\ldots,x_n) dx_1 \cdots dx_n = \int_A U(\vec{x}) d\vec{x} \; ,$$

using the notation $\vec{x} = (x_1,\ldots,x_n) \in R^n$ for the n numbers representing the discrete path.

Particularly when Δt is small, the X_k will be strongly correlated with each other, though none is entirely determined by the others. It does not make sense to say the x_k are correlated because they are not random variables. It is true, as we are about to see, that U is very small if x_k is far from the neighboring values x_{k-1} and x_{k+1}, corresponding to the fact that is is unlikely for X_k to be far from the values X_{k-1} and X_{k+1}.

After this buildup, here is the calculation. From the forward Euler formula (8.3), it is clear that conditioned on $\overline{X}_k$, $\overline{X}_{k+1}$ is a gaussian random variable with mean $\overline{X}_k + a(\overline{X}_k, t_k)\Delta t$ and variance $\sigma(\overline{X}_k, t_k)^2 \Delta t$. Conditioned on $\mathcal{F}_k$, the only thing random in (8.3) is ΔW_k, which is gaussian with mean zero and variance Δt. Therefore, the formula for $u(x_{k+1} \mid x_k; t_k, \Delta t)$ is just the gaussian formula for x_{k+1} with the appropriate mean and variance:

$$u(x_{k+1} \mid x_k; t_k, \Delta t) = \frac{1}{\sqrt{2\pi\sigma(x_k,t_k)^2\Delta t}} \exp\left(\frac{-(x_{k+1} - x_k - \Delta t a(x_k,t_k))^2}{2\sigma(x_k,t_k)^2 \Delta t} \right) \quad .$$