Stochastic Calculus Notes, Lecture 1

Last modified September 12, 2004

1 Overture

1.1. Introduction: The term *stochastic* means "random". Because it usually occurs together with "process" (stochastic process), it makes people think of something something random that changes in a random way over time. The term *calculus* refers to ways to calculate things or find things that can be calculated (e.g. derivatives in the differential calculus). Stochastic calculus is the study of stochastic processes through a collection of powerful ways to calculate things. Whenever we have a question about the behavior of a stochastic process, we will try to find an expected value or probability that we can calculate that answers our question.

1.2. Organization: We start in the *discrete* setting in which there is a finite or countable (definitions below) set of possible outcomes. The tools are summations and matrix multiplication. The main concepts can be displayed clearly and concretely in this setting. We then move to continuous processes in continuous time where things are calculated using integrals, either ordinary integrals in R^n or abstract integrals in probability space. It is impossible (and beside the point if it were possible) to treat these matters with full mathematical rigor in these notes. The reader should get enough to distinguish mathematical right from wrong in cases that occur in practical applications.

1.3. Backward and forward equations: Backward equations and forward equations are perhaps the most useful tools for getting information about stochastic processes. Roughly speaking, there is some number, f, that we want to know. For example f could be the expected value of a portfolio after following a proposed trading strategy. Rather than compute f directly, we define an array of related expected values, $f(x,t)$. The *tower property* implies relationships, backward equations or forward equations, among these values that allow us to compute some of them in terms of others. Proceeding from the few known values (*initial conditions* and *boundary conditions*), we eventually find the f we first wanted. For discrete time and space, the equations are matrix equations or recurrence relations. For continuous time and space, they are partial differential equations of *diffusion* type.

1.4. Diffusions and Ito calculus: The Ito calculus is a tool for studying continuous stochastic processes in continuous time. If $X(t)$ is a differentiable function of time, then $\Delta X = X(t+\Delta t) - X(t)$ is of the order of[1] Δt. Therefore $\Delta f(X(t)) = f(X(t+\Delta t)) - f(X(t)) \approx f'\Delta X$ to this accuracy. For an Ito process, ΔX is of the order of $\sqrt{\Delta t}$, so $\Delta f \approx f'\Delta X + \frac{1}{2}f''\Delta X^2$ has an error

[1] This means that there is a C so that $|X(t+\Delta t - X(t)| \leq C\,|\Delta|$ for small Δt.

smaller than Δt. In the special case where $X(t)$ is Brownian motion, it is often permissible (and the basis of the Ito calculus) to replace ΔX^2 by its mean value, Δt.

2 Discrete probability

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 problems but it is just the thing for describing stochastic processes and decision problems under incomplete information.

2.1. Probability space: Do an "experiment" or "trial", get an "outcome", ω. The set of all possible outcomes is Ω, which is the *probability space*. The Ω is *discrete* if it is finite or countable (able to be listed in a single infinite numbered list). The outcome ω is often called a *random variable*. I avoid that term because I (and most other people) want to call functions $X(\omega)$ random variables, see below.

2.2. Probability: 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 actually to 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 means something that is certain to happen while probability 0 is for something that cannot happen. "Probability zero $\Rightarrow$ impossible" is only strictly true for discrete probability.

2.3. Event: An *event* is 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

$$P(A) = \sum_{\omega \in A} P(\omega) \; . \tag{1}$$

Usually, we specify an event in some way other than listing all the outcomes in it (see below). 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.

2.4. Countable and uncountable (technical detail): A probability space (or

any set) that is not countable is called "uncountable". This distinction was formalized by the late nineteenth century mathematician Georg Cantor, who showed that the set of (real) numbers in the interval $[0, 1]$ is not countable. Under the uniform probability density, $P(\omega) = 0$ for any $\omega \in [0, 1]$. It is hard to imagine that the probability formula (1) is useful in this case, since every term in the sum is zero. The difference between continuous and discrete probability is the difference between integrals and sums.

2.5. 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 $\#(\Omega) = |\Omega| = 16$. We suppose that 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

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

2.6. Set operations: Events are 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 $A \cap B$, because the outcomes that make both A and B happen are those that are in both events. 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. 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 $A \cap B = \emptyset$. Event A is contained in event B, $A \subseteq B$, if every outcome in A is also in B. For example, if the event A is as above and B is the event that the first toss is H, then $A \subseteq B$.

2.7. Basic facts: Each of these facts is a consequence of the representation $P(A) = \sum_{\omega \in A} P(\omega)$. First $P(A) \leq P(B)$ if $A \subseteq B$. Also, $P(A) + P(B) = P(A \cup B)$ if $P(A \cap B) = 0$, but not otherwise. If $P(\omega \neq 0$ for all $\omega \in \Omega$, then $P(A \cap B) = 0$ only wehn A and B are distoint. Clearly, $P(A) + P(A^c) = P(\Omega) = 1$.

2.8. Conditional probability: The probability of outcome A given that B has occurred is the *conditional probability* (read "the probability of A given B",

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \ . \tag{2}$$

This is the fraction 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 $P(A \cap B) = P(A \mid B)P(B)$.

2.9. Independence: Events A and B are *independent* if $P(A \mid B) = P(A)$. That is, knowing whether or not B occurred does not change the probability of A. In view of Bayes' rule, this is expressed as

$$P(A \cap B) = P(A) \cdot P(B) \ . \tag{3}$$

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

2.10. Working with conditional probability: Let us fix the event B, and discuss the conditional probability $\widetilde{P}(\omega) = P(\omega \mid B)$, which also is a probability (assuming $P(B) > 0$). There are two slightly different ways to discuss $\widetilde{P}$. One way is to take B to be the probability space and define

$$\widetilde{P}(\omega) = \frac{P(\omega)}{P(B)}$$

for all $\omega \in B$. Since B is the probability space for $\widetilde{P}$, we do not have to define $\widetilde{P}$ for $\omega \notin B$. This $\widetilde{P}$ is a probability because $\widetilde{P}(\omega) \geq 0$ for all $\omega \in B$ and $\sum_{\omega \in B} \widetilde{P}(\omega) = 1$. The other way is to keep Ω as the probability space and set the conditional probabilities to zero for $\omega \notin B$. If we know the event B happened, then the probability of an outcome not in B is zero.

$$P(\omega \mid B) = \begin{cases} \frac{P(\omega)}{P(B)} & \text{for } \omega \in B, \\ 0 & \text{for } \omega \notin B. \end{cases} \tag{4}$$

Either way, we restrict to outcomes in B and "renormalize" the probabilities by dividing by $P(B)$ so that they again sum to one. Note that (4) is just the general conditional probability formula (2) applied to the event $A = \{\omega\}$.

We can condition a second time by conditioning $\widetilde{P}$ on another event, C. It seems natural that $\widetilde{P}(\omega \mid C)$, which is the conditional probability of ω given that

C, occurred given that B occurred, should be be the P conditional probability of ω given that both B and C occurred. Bayes' rule verifies this intuition:

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

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. This *tower property* underlies the many recurrence relations that allow us to get answers in practical situations.

2.11. 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* $\mathcal{F}$.

2.12. Example 1 of an $\mathcal{F}$: Suppose we learn the outcomes of the first two tosses. One event measurable in $\mathcal{F}$ is (with some abuse of notation)

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

An example of an event not determined by this $\mathcal{F}$ is the event of no more than one H:

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

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

2.13. Example 2 of an $\mathcal{F}$: 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 measurable in this algebra, such as "less than 3 H", but, in some sense, the events listed *generate* the algebra.

2.14. $\sigma-$algebra: An algebra of sets is a $\sigma-$algebra (pronounced "sigma algebra") if it is *closed under countable intersections*, which means the following. Suppose $A_n \in \mathcal{F}$ is a countable family of events measurable in $\mathcal{F}$, and $A = \cap_n A_n$ is the set of outcomes in all of the A_n, then $A \in \mathcal{F}$, too. The reader can check that an algebra closed under countable intersections is also closed under countable unions, and conversely. An algebra is automatically a $\sigma-$algebra if Ω is finite. If Ω is infinite, an algebra might or might not be a $\sigma-$algebra.[2] In a $\sigma-$algebra, it is possible to take limits of infinite sequences of events, just as it is possible to take limits of sequences of real numbers. We will never (again) refer to an algebra of events that is not a $\sigma-$algebra.

2.15. Terminology: What we call "outcome" is usually called "random variable". I did not use this terminology 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: $F(x) = \sum_{X(\omega) \leq x} P(\omega)$.

2.16. 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 $A_x = \{\omega \mid X(\omega) \leq x\}$ then $P(X \leq x) = P(A_x)$. For example, if there are

[2] Let Ω be the set of integers and $A \in \mathcal{F}$ if A is finite or A^c is finite. This $\mathcal{F}$ is an algebra (check), but not a $\sigma-$algebra. For example, if A_n leaves out only the first n odd integers, then A is the set of even integers, and neither A nor A^c is finite.

two functions, X_1 and X_2, we might try to calculate the probability that they are equal, $P(X_1 = X_2)$. Strictly speaking, this is the probability of the set of ω so that $X_1(\omega) = X_2(\omega)$.

2.17. 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 $B_x = \{\omega \ : \ X(\omega) = x\}$. In discrete probability, B_x will be the empty set for almost all x values and will not be empty 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 B_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 Example 2 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 (find an x and $B_x \notin \mathcal{F}$ to show this).

2.18. 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 in A_4 or A_5", which would be written $A = (A_1 \cap (A_2 \cup A_3)) \cap (A_4 \cup A_5)^c$. This is the same as 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 both the first two are 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. To generate a $\sigma-$ algebra, we mayhave to allow infinitely many set operations, but a precise discussion of this would be "off message".

2.19. Generating by a function: A function $X(\omega)$ defines an algebra of sets generated by the sets B_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 1 and 3 or greater than 4.

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 B_x is not empty, then there is some number, $u(x)$, so that $Y(\omega) = u(x)$ for every $\omega \in B_x$. 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, $u(x) = 2x - 4$.

2.20. Equivalence relation: A $\sigma-$algebra, $\mathcal{F}$, determines an *equivalence relation.* Outcomes ω_1 and ω_2 are equivalent, written $\omega_1 \sim \omega_2$, if the information in $\mathcal{F}$ does not distinguish ω_1 from ω_2. More formally, $\omega_1 \sim \omega_2$ if $\omega_1 \in A \Rightarrow \omega_2 \in A$ for every $A \in \mathcal{F}$. For example, in Example 2 above, THTT $\sim$ TTTH. Because $\mathcal{F}$ is an algebra, $\omega_1 \sim \omega_2$ also implies that $\omega_1 \notin A \Rightarrow \omega_2 \notin A$ (think this through). Note that it is possible that $A_\omega = A_{\omega'}$ while $\omega \neq \omega'$. This happens when $\omega \sim \omega'$.

The *equivalence class* of outcome ω is the set of outcomes equivalent to ω in $\mathcal{F}$, indistinguishable from ω using the information available in $\mathcal{F}$. If A_ω is the equivalence class of ω, then $A_\omega \in \mathcal{F}$. (Proof: for any ω' not equivalent to ω in $\mathcal{F}$, there is at least one $B_{\omega'} \in \mathcal{F}$ with $\omega \in B_{\omega'}$ but $\omega' \notin B_{\omega'}$. Since there are (at most) countably many ω', and $\mathcal{F}$ is a $\sigma-$algebra, $A_\omega = \cap_{\omega'} B_{\omega'} \in \mathcal{F}$. This A_ω contains every ω_1 that is equivalent to ω (why?) and only those.) In Example 2, the equivalence class of THTT is the event {HTTT, THTT, TTHT, TTTH}.

2.21. Partition: A *partition* of Ω is a collection of events, $\mathcal{P} = \{B_1, B_2, \ldots\}$ so that every outcome $\omega \in \Omega$ is in exactly one of the events B_k. The $\sigma-$algebra generated by $\mathcal{P}$, which we call $\mathcal{F}_\mathcal{P}$, consists of events that are unions of events in $\mathcal{P}$ (Why are complements and intersections not needed?). For any partition $\mathcal{P}$, the equivalence classes of $\mathcal{F}_\mathcal{P}$ are the events in $\mathcal{P}$ (think this through). Conversely, if $\mathcal{P}$ is the partition of Ω into equivalence classes for $\mathcal{F}$, then $\mathcal{P}$ generates $\mathcal{F}$. In Example 2 above, the sets $B_k = \{k\}$ form the partition corresponding to $\mathcal{F}$. More generally, the sets $B_x = \{\omega \mid X(\omega) = x\}$ that are not empty are the partition corresponding to $\mathcal{F}_X$. In discrete probability, partitions are a convenient way to understand conditional expectation (below). The ininformation in $\mathcal{F}_\mathcal{P}$ is the knowledge of which of the B_j happened. The remaining uncertainty i swhich of the $\omega \in B_j$ happened.

2.22. 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 telling how X was generated.) This is the "average" value in the sense that if you could perform the "experiment" of sampling X many times then 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).

2.23. 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

$$x_{\rm opt} = E[X] \ . \tag{5}$$

2.24. Classical conditional expectation: There are two senses of the term *conditional expectation*. We start with the original *classical* sense then turn to the related but different *modern* sense often used in stochastic processes. Conditional expectation is defined from conditional probability in the obvious way

$$E[X|B] = \sum_{\omega \in B} X(\omega) P(\omega|B) \ . \tag{6}$$

For example, we can calculate

$$E[\#\text{of H in 4 tosses} \ \mid \ \text{at least one H}] \ .$$

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 $X(\omega)$ be the number of H in ω. Unconditionally, $E[X] = 2$, which means

$$\frac{1}{16} \sum_{x \in \Omega} X(\omega) = 2 \ .$$

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

$$\sum_{\omega \in \Omega} X(\omega) P(\omega) = \sum_{\omega \in B} X(\omega) P(\omega) \ ,$$

and therefore

$$\frac{1}{16} \sum_{\omega \in B} X(\omega) P(\omega) \ = \ 2$$

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

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

2.25. Law of total probability: Suppose $\mathcal{P} = \{B_1, B_2, \ldots\}$ is a partition of Ω. The *law of total probability* is the formula

$$E[X] = \sum_k E[X \mid B_k] P(B_k) \; . \tag{7}$$

This is easy to understand: exactly one of the events B_k happens. The expected value of X is the sum over each of the events B_k of the expected value of X given that B_k happened, multiplied by the probability that B_k did happen. The derivation is a simple combination of the definitions of conditional expectation (6) and conditional probability (4):

$$\begin{aligned}
E[X] &= \sum_{\omega \in \Omega} X(\omega) P(\omega) \\
&= \sum_k \left(\sum_{\omega \in B_k} X(\omega) P(\omega) \right) \\
&= \sum_k \left(\sum_{\omega \in B_k} X(\omega) \frac{P(\omega)}{P(B_k)} \right) P(B_k) \\
&= \sum_k E[X \mid B_k] P(B_k) \; .
\end{aligned}$$

This fact underlies the recurrence relations that are among the primary tools of stochastic calculus. It will be reformulated below as the *tower property* when we discuss the modern view of conditional probability.

2.26. Modern conditional expectation: The modern conditional expectation starts with an algebra, $\mathcal{F}$, rather than just the set B. 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 (in the least squares sense) of X given the information in $\mathcal{F}$. If $X \in \mathcal{F}$, then the value of $X(\omega)$ is determined by the information in $\mathcal{F}$, so $Y = X$.

In the classical case, the information is the occurrance or non occurrance of a single event, B. That is, the algebra, $\mathcal{F}_B$, consists only of the sets B, B^c, $\emptyset$, and Ω. For this $\mathcal{F}_B$, 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, $\mathcal{P} = \{B_1, B_2, \ldots\}$, of Ω. The conditional expectation $Y(\omega) = E[X \mid \mathcal{F}]$ is defined to be

$$Y(\omega) = E[X \mid B_j] \text{ if } \omega \in B_j \ , \tag{8}$$

where $E[X \mid B_j]$ is classical conditional expectation (6). A single set B defines a partition: $B_1 = B$, $B_2 = B^c$, so this agrees with the earlier definition in that case. The information in $\mathcal{F}$ is only which of the B_j occurred. The modern conditional expectation replaces X with its expected value over the set taht occurred. This is the expected value of X given the information in $\mathcal{F}$.

2.27. Example of modern conditional expectation: Take Ω to be sequences of 4 coin tosses. Take $\mathcal{F}$ to be the algebra of Example 2 determined by the number of H tosses. Take $X(\omega)$ to be the number of H tosses before the first T (e.g. $X(\text{HHTH}) = 2$, $X(\text{TTTT}) = 0$, $X(\text{HHHH}) = 4$, etc.). With the usual abuse of notation, we calculate (below): $Y(\{0\}) = 0$, $Y(\{1\}) = 1/4$, $Y(\{2\}) = 2/3$, $Y(\{3\}) = 3/2$, $Y(\{4\}) = 4$. Note, for example, that because HHTT and HTHT are equivalent in $\mathcal{F}$ (in the equivalence class $\{2\}$), $Y(\text{HHTT}) = Y(\text{HTHT}) = 1/4$ even though $X(\text{HHTT}) \neq X(\text{HTHT})$. The common value of Y is its average

value of X over the outcomes in the equivalence class.

$\{0\}$	TTTT 0 expected value $= 0$
$\{1\}$	HTTT THTT TTHT TTTH 1 0 0 0 expected value $= (1+0+0+0)/4 = 1/4$
$\{2\}$	HHTT HTHT HTTH THHT THTH TTHH 2 1 1 0 0 0 expected value $= (2+1+1+0+0+0)/6 = 2/3$
$\{3\}$	HHHT HHTH HTHH THHH 3 2 1 0 expected value $= (3+2+1+0)/4 = 3/2$
$\{4\}$	HHHH 4 expected value $= 4$

2.28. Best approximation property: Suppose we have a random variable, $X(\omega)$, that is not measurable with respect to the $\sigma-$algebra $\mathcal{F}$. That is, the information in $\mathcal{F}$ does not completely determine the value of X. The conditional expectation, $Y(\omega) = E[X \mid \mathcal{F}]$, among all functions measurable with respect to $\mathcal{F}$, is the closest to X in the least squares sense. That is, if $Z \in \mathcal{F}$, then

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

In fact, this best approximation property will be the definition of conditional expectation in situations where the partition definition is not directly applicable. The best approximation property for modern conditional expectation is a consequence of the best approximation for classical conditional expectation. The least squares error is the sum of the least squares errors over each B_k in the partition defined by $\mathcal{F}$. We minimize the least squares error in B_k by choosing $Y(B_k)$ to be the average of X over B_k (weighted by the probabilities $P(\omega)$ for $\omega \in B_k$). By choosing the best approximation in each B_k, we get the best approximation overall.

This can be expressed in the terminology of linear algebra. The set of functions (random variables) X is a vector space (Hilbert space) with inner product

$$\langle X, Y \rangle = \sum_{\omega \in \Omega} X(\omega) Y(\omega) P(\omega) = E\left[XY\right] \ ,$$

so $\|X - Y\|^2 = E\left[(X-Y)^2\right]$. The set of functions measurable with respect to $\mathcal{F}$ is a subspace, which we call $\mathcal{S}_{\mathcal{F}}$. The conditional expectation, Y, is the orthogonal projection of X onto $\mathcal{S}_{\mathcal{F}}$, which is the element of $\mathcal{S}_{\mathcal{F}}$ that closest to X in the norm just given.

2.29. Tower property: Suppose $\mathcal{G}$ is a $\sigma-$algebra that has less information than $\mathcal{F}$. That is, every event in $\mathcal{G}$ is also in $\mathcal{F}$, but events in $\mathcal{F}$ need not be in $\mathcal{G}$. This is expressed simply (without abuse of notation) as $\mathcal{G} \subseteq \mathcal{F}$. Consider the (modern) conditional expectations $Y = E[X \mid \mathcal{F}]$ and $Z = E[X \mid \mathcal{G}]$. The *tower property* is the fact that $Z = E[Y \mid \mathcal{G}]$. That is, conditioning in one step gives the same result as conditioning in two steps. As we said before, the tower property underlies the backward equations that are among the most useful tools of stochastic calculus.

The tower property is an application of the law of total probability to conditional expectation. Suppose $\mathcal{P}$ and $\mathcal{Q}$ are the partitions of Ω corresponding to $\mathcal{F}$ and $\mathcal{G}$ respectively. The partition $\mathcal{P}$ is a *refinement* of $\mathcal{Q}$, which means that each $C_k \in \mathcal{Q}$ itself is partitioned into events $\{B_{k,1}, B_{k,2}, \ldots\}$, where the $B_{k,j}$ are elements of $\mathcal{P}$. Then (see "Working with conditional probability") for $\omega \in C_k$, we want to show that $Z(\omega) = E[Y \mid C_k]$:

$$\begin{aligned} Z(\omega) &= E[X \mid C_k] \\ &= \sum_j E[X \mid B_{jk}] P(B_{jk} \mid C_k) \\ &= \sum_j Y(B_{jk}) P(B_{jk} \mid C_k) \\ &= E[Y \mid C_k] \ . \end{aligned}$$

The linear algebra projection interpretation makes the tower property seem obvious. Any function measurable with respect to $\mathcal{G}$ is also measurable with respect to $\mathcal{F}$, which means that the subspace $\mathcal{S}_{\mathcal{G}}$ is contained in $\mathcal{S}_{\mathcal{F}}$. If you project X onto $\mathcal{S}_{\mathcal{F}}$ then project the projection onto $\mathcal{S}_{\mathcal{G}}$, you get the same thing as projecting X directly onto $\mathcal{S}_{\mathcal{G}}$ (always orthogonal projections).

2.30. Modern conditional probability: Probabilities can be defined as expected values of characteristic functions (see below). Therefore, the modern definition of conditional expectation gives a modern definition of conditional probability. For any event, A, the *indicator function*, $\mathbf{1}_A(\omega)$, (also written $\chi_A(\omega)$, for "characteristic function", terminology less used by probabilists because characteristic function means something else to them) is defined by $\mathbf{1}_A(\omega) = 1$ if $\omega \in A$, and $\mathbf{1}_A(\omega) = 0$ if $\omega \notin A$. The obvious formula $P(A) = E[\mathbf{1}_A]$ is the

representation of the probability as an expected value. The modern conditional probability then is $P(A \mid \mathcal{F}) = E[\mathbf{1}_A \mid \mathcal{F}]$. Unraveling the definitions, this is a function, $Y_A(\omega)$, that takes the value $P(A \mid B_k)$ whenever $\omega \in B_k$. A related statement, given for practice with notation, is

$$P(A \mid \mathcal{F})(\omega) = \sum_{B_k \in \mathcal{P}_\mathcal{F}} P(A \mid B_k)\mathbf{1}_{B_k}(\omega) \; .$$

3 Markov Chains, I

3.1. Introduction: Discrete time Markov[3] chains are a simple abstract class of discrete random processes. Many practical models are Markov chains. Here we discuss Markov chains having a finite *state space* (see below).

Many of the general concepts above come into play here. The probability space Ω is the space of paths. The natural states of partial information are described by the algebras $\mathcal{F}_t$, which represent the information obtained by observing the chain up to time t. The tower property applied to the $\mathcal{F}_t$ leads to backward and forward equations. This section is mostly definitions. The good stuff is in the next section.

3.2. Time: The time variable, t, will be an integer representing the number of time units from a starting time. The actual time to go from t to $t+1$ could be a nanosecond (for modeling computer communication networks) or a month (for modeling bond rating changes), or whatever. To be specific, we usually start with $t = 0$ and consider only non negative times.

3.3. 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 at future times $t+1$, $t+2$, ... This is illustrated with the hidden Markov model below. 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, ..., s_m , or simply $1, 2, \ldots, m$. depending on the context.

3.4. Path space: The sequence of states X_0, X_1, ..., X_T, is a *path*. The set of paths is *path space*. It is possible and often convenient to use the set of paths as the probability space, Ω. When we do this, the path $X = (X_0, X_1, \ldots, X_T) = (X(0), X(1), \ldots, X(T))$ plays the role that was played by the outcome ω in the general theory above. We will soon have a formula for the $P(X)$, probability of path X, in terms of *transition probabilities*.

[3]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.

In principle, it should be possible to calculate the probability of any event (such as $\{X(2) \neq s\}$, or $\{X(t) = s_1 \text{ for some } t \leq T\}$) by listing all the paths (outcomes) in that event and summing their probabilities. This is rarely the easiest way. For one thing, the path space, while finite, tends to be enormous. For example, if there are $m = |\mathcal{S}| = 7$ states and $T = 50$ times, then the number of paths is $\|\Omega\| = m^T = 7^{50}$, which is about 1.8×10^{42}. This number is beyond computers.

3.5. Algebras $\mathcal{F}_t$ and $\mathcal{G}_t$: The information learned by observing a Markov chain up to and including time t is $\mathcal{F}_t$. Paths X_1 and X_2 are equivalent in $\mathcal{F}_t$ if $X_1(s) = X_2(s)$ for $0 \leq s \leq t$. Said only slightly differently, the equivalence class of path X is the set of paths X' with $X'(s) = X(s)$ for $0 \leq s \leq t$. The $\mathcal{F}_t$ form an increasing family of algebras: $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$. (Event A is in $\mathcal{F}_t$ if we can tell whether A occurred by knowing $X(s)$ for $0 \leq s \leq t$. In this case, we also can tell whether A occurred by knowing $X(s)$ for $0 \leq s \leq t+1$, which is what it means for A to be in $\mathcal{F}_{t+1}$.)

The algebra $\mathcal{G}_t$ is generated by $X(t)$ only. It encodes the information learned by observing X at time t only, not at earlier times. Clearly $\mathcal{G}_t \subseteq \mathcal{F}_t$, but $\mathcal{G}_t$ is not contained in $\mathcal{G}_{t+1}$, because $X(t+1)$ does not determine $X(t)$.

3.6. Nonanticipating (adapted) functions: The underlying outcome, which was called ω, is now called X. A function of a the outcome, or function of a random variable, will now be called $F(X)$ instead of $X(\omega)$. Over and over in stochastic processes, we deal with functions that depend on both X and t. Such a function will be called $F(X,t)$. The simplest such function is $F(X,t) = X(t)$. More complicated functions are: (*i*) $F(X,t) = 1$ if $X(s) = 1$ for some $s \leq t$, $F(X,t) = 0$ otherwise, and (*ii*) $F(X,t) = \min(s > t)$ with $X(s) = 1$ or $F(X,t) = T$ if $X(s) \neq 1$ for $t < s \leq T$.

A function $F(X,t)$ is *nonanticipating* (also called *adapted*, though the notions are slightly different in more sophisticated situations) if, for each t, the function of X given by $F(X,t)$ is measurable with respect to $\mathcal{F}_t$. This is the same as saying that $F(X,t)$ is determined by the values $X(s)$ for $s \leq t$. The function (*i*) above has this property but (*ii*) does not.

Nonanticipating functions are important for several reasons. In time, we will see that the Ito integral makes sense only for nonanticipating functions. Moreover, functions $F(X,t)$ are a model of decision making under uncertainty. That F is nonanticipating means that the decision at time t is made based on information available at time t and does not depend on future information.

3.7. Markov property: Informally, the *Markov property* is that $X(t)$ is all the information about the past that is helpful in predicting the future. In classical terms, for example,

$$P(X(t+1) = k | X(t) = j) = P(X(t+1) = k | X(t) = j, X(t-1) = l, \text{etc.}) \; .$$

In modern notation, this may be stated

$$P(X(t+1) = k \mid \mathcal{F}_t) = P(X(t+1) = k \mid \mathcal{G}_t) \; . \tag{9}$$

Recall that both sides are functions of the outcome, X. The function on the right side, to be measurable with respect to $\mathcal{G}_t$ must be a function of $X(t)$ only (see "Generating by a function" in the previous section). The left side also is a function, but in general could depend on all the values $X(s)$ for $s \leq t$. The equality (9) states that this function depends on $X(t)$ only.

This may be interpreted as the absence of hidden variables, variables that influence the evolution of the Markov chain but are not observable or included in the state description. If there were hidden variables, observing the chain for a long period might help identify them and therefore change our prediction of the future state. The Markov property (9) states, on the contrary, that observing $X(s)$ for $s < t$ does not change our predictions.

3.8. Transition probabilities: The conditional probabilities (9) are *transition probabilities*:

$$P_{jk} = P\left(X(t+1) = k \mid X(t) = j\right) = P(j \to k \text{ in one step}) \; .$$

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 1. 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 $t+1$.

3.9. Path probabilities: The Markov property leads to a formula for the probabilities of individual path outcomes $P(X)$ as products of transition probabilities. We do this here for a stationary Markov chain to keep the notation simple. First, suppose that the probabilities of the initial states are known, and call them

$$f_0(j) = P(X(0) = j) \; .$$

The Bayes' rule (2) implies that

$$\begin{aligned} &P(X(1) = k \text{ and } X(0) = j) \\ &\quad = \quad P(X(1) = k \mid X(0) = j) \cdot P(X(0) = j) = f_0(j) P_{jk} \; . \end{aligned}$$

Using this argument again, and using (9), we find (changing the order of the factors on the last line)

$$\begin{aligned} &P(X(2) = l \text{ and } X(1) = k \text{ and } X(0) = j) \\ &\quad = \quad P(X(2) = l \mid X(1) = k \text{ and } X(0) = j) \cdot P(X(1) = k \text{ and } X(0) = j) \\ &\quad = \quad P(X(2) = l \mid X(1) = k) \cdot P(X(1) = k \text{ and } X(0) = j) \\ &\quad = \quad f_0(j) P_{jk} P_{kl} \; . \end{aligned}$$

This can be extended to paths of any length.

One way to express the general formula uses a notational habit common in probability, using upper case letters to represent a random value of a variable and lower case for generic values of the same quantity (see "Terminology", Section 2, but note that the meaning of X has changed). We write $x = (x(0), x(1), \cdots, x(T))$ for a generic path, and seek $P(x) = P(X = x) = P(X(0) = x(0), X(1) = x(1), \cdots)$. The argument above shows that this is given by

$$P(x) = f_0(x(0))P_{x(0),x(1)} \cdots P_{x(T-1),x(T)} = f_0(x(0)) \prod_{t=0}^{T-1} P_{x(t),x(t+1)} \ . \qquad (10)$$

3.10. Transition matrix: The transition probabilities form an $m \times m$ matrix, P (an unfortunate conflict of notation), called the *transition matrix*. The (j, k) entry of P is the transition probability $P_{jk} = P(j \to k)$. 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 help in the analysis of Markov chains. As we will see in the next lecture, 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). Also, as discussed later, *steady state* probabilities form an eigenvector of P corresponding to eigenvalue $\lambda = 1$.

3.11. Example 3, coin flips: The state space has $m = 2$ states, called U (up) and D (down). Writing H and T would conflict with T being the length of the chain. The coin starts in the U position, which means that $f_0(\mathrm{U}) = 1$ and $f_0(\mathrm{D}) = 0$. At every time step, the coin turns over with 20% probability, so 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 $P(X(4) = D) = P(X(0) = U \to X(4) = D) = P^4_{UD} = .5648$. The eigenvalues of P are $\lambda_1 = 1$ and $\lambda_2 = .6$, the former required by theory. Numerical experimentation should can convince the reader that

$$\left\| P^s - \begin{pmatrix} .5 & .5 \\ .5 & .5 \end{pmatrix} \right\| = const \cdot \lambda_2^s \ .$$

Take $T = 3$ and let A be the event UUzU, where the state $X(2) = z$ is unknown. There are two outcomes (paths) in A:

$$A = \{\text{UUUU}, \text{UUDU}\} \ ,$$

so $P(A) = P(\text{UUUU}) + P(\text{UUDU})$. The individual path probabilities are calculated using (10):

$$\text{U} \xrightarrow{.8} \text{U} \xrightarrow{.8} \text{U} \xrightarrow{.8} \text{U} \text{ so } P(\text{UUUU}) = 1 \times .8 \times .8 \times .8 = .512 \ .$$

$$\text{U} \xrightarrow{.8} \text{U} \xrightarrow{.2} \text{D} \xrightarrow{.2} \text{U} \text{ so } P(\text{UUDU}) = 1 \times .8 \times .2 \times .2 = .032 \ .$$

Thus, $P(A) = .512 + .032 = .544$.

3.12. Example 4: 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 is 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 we arbitrarily number 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 $\rightarrow$ UF. We first have a U $\rightarrow$ U (non) transition then an F $\rightarrow$ (non) transition. The probability is then $P(U \rightarrow U \mid F) \cdot P(F \rightarrow F) = .8 \cdot .7 = .56$. The other entries can be found in a similar way. The transitions are:

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

The resulting transition matrix is

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

If we start with U but equally likely F or S, and want to know the probability of being D after 4 time periods, the answer is

$$.5 \cdot \left(P_{12}^4 + P_{14}^4 + P_{32}^4 + P_{34}^4\right)$$

because states 1 = UF and 3 = US are the (equally likely) possible initial U states, and $2 = DF$ and $4 = DS$ are the two D states. We also could calculate $P(UUzU)$ by adding up the probabilities of the 32 (list them) paths that make up this event.

3.13. Example 5, incomplete state information: In the model of example 4 we might be able to observe the U–D status but not the F–S status. Let $X(t)$ be the state of the Example 4 model above at time t. Suppose $Y(t) = U$ 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 its F–S status.

For example, suppose that the four states (UF, DF, US, DS) at time $t = 0$ are equally likely, that we know $Y(1) = U$ and we want to guess whether $Y(2)$ will again be U. If $Y(0)$ is D then we are more likely to have the F coin so a $Y(1) = U \to Y(2) = D$ transition is more likely. That is, with $Y(1)$ fixed, $Y(0) = D$ makes it less likely to have $Y(2) = U$. This is a violation of the Markov property brought about by incomplete state information. Models of this kind are called *hidden Markov* models. Statistical estimation of the unobserved variable is a topic for another day.

Thanks to Laura K and Craig for pointing out mistakes and confusions in earlier drafts.