Week 2
Discrete Markov chains

Jonathan Goodman

September 17, 2012

1 Introduction to the material for the week

This week we discuss Markov random processes in which there is a list of possible states. We introduce three mathematical ideas: a $\sigma-algebra$ to represent a state of partial information, *measurability* of a function with respect to a discrete $\sigma-$algebra, and a *filtration* that represents gaining information over time. Filtrations are a convenient way to describe the Markov property and to give the general definition of a martingale, the latter a few weeks from now. Associated with Markov chains are the backward and forward equations that describe the evolution of probabilities and expectation values over time. Forward and backward equations are one of the main "calculate things" methods of stochastic calculus.

A *stochastic process* in *discrete time* is a sequence $(X_1, X_2, \ldots)$, where X_n is the *state* of the system at time n. The *path* up to time T is $X_{[1:T]} = (X_1, X_2, \ldots, X_T)$. The state X_n must be in the *state space*, $\mathcal{S}$. Last week $\mathcal{S}$ was $\mathbb{R}^d$, for a linear Gaussian process. This week, $\mathcal{S}$ is either a finite set $\mathcal{S} = \{x_1, x_2, \ldots x_m\}$, or an infinite *countable* set of the form $\mathcal{S} = \{x_1, x_2, \ldots\}$. A set such as $\mathcal{S}$ is *discrete* if it is finite or countable. The set of all real numbers, like the Gaussian state space $\mathbb{R}^d$, is not discrete because the real numbers are not countable (a famous theorem of Georg Cantor). Spaces that are not discrete may be called *continuous*. If we define X_t for all times t, then X_t is a *continuous time* process. If X_n is defined only for integers n (or another discrete set of times), then we have a *discrete time* process. This week is about discrete time discrete state space stochastic processes.

We will be interested in discrete Markov chains partly for their own sake and partly because they are a setting where the general definitions are easy to give without much mathematical subtlety. The concepts of measurability and filtration for continuous time or continuous state space are more technical and subtle than we have time for in this class. The same is true of backward and forward equations. They are rigorous this week, but heuristic when we go to continuous time and state space. That is the "$\Delta X \rightarrow 0$ and $\Delta t \rightarrow 0$" aspect of stochastic calculus.

The main examples of Markov chain will be random walk and mean reverting random walk. There are discrete versions of Brownian motion and the Ornstein Uhlenbeck process respectively.

2 Basic probability

This section gives some basic general definitions in probability theory in a setting where they are not technical. Look to later sections for more examples. Philosophically, a *probability space*, Ω, is the set of all possible *outcomes* of a "probability experiment". Mathematically, Ω is just a set. In abstract discussions, we usually use ω to denote an element of Ω. In concrete settings, the elements of Ω have more concrete descriptions. This week, Ω will usually be the *path space* consisting of all paths $x_{[1:T]} = (x_1, \ldots, x_T)$, with $x_n \in \mathcal{S}$ for each n. If $\mathcal{S}$ is discrete and T is finite then the path space is discrete. If T is infinite, Ω is not discrete. The discussion below needs to be given more carefully in that case, which we do not do in this class.

An *event* is the answer to a yes/no question about the outcome ω. Equivalently, an event is a subset of the probability space: $A \subseteq \Omega$. You can interpret A as the set of outcomes where the answer is "yes", and $A^c = \{\omega \in \Omega | \omega \notin A\}$ is the complementary set where the answer is "no". We often describe an event using a version of set notation where the informal definition of the event goes inside curly braces, such as $\{X_3 \neq X_5\}$ to describe $\left\{x_{[1:T]} | x_3 \neq x_5\right\}$.

A $\sigma-$*algebra* is a mathematical model of a state of partial knowledge about the outcome. Informally, if $\mathcal{F}$ is a $\sigma-$algebra and $A \subseteq \Omega$, we say that $A \in \mathcal{F}$ if we know whether $\omega \in A$ or not. The most used $\sigma-$algebra in stochastic processes is the one that represents knowing the first n states in a path. This is called $\mathcal{F}_n$. To illustrate this, if $n \geq 2$, then the event $A = \{X_1 = X_2\} \in \mathcal{F}_n$. If we know both X_1 and X_2, then we know whether $X_1 = X_2$. On the other hand, we do not know whether $X_n = X_{n+1}$, so $\{X_n \neq X_{n+1}\} \notin \mathcal{F}_n$.

Here is the precise definition of a $\sigma-$algebra. The *empty set* is written $\emptyset$. It is the event with no elements. We say $\mathcal{F}$ is a $\sigma-$algebra if (explanations of the axioms in parentheses):

(i) $\Omega \in \mathcal{F}$ and $\emptyset \in \Omega$.
(Regardless of how much you know, you know whether $\omega \in \Omega$, it is, and you know whether $\omega \in \emptyset$, it isn't.)

(ii) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.
(If you know whether $\omega \in A$ then you know whether $\omega \notin A$. It is the same information.)

(iii) IF $A \in \mathcal{F}$ and $B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$ and $A \cap B \in \mathcal{F}$.
(If you can answer the questions $\omega \in A$? and $\omega \in B$?, then you can answer the question $\omega \in (A \text{ or } B)$? If $\omega \in A$ or $\omega \in B$, then $\omega \in (A \cup B)$.)

(iv) If $A_1, A_2, \ldots$ is a sequence of events, then $\cup A_k \in \mathcal{F}$.
(If any of the A_k is "yes", then the whole thing is "yes". The only way to

get “no” for $\omega \in \cup A_k \in \mathcal{F}$?, is for every one of the A_k to be “no”.)

This is not a minimal list. There are some redundancies. For example, if you have axiom (*ii*), and $\Omega \in \mathcal{F}$, then it follows that $\emptyset = \Omega^c \in \mathcal{F}$. The last axiom is called *countable additivity.* You need countable additivity to do $\Delta t \to 0$ stochastic calculus.

A *function of a random variable*, sometimes called a *random variable* is a real valued (later, vector valued) function of $\omega \in \Omega$. For example, if Ω is path space and the outcome is a path, and $a \in \mathcal{S}$ is a specific state, then the function could be

$$f(x_{[1:T]}) = \begin{cases} \min\{n \mid x_n = a\} & \text{if there is such an } n \\ T & \text{otherwise.} \end{cases}$$

This is called a *hitting time* and is often written τ_a. It would be more complete to write $\tau_a(x_{[1:T]})$, but it is common not to write the function argument. Such a function has a discrete set of values when Ω is discrete. If there is a finite or infinite list of all elements of Ω, then there is a finite or infinite list of possible values of f. Some of the definitions below simpler and less technical when Ω is discrete.

A function of a random variable, $f(\omega)$, is *measurable* with respect to $\mathcal{F}$ if the value of f can be determined from the information in $\mathcal{F}$. If Ω is discrete, this means that for any number F, the question $f(\omega) \overset{?}{=} F$ can be answered using the information in $\mathcal{F}$. More precisely, it means that for any F, the event $A_F = \{\omega \in \Omega | f(\omega) = F\}$ is an element of $\mathcal{F}$. To be clear, $A_F = \emptyset$ for most values of F because there is a list finite or countable list of F values for which $A_F \neq \emptyset$.

Let $\mathcal{F}_n$ be a family of $\sigma-$algebras, defined for $n = 0, 1, 2, \ldots$. They form a *filtration* if $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for each n. This is a very general model of acquiring new information at time n. The only restriction is that in a filtration you do not forget anything. If you know the answer to a question at time n, then you still know at time $n+1$. The set of questions you can answer at time n is a subset of the set of questions you can answer at time $n+1$. It is common that $\mathcal{F}_0$ is the trivial $\sigma-$algebra, $\mathcal{F}_0 = \{\emptyset, \Omega\}$, in which you can answer only trivial questions. The most important filtration for us has Ω being the path space and $\mathcal{F}_n$ knowing the path up to time n. This is called the *natural* filtration, or the filtration *generated* by the process X_n, in which $\mathcal{F}_n$ knows the values $X_1, \ldots, X_n$.

Suppose $\mathcal{F}_n$ is a filtration and $f_n(\omega)$ is a family of functions. We say that the functions are *progressively measurable*, or *non-anticipating*, or *adapted* to the filtration, or *predictable*, if f_n is measurable with respect to $\mathcal{F}_n$ for each n. There subtle differences between these concepts for continuous time processes, differences we will ignore in this class. Adapted functions are important in several ways. In the Ito calculus that is the core of this course, the integrand in the Ito integral must be adapted. In stochastic control problems, you try to find decide on a control at time n using only the information available at time n. A realistic stochastic control must be non-anticipating.

Partitions are a simple way to describe $\sigma-$algebras in discrete probability. A partition, $\mathcal{P}$, is a collection of events that is mutually exclusive and collectively exhaustive. That means that if $\mathcal{P} = \{A_1, \ldots\}$, then:

(*i*) $A_i \cap A_j = \emptyset$ whenever $i \neq j$.
(mutually exclusive.)

(*ii*) $\cup A_i = \Omega$.
(collectively exhaustive)

The events A_i are the *elements* of the partition $\mathcal{P}$. They form a partition if each $\omega \in \Omega$ is a member of exactly one partition element. A partition may have finitely or countably many elements. One example of a partition, if Ω is the path space, has one partition element for every state $y \in \mathcal{S}$. We say $x_{[1,T]} \in A_y$ if and only if $x_1 = y$. Of course, we could partition using the value x_2, etc.

In discrete probability, there is a one to one correspondence between partitions and $\sigma-$algebras. If $\mathcal{F}$ is a $\sigma-$algebra, the corresponding partition, informally, is the finest grained information contained in $\mathcal{F}$. To say this more completely, we say that $\mathcal{F}$ *distinguishes* ω from ω' if there is an $A \in \mathcal{F}$ so that $\omega \in A$ and $\omega' \notin A$. For example, let $\mathcal{F}_2$ part of the natural filtration of path space. Suppose $\omega = (y_1, y_2, y_3, \ldots)$, and $\omega' = (y_1, y_2, z_3, \ldots)$. Then $\mathcal{F}_2$ does not distinguish ω from ω'. The information in $\mathcal{F}_2$ cannot answer the question $y_3 \overset{?}{=} z_3$. For any $\omega \in \Omega$, the set of ω' that cannot be distinguished from ω is an event, called the *equivalence class* of ω. The set of all equivalence classes forms a partition of Ω (check properties (*i*) and (*ii*) above – if two equivalence classes overlap, then they are the same). If B_ω is the equivalence class of ω, then

$$B_\omega = \cap A \quad , \quad \text{with } \omega \in A \quad , \quad \text{and } A \in \mathcal{F} \; .$$

Therefore, $B_\omega \in \mathcal{F}$ (countable additivity of $\mathcal{F}$). Clearly, if $A \in \mathcal{F}$, then A is the union of all the equivalence classes contained in A. Therefore, if you have $\mathcal{P}$, you can create $\mathcal{F}$ buy taking all countable unions of elements of $\mathcal{P}$.

The previous paragraph is full if lengthy but easy verifications that you may not go through completely or remember long. What you should remember is what a partition is and how it carries the information in a $\sigma-$algebra. The information in $\mathcal{F}$ does not tell you which outcome happened, but it does tell you which partition element it was in. A function f is measurable with respect to $\mathcal{F}$ if and only if it is constant on each partition element of the corresponding $\mathcal{P}$. The information in $\mathcal{F}$ determines the partition element, B_j, which determines the value of f.

A *probability distribution* on Ω is an assignment of a probability to each outcome in Ω. If $\omega \in \Omega$, then $P(\omega)$ is the probability of ω. Naturally, $P(\omega) \geq 0$ for all ω and $\sum_{\omega\in\Omega} P(\omega) = 1$. If $A \subset \Omega$ is an event, then $P(A) = \sum_{\omega\in A} P(\omega)$. If $f(\omega)$ is a function of a random variable, then

$$\mathrm{E}[f] = \sum_{\omega\in\Omega} f(\omega) P(\omega) \; .$$

3 Conditioning

Conditioning is about how probabilities change as you get more information. The simplest conditional expectation tells you how the probability of event A changes if you know the event B happened. The formula is often called *Bayes' rule.*

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \cap B)}{P(B)} \; . \tag{1}$$

As a simple check, note that if $A \cap B = \emptyset$, then the event B rules A out completely. Bayes' rule (1) gets this right, as $P(\emptyset) = 0$ in the numerator. Bayes' rule does not say how to define conditional probability if $P(B) = 0$. This is a serious drawback in continuous probability. For example, if (X_1, X_2) is a bivariate normal, then $P(X_1 = 4) = 0$, but we saw last week how to calculate $P(X_2 > 0|X_1 = 4)$ (say). The conditional probability of a particular outcome is given by Bayes' rule too. Just take A to be the event that ω happened, which is written $\{\omega\}$:

$$P(\omega|B) = \begin{cases} P(\omega)\,/P(B) & \text{if } \omega \in B \\ 0 & \text{if } \omega \notin B \end{cases}$$

The conditional expected value is

$$E[f|B] = \sum_{\omega \in B} f(\omega) P(\omega|B) = \frac{\sum_{\omega \in B} f(\omega) P(\omega)}{P(B)} \; . \tag{2}$$

You can check that this formula gives the right answer when $f(\omega) = 1$ for all ω. We indeed get $E[1|B] = 1$.

You can think of expectation as something you say about a random function when you know nothing but the probabilities of various outcomes. There is a version of conditional expectation that describes how your understanding of f will change when you get the information in $\mathcal{F}$. Suppose $\mathcal{P} = (B_1, B_2, \ldots)$ is the partition that is determined by $\mathcal{F}$. When you learn the information in $\mathcal{F}$, you will learn which of the B_j happened. The conditional expectation of f, conditional on $\mathcal{F}$, is a function of ω determined by this information.

$$E[f|\mathcal{F}]\,(\omega) = E[f|B_j] \quad \text{if } \omega \in B_j \; . \tag{3}$$

To say this differently, if $g = E[f|\mathcal{F}]$, then g is a function of ω. If $\omega \in B_j$, then $g(\omega) = E[f|B_j]$.

You can see that the conditional expectation, $g = E[f|\mathcal{F}]$, is constant on partition elements $B_j \in \mathcal{P}$. This implies that g is measurable with respect to $\mathcal{F}$, which is another way of saying that $E[f|\mathcal{F}]$ is determined by the information in $\mathcal{F}$. The ordinary expectation is conditional expectation with respect to the trivial $\sigma-$algebra $\mathcal{F}_0 = \{\emptyset, \Omega\}$. The corresponding partition has only one element, Ω. The conditional expectation has the same value for every element of Ω, and that value is $E[f]$.

The *tower property* is a fact about conditional expectations. It leads to *backward equations* which is a powerful way to calculate conditional expectations.

Suppose $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$ is a filtration, and $f_n = \mathrm{E}[f|\mathcal{F}_n]$. The tower property is

$$f_n = \mathrm{E}[f_{n+1}|\mathcal{F}_n] \ . \tag{4}$$

This is a consequence of a simpler and more general statement. Suppose $\mathcal{G}$ is a $\sigma-$algebra with more information than $\mathcal{F}$, which means $\mathcal{F} \subseteq \mathcal{G}$. Suppose f is some function, $g = \mathrm{E}[f|\mathcal{G}]$, and $h = \mathrm{E}[f|\mathcal{F}]$. Then $h = \mathrm{E}[g|\mathcal{F}]$. To say this another way, you can condition from f down to h directly, which is $\mathrm{E}[f|\mathcal{F}]$, or you can do it in two stages, which is $f \longrightarrow g = \mathrm{E}[f|\mathcal{G}] \longrightarrow \mathrm{E}[g|\mathcal{F}]$. The result is the same.

One proof of the tower property makes use the partitions associated with $\mathcal{F}$ and $\mathcal{G}$. The partition for $\mathcal{G}$ is a *refinement* of the partition for $\mathcal{F}$. This means that you make the partition elements for $\mathcal{G}$ by cutting up partition elements of $\mathcal{F}$. Every C_i that is a partition element of $\mathcal{G}$ is completely contained in one of the partition elements of $\mathcal{F}$. Said another way, if ω and ω' are two elements of C_i, they are indistinguishable using the information in $\mathcal{G}$, which surely makes them indistinguishable using $\mathcal{F}$, which is less information. This is why C_i cannot contain outcomes from different B_j.

Now it is just a calculation. Let $h'(\omega) = \mathrm{E}[g|\mathcal{F}]$. For $\omega \in B_j$, it is intuitively clear (and we will verify) that

$$\begin{aligned}
h'(\omega) &= \sum_{C_i \subset B_j} \mathrm{E}[g|C_i]\,\mathrm{P}(C_i|B_j) \qquad (5)\\
&= \sum_{C_i \subset B_j} \mathrm{E}[f|C_i]\,\mathrm{P}(C_i|B_j)\\
&= \sum_{C_i \subset B_j} \left(\sum_{\omega \in C_i} f(\omega)\mathrm{P}(\omega|C_i) \right) \mathrm{P}(C_i|B_j)\\
&= \sum_{C_i \subset B_j} \left(\sum_{\omega \in C_i} \frac{f(\omega)\mathrm{P}(\omega)}{\mathrm{P}(C_i)} \right) \frac{\mathrm{P}(C_i)}{\mathrm{P}(B_j)}\\
&= \sum_{C_i \subset B_j} \left(\sum_{\omega \in C_i} f(\omega)\mathrm{P}(\omega) \right) \frac{1}{\mathrm{P}(B_j)}\\
&= \sum_{\omega \in B_j} f(\omega) \frac{\mathrm{P}(\omega)}{\mathrm{P}(B_j)}\\
&= h(\omega) \ .
\end{aligned}$$

The first line (5) is a convenient way to think about the partition produced by the $\sigma-$algebra $\mathcal{G}$. The partition elements C_i play the role of elementary outcomes $\omega \in \Omega$. The partition $\mathcal{P}$ plays the role of the probability space Ω. Instead of $\mathrm{P}(\omega)$, you have $\mathrm{P}(C_i)$. If the function g is measurable with respect to $\mathcal{G}$, then g has the same value for each $\omega \in C_i$, so you might as well call this $g(C_i)$. And of course, since g is constant is C_i, if $\omega \in C_i$, then $g(\omega) = \mathrm{E}[g|C_i]$.

We justify (5), for $\omega \in B_j$, using

$$\begin{aligned} h'(\omega) &= \mathrm{E}[g|B_j] \\ &= \sum_{\omega \in B_j} g(\omega)\mathrm{P}(\omega|B_j) \\ &= \sum_{C_i \in B_j} \left(\sum_{\omega \in C_i} g(\omega)\mathrm{P}(\omega) \right) \frac{1}{\mathrm{P}(B_j)} \\ &= \sum_{C_i \in B_j} g(C_i) \left(\sum_{\omega \in C_i} \mathrm{P}(\omega) \right) \frac{1}{\mathrm{P}(B_j)} \\ &= \sum_{C_i \in B_j} g(C_i) \frac{\mathrm{P}(C_i)}{\mathrm{P}(B_j)} \ . \end{aligned}$$

The last line is the same as (5).

4 Markov chains

This section, like the previous two, lacks examples. You might want to read the next section together with this one for examples.

A *Markov chain* is a stochastic process where the present is all the information about the past that is relevant for predicting the future. The $\sigma-$algebra definitions in the previous sections express these ideas easily. Here is a definition of the natural filtration $\mathcal{F}_n$. Let $x_{[1:T]}$ and $x'_{[1:T]}$ be two paths in the path space, Ω. Suppose $n \leq T$. We say the paths are indistinguishable at time n if $x_k = x'_k$ for $k = 1, 2, \ldots, n$. This definition of indistinguishability gives rise to a partition of Ω, with two paths being in the same partition element if they are indistinguishable. The $\sigma-$algebra corresponding to this partition is $\mathcal{F}_n$. A function $f(x_{[1:T]})$ is measurable with respect to $\mathcal{F}_n$ if it is determined by the first n states $(x_1, \ldots, x_n)$. More precisely, if $x_k = x'_k$ for $k = 1, 2, \ldots, n$, then $f(x_{[1:T]}) = f(x'_{[1:T]})$. The $\sigma-$algebra that "knows" only the value of x_n is $\mathcal{G}_n$. A path function is measurable with respect to $\mathcal{G}_n$ if and only if it is determined by the value x_n alone.

Let Ω be the path space and $\mathrm{P}(\cdot)$ a probability distribution on Ω. Then $\mathrm{P}(\cdot)$ has the *Markov property* if, for all $x \in \mathcal{S}$ and $n = 1, \ldots, T-1$,

$$\mathrm{P}(X_{n+1} = x \mid \mathcal{F}_n) = \mathrm{P}(X_{n+1} = x \mid \mathcal{G}_n) \ . \tag{6}$$

Unwinding all the definitions, this is the same as saying that for any path up to time n, $(x_1, \ldots, x_n)$,

$$\mathrm{P}(X_{n+1} = x \mid X_1 = x_1, \ldots, X_n = x_n) = \mathrm{P}(X_{n+1} = x \mid X_n = x_n) \ .$$

You might complain that we have defined conditional expectation but not conditional probability in (6). The answer is a trick for defining probability from expectation. The *indicator function* of an event $A \subseteq \Omega$ is $\mathbf{1}_A(\omega)$, which is equal to 1 if $\omega \in A$ and 0 otherwise. Then $\mathrm{P}(A) = \mathrm{E}[\mathbf{1}_A]$. In particular, if $A = \{\omega\}$, then, using the notation slightly incorrectly, $\mathrm{P}(\omega) = \mathrm{E}[\mathbf{1}_\omega]$. This applies to conditional expectation too: $\mathrm{P}(\omega|\mathcal{F}) = \mathrm{E}[\mathbf{1}_\omega|\mathcal{F}]$. But you should be alert to the fact that the latter statement is more complicated, in that both sides are functions on Ω (measurable with respect to $\mathcal{F}$) rather than just numbers. The simple notation hides the complexity.

The probabilities in a Markov chain are determined by *transition probabilities*, which are the numbers defined by the right side of (6). The probability $\mathrm{P}(X_{n+1} = x \mid \mathcal{G}_n)$ is a measurable function of $\mathcal{G}_n$, which means that they are a function of x_n, which we call y to simplify notation. The transition probabilities are

$$p_{n,yx} = \mathrm{P}(X_{n+1} = x \mid X_n = y) \ . \tag{7}$$

You can remember that it is $p_{n,yx}$ instead of $p_{n,xy}$ by saying that $p_{n,yx} = \mathrm{P}(y \to x)$ is the probability of of a y to x transition in one step.

You can use the definition (7) even if the process does not have the Markov property. What is special about Markov chains is that the numbers (7) determine all other probabilities. For example, we will show that

$$\mathrm{P}(X_{n+2} = x \text{ and } X_{n+1} = y \mid X_n = z) = p_{n+1,yx}p_{n,zy} \ . \tag{8}$$

What is behind this, besides the Markov property, is a general fact about conditioning. If A, B, and C are any three events, then

$$\mathrm{P}(A \text{ and } B \mid C) = \mathrm{P}(A \mid B \text{ and } C) \cdot \mathrm{P}(B \mid C) \ .$$

Without the Markov property, this leads to

$$\begin{aligned}\mathrm{P}(X_{n+2} = x \text{ and } X_{n+1} = y \mid X_n = z) = {} & \mathrm{P}(X_{n+2} = x \mid X_{n+1} = y \text{ and } X_n = z) \\ & \cdot \mathrm{P}(X_{n+1} = y \mid X_n = z) \ .\end{aligned}$$

According to the Markov property, the first probability on the right is $\mathrm{P}(X_{n+2} = x \mid X_{n+1} = y)$, which gives (8).

There is something in the spirit of (8) that is crucial for the backward equation below. This is that the Markov property applies to the whole future path, not just one step into the future. For examaple, (8) implies that $\mathrm{P}(X_{n+2} = x \mid \mathcal{F}_n) = \mathrm{P}(X_{n+1} = x \mid \mathcal{G}_n)$. The same line of reasoning justifies the stronger statement that for any $x_{n+1}, \ldots, x_T$,

$$\begin{aligned}\mathrm{P}(X_{n+1} = x_{n+1}, \ldots, X_T = x_T \mid \mathcal{F}_n) &= \mathrm{P}(X_{n+1} = x_{n+1}, \ldots, X_T = x_T \mid \mathcal{G}_n) \\ &= \prod_{k=n}^{T-1} p_{k,x_k,x_{x+1}}\end{aligned}$$

Going one more step, consider a function f that depends only on the future: $f = f(x_{n+1}, x_{n_2}, \ldots, x_T)$. Then

$$\mathrm{E}[f \mid \mathcal{F}_n] = \mathrm{E}[f \mid \mathcal{G}_n] \ . \tag{9}$$

If you like thinking about $\sigma-$algebras, you could say this by defining the *future* algebra, $\mathcal{H}_n$, that is determined only by information in the future. Then (9) holds for any f that is measurable with respect to $\mathcal{H}_n$.

A Markov chain is *homogeneous* if the transition probabilities do not depend on n. Most of the Markov chains that arise in modeling are homogeneous. Much of the theory is for homogeneous Markov chains. From now on, unless we explicitly say otherwise, we will assume that a Markov chain is homogeneous. Transition probabilities will be p_{yx} for every n.

The *forward equation* is the equation that describes how probabilities evolve over time in a Markov chain. Last week we saw that we could evolve the mean and variance of a linear Gaussian discrete time process ($X_{n+1} = AX_n + BZ_n$) using $\mu_{n+1} = A\mu_n$ and $C_{n+1} = AC_nA^t + BB^t$. This determines $X_{n+1} \sim \mathcal{N}(\mu_{n+1}, C_{n+1})$ from the information $X_n \sim \mathcal{N}(\mu_n, C_n)$. It is possible to formulate a more general forward equation that gives the distribution of X_{n+1} even if X_n is not Gaussian. But we do not need that here.

Let p_{yx} be the transition probabilities of a discrete state space Markov chain. Let $u_n(y) = \mathrm{P}(X_n = y)$. The forward equation is a formula for the numbers u_{n+1} in terms of the numbers u_n. The derivation is simple

$$\begin{aligned} u_{n+1}(x) &= \mathrm{P}(X_{n+1} = x) \\ &= \sum_{y\in\mathcal{S}} \mathrm{P}(X_{n+1} = x \mid X_n = y)\, \mathrm{P}(X_n = y) \\ u_{n+1}(x) &= \sum_{y\in\mathcal{S}} u_n(y) p_{yx} \qquad (10) \end{aligned}$$

The step from the first to second line uses what is sometimes called the *law of total probability*. The terms are rearranged in the last line for the following reason ...

We reformulate the Markov chain forward equation in matrix/vector terms. Suppose the state space is finite and $\mathcal{S} = \{x_1 \ldots, x_m\}$. We will say "state j" instead of "state x_j", etc. We write $u_{n,j} = \mathrm{P}(X_n = j)$ instead of $u_n(x_j) = \mathrm{P}(X_n = x_j)$. We collect the probabilities into a row vector $u_n = (u_{n,1}, \ldots, u_{n,m})$. The *transition matrix*, P, is the $m \times m$ matrix of transition probabilities. The (i,j) entry of P is $p_{ij} = \mathrm{P}(i \to j) = \mathrm{P}(X_{n+1} = j | X_n = i)$. The forward equation is

$$u_{n+1,j} = \sum_{i=1}^{n} u_{n,i} p_{ij} \ .$$

In matrix terms, this is just

$$u_{n+1} = u_n P \ . \qquad (11)$$

The row vector u_{n+1} is the product of the row vector u_n and the transition matrix P. It is a tradition to make u_n a row vector and put it on the left. You will come to appreciate the wisdom of this unusual choice over the coming weeks.

Once we have a linear algebra formulation, many tools of linear algebra become available. For example, powers of the transition matrix trace the evolution of u over several steps:

$$u_{n+2} = u_{n+1}P = (u_n P)\, P = u_n P^2 \; .$$

Clearly $u_{n+k} = u_n P^k$ for any k. This means we can study the evolution of Markov chain probabilities using the eigenvalues and eigenvectors of the transition matrix P.

The other major equation is the *backward equation*, which propagates conditional expectations backward in time. Take $\mathcal{F}_n$ to be the natural filtration generated by the path up to time n. Take $f(x_{[1:T]}) = V(x_T)$. This is a *final time payout*. The function is completely determined by the state of the system at time T. We want to characterize $f_n = \mathrm{E}[f|\mathcal{F}_n]$ and see how to calculate it using V and P. The characterization comes from the Markov property (9), which implies that f_n is a function of x_n. The backward equation determines this function.

The backward equation for Markov chains follows from the tower property (4). Use the transition probabilities (7). Since $\mathrm{E}[f_{n+1} \mid \mathcal{F}_n]$ is measurable with respect to $\mathcal{G}_n$, the expression $\mathrm{E}[f_{n+1} \mid \mathcal{F}_n]\,(x_n)$ makes sense:

$$\begin{aligned} f_n(x_n) &= \mathrm{E}[f_{n+1} \mid \mathcal{F}_n]\,(x_n) \\ &= \sum_{x_{n+1}\in\mathcal{S}} \mathrm{P}(X_{n+1} = x_{n+1} | X_n = x_n)\, f_{n+1}(x_{n+1}) \\ &= \sum_{x_{n+1}\in\mathcal{S}} p_{x_n x_{n+1}} f_{n+1}(x_{n+1}) \; . \end{aligned}$$

It is simpler to write the last formula for generic $x_n = x$ and $x_{n+1} = y$.

$$f_n(x) = \sum_{y\in\mathcal{S}} p_{xy} f_{n+1}(y) \; . \tag{12}$$

This is one form of the backward equation.

The backward equation gets its name from the fact that it determines f_n from f_{n+1}. If you think of n as a time variable, then time runs backwards for the equation. To find the solution, you start with the *final condition* $f_T(x) = V(x)$, then compute f_{T-1} using (12), and continue.

The backward equation may be expressed in matrix/vector terms. As we did when doing this for the forward equation, we suppose the state space is $\mathcal{S} = \{1, 2, \ldots, m\}$. Then we define the column vector $f_n \in \mathbb{R}^m$ whose components are $f_{n,j} = f_n(j)$. The elements of the transition matrix are $p_{ij} = \mathrm{P}(i \to j)$. The right side of (12) is the matrix/vector product Pf_{n+1}, so the equation is

$$f_n = P f_{n+1} \; . \tag{13}$$

The forward and backward equations use the same matrix P, but the forward equation multiplies from the left by a row vector of probabilities, while the backward equation multiplies from the right by a column vector of conditional probabilities.

The transition matrix if a homogeneous Markov chain is an $m \times m$ matrix. You can ask which matrices arise in this way. A matrix that can be the transition matrix for a Markov chain is called a *stochastic matrix.* There are two obvious properties that characterize stochastic matrices. The first is that $p_{ij} \geq 0$ for all i and j. The transition probabilities are probabilities, and probabilities cannot be negative. The second is that

$$\sum_{j=1}^{m} p_{ij} = 1 \quad \text{for all } i = 1, \ldots, m. \tag{14}$$

This is because $\mathcal{S}$ is a complete list of the possible states at time $n+1$. If $X_n = i$, then X_{n+1} is one of the states $1, 2, \ldots, m$. Therefore, the probabilities for the landing states (the state at time $n+1$) add up to one.

If you know P is a stochastic matrix, you know two things about its eigenvalues. One of those things is that $\lambda = 1$ is an eigenvalue. The proof of this is to give the corresponding eigenvector, which is the column vector of all ones: $\mathbf{1} = (1, 1, \ldots, 1)^t$. If $g = P\mathbf{1}$, then the components of g are, using (14),

$$g_i = \sum_{j=1}^{m} p_{ij} \mathbf{1}_j = \sum_{j=1}^{m} p_{ij} = 1 \ ,$$

for all i. This shows that $g = P\mathbf{1} = \mathbf{1}$. This result is natural in the Markov chain setting. The statement $f_{n+1} = \mathrm{E}[V(X_T) \mid \mathcal{F}_{n+1}] = 1$ for all $x_j \in \mathcal{S}$ means that given the information in $\mathcal{F}_{n+1}$, the expected value equals 1 no matter what. But $\mathcal{F}_n$ has less information. All you know at time n is that in the next step you will go to a state where the expected value is 1. But this makes the expected value at time n equal to 1 already.

The other thing you know is that if λ is an eigenvalue of P then $|\lambda| \leq 1$. This is a consequence of the *maximum principle* for P, which we now explain. Suppose $f \in \mathbb{R}^m$ is any vector and $g = Pf$. The maximum principle for P is

$$\max_i g_i \leq \max_j f_j \ . \tag{15}$$

Some simple reasoning about P proves this. First observe that if $h \in \mathbb{R}^m$ and $h_j \geq 0$ for all j, then

$$(Ph)_i = \sum_{j=1}^{m} p_{ij} h_j \geq 0 \quad \text{for all } i.$$

This is because all the terms on the right, p_{ij} and h_j are non-negative. Now let $M = \max f_j$. Then define h by $h_j = M - f_j$, so that $M\mathbf{1} = f + h$. Since $P(M\mathbf{1}) = M\mathbf{1}$, we know $g_i + (Ph)_i = M$. Since $(Ph)_i \geq 0$, this implies that $g_i \leq M$, which is the statement (15). Using similar arguments,

$$|g_i| \leq \sum_j p_{ij} |f_j| \leq \left(\sum_j p_{ij} \right) \max_j |f_j| \ ,$$

you can show that even if f is complex,

$$\max_i |g_i| \leq \max_j |f_j| \ . \tag{16}$$

This implies that if λ is an eigenvalue of P, then $|\lambda| \leq 1$. That is because the equation $Pf = \lambda f$ with $|\lambda| > 1$, violates (16) by a factor of $|\lambda|$.

A stochastic matrix is called *ergodic* if:

(i) The eigenvalue $\lambda = 1$ is simple.

(ii) If $\lambda \neq 1$ is an eigenvalue of P, then $|\lambda| < 1$.

A Markov chain is ergodic if its transition matrix is ergodic. (Warning: the true definition of ergodicity applies to Markov chains. There is a theorem stating that if $\mathcal{S}$ is finite, then the Markov chain is ergodic if and only if the eigenvalues of P satisfy the conditions above. Our definitions are not completely wrong, but they might be misleading.) Most of our examples are ergodic.

Last week we studied the issue of whether the Markov process (a linear Gaussian process last week, a discrete Markov chain this week) has a *statistical steady state* that it approaches as $n \to \infty$. You can ask the same question about discrete Markov chains. A probability distribution, π, is *stationary*, or *steady state*, or *statistical* steady state, if $u_n = \pi \Longrightarrow u_{n+1} = \pi$. That is the same as saying that $X_n \sim \pi \Longrightarrow X_{n+1} \sim \pi$. The forward equation (11) implies that a stationary probability distribution must satisfy the equation $\pi = \pi P$. This says that π is a *left eigenvector* of P with eigenvalue $\lambda = 1$. We know there is at least one left eigenvector with eigenvalue $\lambda = 1$ because $\lambda = 1$ is a right eigenvalue with eigenvector $\mathbf{1}$.

If $\mathcal{S}$ is finite, it is a theorem that the chain is ergodic if and only if it satisfies both of the following conditions

(i) There is a unique row vector π with $\pi = \pi P$ and $\sum_i \pi_i = 1$.

(ii) Let u_1 be any distribution on m states and $X_1 \sim u_1$. If $X_n \sim u_n$, then $u_n \to \pi$ as $n \to \infty$.

Without discussing these issues thoroughly, you can see the relation between these theorems about probabilities and the statements about eigenvalues of P above. If P has a unique eigenvalue equal to one and the rest less than one, then $u_{n+1} = u_n P$ has $u_n \to$ (the eigenvector), as $n \to \infty$. But π is the eigenvector corresponding to eigenvalue $\lambda = 1$. It might be that $u_n \to c\pi$, but we know $c = 1$ because $\sum_i u_{n,i} = 1$, and similarly for π.

5 Discrete random walk

This section discusses Markov chains where the states are integers in some range and the only transitions are $i \to i$ or $i \to i \pm 1$. The non-zero transition probabilities are $a_i = \mathrm{P}(i \to i-1) = p_{i,i-1}$, and $b_i = \mathrm{P}(i \to i) = p_{i,i}$, and

$c_i = \mathrm{P}(i \to i+1) = p_{i,i+1}$. These are called *random walk*, particularly if $\mathcal{S} = \mathbb{Z}$ and the transition probabilities are independent of i. A *reflecting* random walk is one where moves to $i < 0$ are blocked (reflected). In this case the state space is the non-negative integers $\mathcal{S} = \mathbb{Z}_+$, and $c_0 = 0$, and $b_0 = b + c$. For $i > 0$, $a_i = a$, $b_i = b$, and $c_i = c$. You can interpret the transition probabilities at $i = 0$ as saying that a *proposed* transition $0 \to -1$ is *rejected*, leading the state to stay at $i = 0$. You cannot tell the difference between a pure random walk and a reflecting walk until the *walker* hits the reflecting boundary at $i = 0$. You could get a random walk on a finite state space by putting a reflecting boundary also at $i = m - 1$ (chosen so that $\mathcal{S} = \{0, 1, \ldots m-1\}$ and $|\mathcal{S}| = m$). The transition probabilities at the right boundary would be $c_{m-1} = c$, $b_{m-1} = b + a$, and $a_{m-1} = 0$. These probabilities reject proposed $m - 1 \to m$ transitions.

The phrase *birth death process* is sometimes used for random walks on the state space $\mathbb{Z}_+$ with transition probabilities that depend in a more general way on i. The term birth/death comes from the idea that X_n is the number of animals at time n. Then a_i is the probability that one dies, c_i is the probability that one is born, and $b_i = 1 - a_i - c_i$ is the probability that the probability does not change. You might think that $i = 0$ would be an *absorbing state* in that $\mathrm{P}(0 \to 1) = 0$ because no animals can be born if there are no animals. Birth/death processes do not necessarily assume this. They permit storks.

Urn processes are a family of probability models that give rise to Markov chains on the state space $\{0, \ldots, m-1\}$. An urn is a large clay jar. In urn processes, we talk about one or more urns each with balls of one or more colors. Here is an example that has one urn, two colors (red and blue) and a probability, q. At each stage there are m balls in the urn (can't stay with $m - 1$ any longer. Some are red and the rest blue. You choose one ball from the urn, each with the same probability to be chosen – a *well mixed* urn. You replace that ball with a new one whose color is red with probability q and blue with probability $1 - q$. Let X_n be the number of red balls at time n. In this process $X_n \to X_n - 1$ if you you choose a red ball and replace it with a blue ball. If you choose a blue ball and replace it with a red ball, then $X_n \to X_n + 1$. If you replace red with red, or blue with blue, then $X_{n+1} = X_n$.

A matrix is *tridiagonal* if all its entries are zero when $|i - j| > 1$. It's non-zeros are on the main diagonal, the one super-diagonal, and one sub-diagonal. This section is about Markov chains whose transition matrix is tri-diagaonal. Consider, for a simple example, the random walk with a reflecting boundary at $i = 0$ and $i = 5$. Suppose the transition probabilities are $a = \frac{1}{2}$, and $b = c = \frac{1}{4}$. The transition matrix is the 6×6 matrix

$$P = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} . \tag{17}$$

There is always confusion about how to number the first row and column.

Markov chain people like to start the numbering with $i = 0$, as we did above. The tradition in linear algebra is to start numbering with $i = 1$. These notes try to keep the reader on her or his toes by doing both, starting with $i = 1$ when describing the entries in P, and starting with $i = 0$ when describing the Markov chain transition probabilities. The transition matrix P above has $p_{12} = \frac{1}{4}$, which indicates that if $X_n = 1$, there is a .25% chance that $X_{n+1} = 2$. Since X_n is not allowed to go lower than 1 or higher than 2, the rest of the probability must be to stay at $i = 1$, which is why $p_{11} = \frac{3}{4}$. In the first column we have $p_{21} = \frac{1}{2}$, which is the probability of a $2 \to 1$ transition. From state $i = 2$ there are three possible transitions, $2 \to 1$ with probability $\frac{1}{2}$ as we just said, $2 \to 2$ with probability $\frac{1}{4}$, and $2 \to 3$, with probability $\frac{1}{4}$. The row sums of this matrix are all equal to one, and so are most of the column sums. But the first column sum is $p_{11} + p_{21} = \frac{5}{4} > 1$ and the last is $p_{56} + p_{66} = \frac{3}{4} < 1$.

My computer says that

$$P^2 = \begin{pmatrix} 0.6875 & 0.2500 & 0.0625 & 0 & 0 & 0 \\ 0.5000 & 0.3125 & 0.1250 & 0.0625 & 0 & 0 \\ 0.2500 & 0.2500 & 0.3125 & 0.1250 & 0.0625 & 0 \\ 0 & 0.2500 & 0.2500 & 0.3125 & 0.1250 & 0.0625 \\ 0 & 0 & 0.2500 & 0.2500 & 0.3125 & 0.1875 \\ 0 & 0 & 0 & 0.2500 & 0.3750 & 0.3750 \end{pmatrix} .$$

For example, $p^{(2)}_{4,2} = \frac{1}{4}$, which says that the probability of going from 4 to 2 in two hops is $\frac{1}{4}$. The only way to do that is $X_n = 4 \to X_{n+1} = 3 \to X_{n+2} = 2$. The probability of this is $p_{43}p_{32} = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$. The probability of $3 \to 3$ in two steps is $p^{(2)}_{33} = \frac{5}{16}$. The three paths that do this, with their probabilities, are $\mathrm{P}(3 \to 2 \to 3) = \frac{1}{2}\frac{1}{4} = \frac{2}{16}$, and $\mathrm{P}(3 \to 3 \to 3) = \frac{1}{4}\frac{1}{4} = \frac{1}{16}$, and $\mathrm{P}(3 \to 4 \to 3) = \frac{1}{4}\frac{1}{2} = \frac{2}{16}$. These add up to $\frac{5}{16}$. The matrix P^2 is the transition matrix for the Markov chain that says "take two hops with the P chain. Therefore its row sums should equal 1, as $p^{(2)}_{11} + p^{(2)}_{12} + p^{(2)}_{13} = \frac{11}{16} + \frac{4}{16} + \frac{1}{16} = 1$.

My computer says that, for $n = 100$ (think $n = \infty$),

$$P^n = \begin{pmatrix} 0.5079 & 0.2540 & 0.1270 & 0.0635 & 0.0317 & 0.0159 \\ 0.5079 & 0.2540 & 0.1270 & 0.0635 & 0.0317 & 0.0159 \\ 0.5079 & 0.2540 & 0.1270 & 0.0635 & 0.0317 & 0.0159 \\ 0.5079 & 0.2540 & 0.1270 & 0.0635 & 0.0317 & 0.0159 \\ 0.5079 & 0.2540 & 0.1270 & 0.0635 & 0.0317 & 0.0159 \\ 0.5079 & 0.2540 & 0.1270 & 0.0635 & 0.0317 & 0.0159 \end{pmatrix} . \tag{18}$$

These numbers have the form $p^{(\infty)}_{i,j} = 2^{-j}/s$, where $s = \sum_{j=1}^{6} 2^{-j}$. The formula for s comes from the requirement that the row sums of $p^{(\infty)}$ are 1. The fact that $p^{(\infty)}_{i,j+1}/p^{(\infty)}_{ij} = \frac{1}{2}$ comes from the following theory. Think of starting with $X_0 = i$. Then $u_{1,ij} = \mathrm{P}(X_1 = j|X_0 = i) = p_{ij}$. Similarly, $u_{n,j} = \mathrm{P}(X_n = j|X_0 = i) = p^{(n)}_{ij}$. If this P is ergodic (it is), then $u_{n,j} \to \pi_j$ as $n \to \infty$. This implies that $p^{(n)}_{ij} \to \pi_j$ as $n \to \infty$ for any i. The P^n in (18) seems to fit that, at least in the fact that all the rows are the same because the limit of $p^{(n)}_{ij}$ is independent of i.

The equation $\pi = \pi P$ determines π. For our P in (17), these equations are

$$\begin{aligned}
\pi_1 &= \pi_1 \tfrac{3}{4} + \pi_2 \ \tfrac{1}{2} \\
\pi_2 &= \pi_1 \tfrac{1}{4} + \pi_2 \tfrac{1}{4} + \pi_3 \tfrac{1}{2} \\
&\ \vdots \\
\pi_5 &= \pi_4 \tfrac{1}{4} + \pi_5 \tfrac{1}{4} + \pi_6 \tfrac{1}{2} \\
\pi_6 &= \pi_5 \tfrac{1}{4} + \pi_6 \ \tfrac{1}{2} \ .
\end{aligned}$$

You can check that a solution is $\pi_j = c2^{-j}$. The value of c comes from the requirement that $\sum_j \pi_j = 1$ (π is a probability distribution).