Plan Markov Chains

1. The Markov Property – Motivation
3. Examples
3. Transition Probabilities

   Next Time: More Markov

   Reading: G&S 6.3, generating function sheet
   Homework 3 due Thursday

Heuristic Definition 1. A stochastic process has the *Markov property* if the future of the process is conditionally independent of the past given the present state.

Heuristic Definition 2. A *Markov chain* is a stochastic process $X = (X_n)_{n\in\mathbb{Z}_\oplus}$ with the Markov property. We assume that the X_n take values in some common space $\mathcal{S}$, called the *state space* of the chain.

Note 3. We allow $\mathcal{S}$ to be countable (a "countable-state Markov chain") or uncountable (a "general-state Markov chain"). The former allows some simplication of the theory.

Note 4. The index set $\mathbb{Z}_\oplus$ could in principle be replaced by any ordered, countable set but we'll stick to $\mathbb{Z}_\oplus$. An important class of processes, including the Weiner process, have a Markovian structure with a continuous index set. We will call these *Markov processes* (or more pedantically continuous-time Markov processes). There is some disagreement about the applicability of the word "chain," but this seems to be the most common convention. If the index set is multi-dimensional, the process is often referred to as a *Markov random field*.

Motivating Examples 5. Markov chains have important applications in a wide range of applications and disciplines.

1. White Noise Let Ξ_n be IID real-valued random variables. Then $\Xi = (\Xi_n)_{n\geq 0}$ is a Markov chain. This is clear intuitively: the future is independent of the past because all the Ξ_n are independent.
   This is true, but if we look closely, we why we might need to be careful with infinite collections. For any suitable h and all $n \geq 0$, we have

$$\mathsf{E}\left[h(\Xi_{n+1}, \Xi_{n+2}, \ldots) \mid \Xi_0, \ldots, \Xi_n\right] = \mathsf{E}h(\Xi_{n+1}, \Xi_{n+2}, \ldots \mid \Xi_n) = \mathsf{E}h(\Xi_{n+1}, \Xi_{n+2}, \ldots). \qquad (1)$$

   But what are suitable h? For each n, we need h measurable with respect to the σ-field $\sigma(X_{n+1}, \ldots,)$
   Define the *tail σ-field* of the process by

$$\mathcal{T} = \bigcap_{n\geq 0} \sigma(\Xi_n, \Xi_{n+1}, \ldots). \qquad (2)$$

   This σ-field, which we need not be trivial, describes events that occur "for large n." Examples?
   Kolmogorov's Zero-One Law. For independent Ξ_n, every $A \in \mathcal{T}$ has $\mathsf{P}(A) = 0$ or 1.

2. Random Walks
   Let Ξ_n be IID real-valued random variables and S_0 be a real-valued random variable with arbitrary distribution that is independent of the Ξ_ns. Then, for $n \geq 1$,

$$S_n = S_{n-1} + \Xi_n. \qquad (3)$$

Then $S = (S_0, S_1, \ldots)$ is a *random walk.*
Note that we can write

$$S_n = S_0 + \sum_{k=1}^{n} \Xi_k, \tag{4}$$

but the recursive definition above has value.
For any suitable function h and every $n \geq 0$, we have

$$\begin{aligned}
&\mathsf{E}\left(h(S_{n+1}, S_{n+2}, \ldots) \mid S_0, \ldots S_n\right) \\
&\qquad = \mathsf{E}(h(S_n + X_{n+1}, S_n + X_{n+1} + X_{n+2}, \ldots) \mid S_0, \ldots, S_n) &(5)\\
&\qquad = \mathsf{E}(h(S_n + X_{n+1}, S_n + X_{n+1} + X_{n+2}, \ldots) \mid S_n) &(6)\\
&\qquad = \mathsf{E}\left(h(S_{n+1}, S_{n+2}, \ldots) \mid S_n\right). &(7)
\end{aligned}$$

So any question we might ask about the future of the process given its entire history depends only on its present state – loosely, the process "forgets" its past. This is the Markov property at work.
As we have seen, we can build many processes from this starting point. For example, Define $H_0 = \max(S_0, 0)$ and for $n \geq 1$, let

$$H_n = \max(H_{n-1} + \Xi_n, 0). \tag{8}$$

Then the process $H = (N_n)_{n\geq 0}$ is called a *random walk on the half-space.* It might serve as a simple model for a dam or a bank-account.

3. Time-Series Models
One simple generalization of the random walk is to define for $n \geq 1$ Then, for $n \geq 1$,

$$S_n = \alpha S_{n-1} + \Xi_n, \tag{9}$$

for some real-valued parameter α. This is called an *order 1 autoregressive or* $AR(1)$ *process.* It has the Markov property as expressed above by the same basic argument.
This extends to $AR(r)$ by defining initial values $(S_{-r+1}, \ldots, S_0)$ to have an arbitrary distribution and then defining for $n \geq 1$,

$$S_n = \alpha_1 S_{n-1} + \alpha_2 S_{n-2} + \cdots + \alpha_r S_{n-r} + \Xi_n, \tag{10}$$

for parameter $\alpha = (\alpha_1, \ldots, \alpha_r) \in \mathbb{R}^r$.
But wait! Can this be Markovian?
Not directly, no. But let $X_n = (S_{n-r+1}, \ldots, S_n)^T$. Then there is an $r \times r$ matrix A such that $X_0 = (S_{-r+1}, \ldots, S_0)$ and for $n \geq 1$

$$X_n = AX_{n-1} + \Xi_n. \tag{11}$$

How familiar. This is a special case of a so-called *state-space model* common in time-series analysis.
Even more generally, an *autoregressive moving-average process*, or $ARMA(r, m)$ is of the form

$$X_n = \sum_{i=1}^{r} \alpha_i X_{n-i} + \sum_{j=1}^{m} \beta_j \Xi_{n-j}. \tag{12}$$

This can be made Markovian in several ways as you'll see on homework.

4. Control systems

   Consider a cruise control system on your car. At time n, the system has access to inputs such as the car's speed, engine status, turning angle, and so on. Put these together in a vector V_n. The system then calculates a set of changes to the system, a vector U_n, designed to keep the vehicle's speed within specified parameters. U_n is called a control vector. The vector $X_n = (U_n, V_n)$ can often be modeled as a Markov chain.

   More generally, we might have a system governed by a model such as

$$V_n = \theta_{n-1} V_{n-1} + U_{n-1} + \Xi_n \tag{13}$$

$$\theta_n = \alpha \theta_{n-1} + \Psi_n, \tag{14}$$

   where (Ξ_n, Ψ_n) are Normal and where U_n depends only on previous Vs.

   The behavior of this system can depend strongly on α, but if α is unknown, we need to statistically infer α and the θ_ns sequentially as we control the process by choosing the Us. We want to understand the behavior of the process, how to estimate the underlying parameters, and how to control the process within target specifications.

5. Queues

   Suppose the time between successive customer arrivals at a service center are IID random variables $\Xi_1, \Xi_2, \ldots$ with $\mathsf{F}\ G$. The arrival time of the nth customer is then given by $S_0 = 0$ and for $n \geq 1$,

$$T_n = T_{n-1} + \Xi_n, \tag{15}$$

   which looks familiar.

   Customers that arrive wait in a queue until they are serviced by a single server. Assume that the nth customer requires service time S_n where the $S_1, S_2, \ldots$ are IID random variables with $\mathsf{F}\ H$.

   Let $N(t)$ be the number of customers in the queue at time t.

   This system is called a $GI/G/1$ queue.

   Although $N(t)$ is a continuous-time process, it can be analyzed fruitfully by considering emphembedded Markov chains.

   Let $N_n = N(T_n-)$, the number of customers *just before* the nth arrival, with $N_0 = 0$. We will see that under some conditions this is a Markov chain with

$$\mathsf{P}\{N_{n+1} = k \mid N_n = j, N_{n-1}, \ldots, N_0\} = p(k-j). \tag{16}$$

   Conditions are needed on either G or H because transitions of N_n can depend on the history in a non-Markovian way.

   Two important cases are as follows:

   – $GI/M/1$ queue: H has an Exponential distribution
   – $M/G/1$ queue: G has an Exponential distribution

   We are interested in understanding the performance of the system in terms of waiting times, efficiency, and stability.

6. Networks

   We can view many computer and communications networks as systems of queues. Similar questions of performance arise.

7. Storage and Insurance Models
   An arrival process like the above can represent the times of withdrawals/additions from/to a dam (or other resource resevoir) or claims on an insurance company.
   The analog of the service times might be IID random variables $Y_1, Y_2, \ldots$, representing the amount by which the dam's supply changes or the amount of a claim.
   We want to understand the long-run stability of the process. What is the chance that a dam will overflow or that an insurance company will go bankrupt.
8. Economic and Financial Models
   A variety of time series and control models are used in Economics and Finance. Example: Currency exchange rates. Here we ask questions about long-run behavior and short-run large deviations of the process. Will a currency crisis ensue? What policies on the part of government will stabilize the system?
9. Population and Genetic Models
   Consider the evolution of populations across generations. Let Z_n represent the number of individuals in a given species at the nth generation with $Z_0 = c$, for some constant.
   Assume that the ith member of a generation has a family of size M_{ni}, the collection of which are IID.
   The process $Z = (Z_n)_{n \geq 0}$ is called a *branching process*. Questions of interest about Z include the probability of eventual extinction, long-run stability of population size. Generalizations to multiple populations and more sophisticated dynamics are in common use.

   A fundamental process in evolutionary biology is the frequencies of different alleles (called gene frequencies) in a population.
   For example, the Hardy-Weinberg equilibrium model gives a case in which there is no change in gene frequencies.
   Given a population sexually reproducing individuals, assume the following.

   A. The population is large (e.g., infinite) and remains so.
   B. No flow of genes into or out of the population
      (that is, no migration).
   C. No mutation.
   D. All genotypes have the same rate of reproductive success.
   E. Mating is purely random.

   Then, gene frequencies in the population will attain an equilibrium.

Genotype Frequencies	Gene (Allele) Frequencies
$D =$ Frequency of AA	$p =$ Frequency of A $= D + H/2$
$H =$ Frequency of Aa	$q =$ Frequency of a $= R + H/2$
$R =$ Frequency of aa	

   Under the Hardy-Weinberg assumptions, after one generation of random mating, each individual's alleles are randomly and independently assigned A or a with probabilities p and q respectively. So,

$$D' = \text{Frequency of AA} = p \cdot p = p^2 \tag{17}$$

$$H' = \text{Frequency of Aa} = p \cdot q + q \cdot p = 2pq \tag{18}$$

$$R' = \text{Frequency of aa} = q \cdot q = q^2, \tag{19}$$

and

$$p' = D' + H'/2 = p^2 + pq = p(p+q) = p \tag{20}$$
$$q' = R' + H'/2 = q^2 + pq = q(p+q) = q. \tag{21}$$

The gene frequencies haven't changed!
In general, we are interested in Evolution: cross-generational change in a population of organisms that involves changes in gene frequency. To do that, we model gene-frequency changes across generations as a Markov chain.

10. Markov Chain Monte Carlo
A fundamental question in Bayesian statistics is how to compute the posterior distribution of parameters in a statistical model.
Suppose that we have a statistical model $\{\mathsf{P}_\theta\}_{\theta\in\Theta}$ given by a likelihood $\ell(\theta) \equiv f_\theta(\boldsymbol{Y})$ for data Y. We put a prior distribution $\pi(\theta)$ on the parameter space. The posterior is given by

$$\pi(\theta \mid \boldsymbol{Y}) = \frac{\ell(\theta)\pi(\theta)}{\int \ell(\eta)\pi(\eta)\, d\eta}. \tag{22}$$

In many realistic models this is difficult to compute exactly.
But it turns out that we can define a Markov chain X_n so that asymptotically the distribution of X_n equals $\pi(\cdot \mid \boldsymbol{Y})$. That is,

$$\mathsf{P}\{X_n \in \mathcal{A}\} \to \pi(A \mid \boldsymbol{Y}). \tag{23}$$

The key questions here is how to compute a chain that is easy to compute and that makes this approximation good for a reasonable sized n.

11. Communication Systems
We'll see a lot more of these when we talk about information theory.

Working Definition 6. A countable-state *Markov chain* with state space $\mathcal{S}$ is an $S^{\mathbb{Z}_\oplus}$-valued stochastic process $X = (X_n)_{n\geq 0}$ such that

$$\mathsf{P}\{X_n = s_n \mid X_{n-1} = s_{n-1}, \ldots, X_1 = s_1, X_0 = s_0\} = \mathsf{P}\{X_n = s_n \mid X_{n-1} = s_{n-1}\}\,. \tag{24}$$

for all $n \geq 1$ and all $s_0, \ldots, s_n \in \mathcal{S}$. If the right-hand side does not depend on n explicitly, then X is said to be *time homogeneous* (or just homogeneous).

Equation (24) gives a version of the Markov property for these processes. We will deal with homogeneous chains unless otherwise indicated.

The behavior of the (time homogeneous) process is consquently governed by two features:

- The initial distribution. Let μ denote the distribution of X_0.
- The transition probabilities. Let $P(s, A) = \mathsf{P}\{X_n \in A \mid X_{n-1} = s\}$.

The object $P(s, A)$ is a transition probability kernel, defined below. When $\mathcal{S}$ is countable, the latter can be written as

$$\mathsf{P}(s, A) = \sum_{s' \in A} \mathsf{P}\{X_n = s' \mid X_{n-1} = s\} = \sum_{s' \in A} P(s, \{s'\}). \tag{25}$$

In this case, we define $P(s, s') \equiv P(s, \{s'\})$ and call it the *transition probability matrix*. This must satisfy $P(s, s') \geq 0$ and $P(s, \mathcal{S}) = 1 = \sum_{s' \in \mathcal{S}} P(s, s')$.

Another statement of the Markov property in the countable case is as follows:

$$\mathsf{P}(X_0 = s_0, X_1 = s_1, \ldots, X_n = s_n) = \mu(s_0) P(s_0, s_1) P(s_1, s_2) \cdots P(s_{n-1}, s_n). \tag{26}$$

The initial distribution and the transition probability matrix determine the probability of every sample path up to time n.

Yet another conceptually satisfying form of the Markov property is that for suitable functions h on the sample paths of the process:

$$\mathsf{E}_\mu \left(h(X_{n+1}, X_{n+2}, \ldots) \mid X_0, \ldots, X_{n-1}, X_n = s \right) = \mathsf{E}_s h(X_1, X_2, \ldots). \tag{27}$$

For general-state Markov chains, we cannot define the transition probability matrix, so we must use the transition probability kernel from which the former is derived. This requires a careful definition.

Definition 7. Let $(\mathcal{S}, \mathcal{B})$ be a measurable space. Let $P(x, A)$ be a function on $\mathcal{S} \times \mathcal{B}$ such that

1. For each $A \in \mathcal{B}$, $x \mapsto P(x, A)$ is a measurable function on $\mathcal{S}$.
2. For each $x \in \mathcal{S}$, $A \mapsto P(x, A)$ is a probability measure on $\mathcal{B}$.

We call such a P a *transition probability kernel.*

Rigor Alert 8. We are essentially ignoring the definition of the σ-field for the random function definition of the stochastic process. For the most part, we can do so because it just works as you might expect, but I'll give more details if and when it becomes necessary.

Definition 9. A stochastic process X on $(\mathcal{S}_+^{\mathbb{Z}}, \mathcal{F})$ is called a *time-homogeneous Markov chain* with transition probability kernel $P(x, A)$ and initial distribution μ if the finite-dimensional distributions of X satisfy

$$\mathsf{P}_\mu(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) = \int_{A_0} \cdots \int_{A_{n-1}} \mu(ds_0) P(s_0, ds_1) \cdots P(s_{n-2}, ds_{n-1}) P(s_{n-1}, A_n). \tag{28}$$

Notes. (i) The P_μ is a technicality, but think of it as reminding us that the initial state has distribution μ. The analogous expected value operator is E_μ. (ii) The $\mu(ds)$ or $P(\cdot, ds)$ notation means integration against the corresponding distribution. If a random variable Y has distribution F, we write $\mathsf{P}\{Y \in A\} = \int_A dF = \int_A F(ds)$.

An equivalent form of the Markov property is that for every bounded and measurable, real-valued function h on the sample paths of the process

$$\mathsf{E}_\mu \left(h(X_{n+1}, X_{n+2}, \ldots) \mid X_0, \ldots, X_{n-1}, X_n \text{ near } s \right) = \mathsf{E}_s h(X_1, X_2, \ldots). \tag{29}$$