Stochastic Calculus Notes, Lecture 1

Last modified September 9, 2002

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 information.

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.

1.2. 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.

1.3. "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 occured $A = \{\omega\}$. 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.

1.4. 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

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

1.5. 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.

1.6. Basic facts: Each of these facts is a consequence of the representation $P(A) = \sum_{\omega\in A}$. First $P(\omega)P(A) \le P(B)$ 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 $P(A) + P(A^c) = P(\Omega) = 1$.

1.7. Conditional probability: The probability of outcome A given that B has occured is

$$P(A \mid B) = \frac{P(A\cap B)}{P(B)}\ . \tag{1}$$

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 $P(A\cap B) = P(A \mid B)P(B)$.

1.8. 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

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

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 (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) 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 $\frac{3}{4}$.

1.9. 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 occured? 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 $B \cap B$ (B and C) all at once.

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 $\mathcal{F}$ occured but we do not have enough information to

determine whether an event not in $\mathcal{F}$ occured. 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.

1.11. 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.

1.12. 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 $6 \cdot \dfrac{1}{16} = \dfrac{3}{8}$. There are other events measureable in this algebra, such as "less than 3 H", but, in some sense, the events listed "generate" the algebra.

1.13. 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 $F(x) = \sum_{X(\omega) \leq x} P(\omega)$.

1.14. 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 $X_1(\omega) = X_2(\omega)$.

1.15. 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.

1.16. Generating an algebra of sets: Suppose there are events $A_1, \ldots, A_k$ 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.

1.17. 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, $u(x) = 2x - 4$.

1.18. Expected value: A random variable (actually, a function of a random variable) $X(\omega)$ has expected value

$$E[X] = \sum_{\emptyset, ega \in \Omega} X(\omega)P(\omega) \ .$$

(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).

1.19. 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

$$x_{\text{opt}} = E[X] \ . \tag{3}$$

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 $\omega =$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

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

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

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 occurance or non occurance 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, ..., B_n, so that each outcome, ω is in one and only one of the events. The sets {4}, {3}, ..., {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

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

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 $\mathcal{F}$.

1.22. 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 $B_1, \ldots, B_n$. 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 $E[X \mid B_j]$.

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.

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.

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 at future times $t+1$, $t+2$, ... This will be more

[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 distrubution of prime numbers as well as on probability.

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, ..., s_m , or simply $1, 2, \ldots, m$. depending on the context.

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.

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, ..., m is a complete list of the possible states at time $t + 1$.

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 (explination below). The "steady state" probabilities form an eigenvector of P.

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 irrelevent 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, $P(X_1 = j) = f_j^1$. The latter may also be written $f(j, 1)$. In general we use notation $f_j^t = P(X_t = j)$. Using f^1 and the P_{jk}, we can calculate the probabilities of paths:

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

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

and so on. Expressed slightly differently, we have

$$P(X) = f_{X_1}^1 \cdot P_{X_1, X_2} \cdot \ \cdots \ \cdot P_{X_{T-1}, X_T} \ . \tag{4}$$

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 $P(X_5 = U) = P(X_1 = U \to X_5 = U) = P_{UU}^5 = .5648$

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) transitin 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

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

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.

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 $Y_8 = U \rightarrow Y_9 = D$ 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.