Stochastic Calculus Notes, Lecture 3

Last modified September 17, 2002

1 Recurrence relations for Markov Chains

1.1. Recapituation 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

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

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

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

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:

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

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 (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.

1.4. Backward equation, matrix version: The equation (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 (1) is equivalent to

$$f_t = P \cdot f_{t+1} \; . \tag{2}$$

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 (2) may be iterated to give

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

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

$$\begin{aligned} u_{t+1}(k) &= P(X_{t+1} = k) \\ &= \sum_{j=1}^{s} P(X_t = j) P(X_{t+1} = k \mid X_t = j) \\ u_{t+1}(k) &= \sum_{j=1}^{s} u_t(j) P_{jk} \end{aligned} \tag{3}$$

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

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

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

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) may be expresses as

$$u_{t+1} = u_t P \; , \tag{4}$$

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 (2) to get, for example, $u_t = u_1 P^{t-1}$.

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.

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 (2) implies the forward evolution (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

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

(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 (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.

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.

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 (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.

2 Martingales and stopping times

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, $\ldots$, 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.

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.

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: $B_{T-1,1} = \{1,2\}$, $B_{T-1,2} = \{3,4\}$, ..., $B_{T-1,2^{T-2}} = \{2^{T-1}-1, 2^{T-1}\}$. 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.

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, ..., Y_t). The reader should check this.

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.

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 $\mathcal{F}_t$.

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 $\mathcal{G}_t$.

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.

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.

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.

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 $E[Y_{t+1} \mid \mathcal{F}_t] = 0$.

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.

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 $t' > t$.

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.

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 $E[X_{\tau'}] = 0$.

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.