Markov Chains (With Some MCMC at the End)

Goals of this lecture:

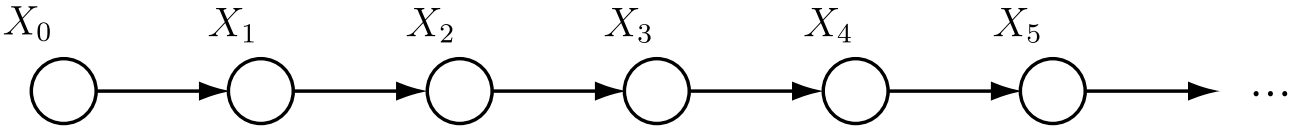
- Introduce Markov chains
- Highlight the properties of Markov chains needed to understand MCMC
- Explain what the limiting distribution of a Markov chain is and why it is useful for Monte Carlo approximation.

Definition of a Markov chain:

X_t , $t = 0, 1, 2 \ldots$, having a joint distribution such that

$$P(X_t|X_{t-1}, X_{t-2}, \ldots, X_0) = P(X_t|X_{t-1}) \ \forall \, t$$

Schematically:



And the variables X_t can be scalars or vectors.

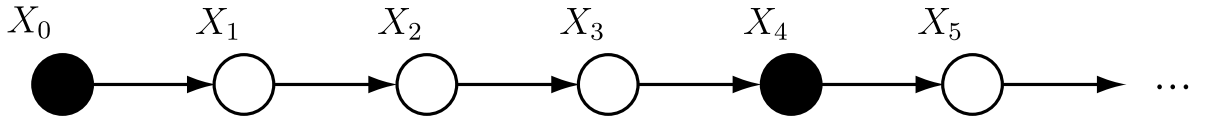
- For most of the *examples* in this lecture, X_t will be univariate.
- When X_t includes all the variables in an MCMC application, however, it is typically high-dimensional.

A genetics example—the Wright-Fisher population:

Let X_t denote the number of alleles of type A in a Wright-Fisher population of size N diploids. Then:

$$P(X_t = j | X_{t-1} = i) = \frac{2N!}{(2N-j)!j!} \left(\frac{i}{2N}\right)^j \left(\frac{2N-i}{2N}\right)^{2N-j}$$

- Important point: conditional independence $\neq$ independence. If you don't know X_{t-1}, then, of course X_t depends on X_{t-2}



- NB: Do not think that MCMC is useful in genetics because the problems in genetics involve Markov chains. The Markov chain underlying Markov chain Monte Carlo and any Markov chains involved in a statistical genetics model may be quite distinct entities.

Transition Probability Matrices:

We can write down the transition probabilities from all values of X_t to all values of X_{t+1} in a matrix:

$$\begin{array}{c} \nearrow \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ 2N \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \cdots & 2N \end{array} \\ \begin{pmatrix} P(0 \to 0) & P(0 \to 1) & P(0 \to 2) & \cdots & P(0 \to 2N) \\ P(1 \to 0) & P(1 \to 1) & P(1 \to 2) & \cdots & P(1 \to 2N) \\ P(2 \to 0) & P(2 \to 1) & P(2 \to 2) & \cdots & P(1 \to 2N) \\ P(3 \to 0) & P(3 \to 1) & P(3 \to 2) & \cdots & P(3 \to 2N) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ P(2N \to 0) & P(2N \to 1) & P(2N \to 2) & \cdots & P(2N \to 2N) \end{pmatrix} \end{array}$$

Here we use $P(i \to j)$ as a shorthand for $P(X_{t+1} = j | X_t = i)$.

Such *transition probability matrices* form the basis for much of the classical analysis of Markov chains.

The above t.p.m. is a bit unwieldy for examples, so consider another simple model. . .

Random walk with scattering boundaries:

Imagine a random walk on the integers from 1 to 5. From 1 or 5, the walk goes to any state with equal probability, and from the remaining states, the walk may take steps of size 0 or 1, but they are biased toward the center.

For example, consider the t.p.m:

$$\mathbf{P} = \begin{array}{c} \nearrow \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} .2 & .2 & .2 & .2 & .2 \\ .2 & .3 & .5 & 0 & 0 \\ 0 & .3 & .4 & .3 & 0 \\ 0 & 0 & .5 & .3 & .2 \\ .2 & .2 & .2 & .2 & .2 \end{pmatrix} \end{array}$$

Computer Demo: `tpm`

Some things to note:

- The rows of $\mathbf{P}$ sum to one—they are probabilities that must sum to one
- The columns need not sum to one
- Probabilities after one step can be computed by matrix multiplication. *i.e.*, if: $\boldsymbol{v_0} = (0, 1, 0, 0, 0)$, then the probabilities of being in states 1,2,. . . ,5 after one step of the chain are given by:

$$\boldsymbol{v_0}\mathbf{P} = \boldsymbol{v_1} = (.2, .3, .5, 0, 0)$$

- Probabilities after two steps are:

$$\boldsymbol{v_1}\mathbf{P} = (\boldsymbol{v_0}\mathbf{P})\mathbf{P} = \boldsymbol{v_2}$$

- and probabilities after n steps are

$$\boldsymbol{v_n} = \boldsymbol{v_0}\mathbf{P}^n$$

  Vectors $\boldsymbol{v_t}$ are taken to be row vectors.

Limiting Distributions:

A class of Markov chains called *ergodic Markov chains*[1] have the property that

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix}$$

where π is (in this discrete case) a vector referred to as the *limiting distribution* of the ergodic Markov chain.

Take our random walk example:

$$\mathbf{P}^1 = \begin{array}{c} \nearrow \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} .2 & .2 & .2 & .2 & .2 \\ .2 & .3 & .5 & 0 & 0 \\ 0 & .3 & .4 & .3 & 0 \\ 0 & 0 & .5 & .3 & .2 \\ .2 & .2 & .2 & .2 & .2 \end{pmatrix} \end{array}, \quad \mathbf{P}^2 = \begin{array}{c} \nearrow \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.12 & 0.20 & 0.36 & 0.20 & 0.12 \\ 0.10 & 0.28 & 0.39 & 0.19 & 0.04 \\ 0.06 & 0.21 & 0.46 & 0.21 & 0.06 \\ 0.04 & 0.19 & 0.39 & 0.28 & 0.10 \\ 0.12 & 0.20 & 0.36 & 0.20 & 0.12 \end{pmatrix} \end{array}$$

[1] Conditions conferring ergodicity will be discussed a little later.

$$\mathbf{P}^3 = \begin{array}{c} \nearrow \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.088 & 0.216 & 0.392 & 0.216 & 0.088 \\ 0.084 & 0.229 & 0.419 & 0.202 & 0.066 \\ 0.066 & 0.225 & 0.418 & 0.225 & 0.066 \\ 0.066 & 0.202 & 0.419 & 0.229 & 0.084 \\ 0.088 & 0.216 & 0.392 & 0.216 & 0.088 \end{pmatrix} \end{array}$$

$$\mathbf{P}^5 = \begin{array}{c} \nearrow \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.075 & 0.219 & 0.412 & 0.219 & 0.075 \\ 0.074 & 0.220 & 0.415 & 0.218 & 0.073 \\ 0.072 & 0.220 & 0.415 & 0.220 & 0.072 \\ 0.073 & 0.218 & 0.415 & 0.220 & 0.074 \\ 0.075 & 0.219 & 0.412 & 0.219 & 0.075 \end{pmatrix} \end{array}$$

$$\mathbf{P}^\infty = \begin{array}{c} \nearrow \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0.07317 & 0.2195 & 0.4146 & 0.2195 & 0.07317 \\ 0.07317 & 0.2195 & 0.4146 & 0.2195 & 0.07317 \\ 0.07317 & 0.2195 & 0.4146 & 0.2195 & 0.07317 \\ 0.07317 & 0.2195 & 0.4146 & 0.2195 & 0.07317 \\ 0.07317 & 0.2195 & 0.4146 & 0.2195 & 0.07317 \end{pmatrix} \end{array}$$

The limiting distribution in this case is

$$\pi = (0.07317,\ 0.2195,\ 0.4146,\ 0.2195,\ 0.07317)$$

- This means that if you start the chain from any of the five states, after a sufficient (and not very many in this case) number of steps, the probability that it will be found in any of the five states is essentially independent of its starting state.
- And, as we have already seen, it means that a "time-average" over the chain converges to the limiting distribution.
- This second point is asserted by the weak law of large numbers for ergodic Markov chains which states that as the number of transitions (or steps) in an ergodic chain tends to infinity, the proportion of time the chain spends in a state tends to the limiting probability (*i.e.*, the appropriate component of π) of that state[2].
- Thus, the states visited by an ergodic Markov chain may be used to compute a Monte Carlo average. That is MCMC.

[2] See Feller (1957)

The use of Markov Chains in Monte Carlo:

Monte Carlo with a Markov chain is pretty much the same as before:

$$\mathbb{E}[g(X)] = \frac{1}{n}\sum_{i=1}^{n} g(x^{(i)})$$

except that now, the $x^{(i)}$ are states visited by a Markov chain having a limiting distribution that we wish to sample from.

To implement this we must be able to construct an appropriate Markov chain. It must:

- be ergodic
- have the right limiting distribution

In the next session, Matthew will explain how that is done using the *Metropolis-Hastings* algorithm.

Important points to keep in mind:

1. If you can simulate *independent* samples, $x^{(1)}, \ldots, x^{(n)}$, then, by all means, do so, and avoid MCMC if you can.
2. MCMC is most useful when the desired distribution to be sampled from is "known only up to scale"—this means that the "shape" of the distribution is known but its normalizing constant is not.

   A germane example of a distribution known only up to scale is the Bayesian posterior distribution of θ given data Y:

   $$P(\theta|Y) = \frac{P(Y|\theta)P(\theta)}{\int_\theta P(Y|\theta)P(\theta)d\theta}$$

   The denominator is a constant w.r.t. θ. It is the (typically unknown) normalizing constant. However, the numerator is the joint density of the data Y and the parameters θ, which is often easy to compute. This is why MCMC is so useful to Bayesians.

Conditions ensuring a chain is ergodic:

A Markov chain is ergodic if:

1. *It has no transient states*, *i.e.*, there are no states with an expected time to recurrence of ∞. (This is not an issue with a finite state space).
2. *It is irreducible*, *i.e.*, any state in the state space is reachable from any other state in the state space in a finite number of steps.
3. *It is aperiodic*, *i.e.*, there exists no pair of states i and j such that the probability of reaching j from i is non-zero only if the number of steps is an integer multiple of some period τ.

The following pages have extra information about Metropolis Hastings sampling that we will cover differently this year. But it might be interesting reading for some!

General balance and the stationary distribution of a Markov chain:

Recall that we wish to construct an ergodic Markov chain with limiting distribution π equal to some distribution that we wish to sample from. How might we do that?

One possibility is via a theorem which tells us about a property of π:

If π is the limiting distribution of the ergodic Markov chain with t.p.m. $\mathbf{P}$, then π is the unique stationary distribution that satisfies the general balance equation

$$\pi\mathbf{P} = \pi$$

So, if we can find $\mathbf{P}$ such that its unique stationary distribution is π, and π is the distribution we wish to sample from, then we can use the chain defined by $\mathbf{P}$ in MCMC.

This is, however, not a tractable approach in any problem of consequence. In general it will be harder to (or just as difficult to) solve the general balance equation for π as it will be to draw independent samples from π.

The main problem with solving the general balance equation is that all possible states in the distribution π must be considered simultaneously. In most interesting problems, that number of states can be astronomically huge.

The solution to this conundrum is to not try to solve the general balance equation directly, but, instead satisfy a "locally-defined" balance condition (so that only two states in the state space need be considered simultaneously—rather than all of the states, simultaneously), AND do this in a way that ensures that general balance is satisfied.

This is quite a lovely thing.

Time-reversible Markov chains and detailed balance:

A special class of Markov chains are called “time-reversible” Markov chains, because if you were to watch them running “backward in time” they would look just the same as the chain running “forward in time.”

The salient feature of such chains is that they satisfy the *detailed balance* (also called the *local balance*) condition.

Detailed balance with respect to π between a pair of states i and j, is satisfied by a Markov chain with a stationary distribution π, and a t.p.m $\mathbf{P}$ having elements $P(x \to y)$ if

$$\pi(i)P(i \to j) = \pi(j)P(j \to i).$$

It is easy to show (by summing over all states i) that if detailed balance holds for every pair of states, then general balance is also satisfied.

The Metropolis-Hastings Algorithm[3]:

The M-H algorithm provides a way to perform steps in a Markov chain that satisfy detailed balance.

Imagine you wish to simulate a dependent sample from a target distribution f, and you are currently in state i. The recipe is:

- Propose changing the state from state i to a new state j. Draw the state j from a proposal distribution, $q(j|i)$ which may be conditional on i.
- Accept and move to the new state j with probability R equal to the lesser of 1 or the *Hastings ratio*:

$$R(i \to j) = \min \left\{ 1, \frac{f(j)}{f(i)} \times \frac{q(i|j)}{q(j|i)} \right\}$$

  If you don't accept the move to state j, then stay where you are.

[3] Metropolis et al. (1953) and Hastings (1970).

M-H algorithm satisfies detailed balance:

To show the M-H algorithm satisfies detailed balance w.r.t. f, we write down $P(i \to j)$ and $P(j \to i)$, for any i and j, under the M-H algorithm.

$P(i \to j)$ is the product of the probabilities that we:

1. Propose going from i to j. $\quad \text{Prob} = q(j|i)$
2. Accept that proposal. $\quad \text{Prob} = R(i \to j) = \min\{1, \frac{f(j)}{f(i)} \frac{q(i|j)}{q(j|i)}\}$

And $P(j \to i)$ can be found similarly [using $R(j \to i)$].

Now, there are two possible cases:

Case I: $f(j)q(i|j) > f(i)q(j|i)$, in which case
$R(i \to j) = 1 \quad$ and $\quad R(j \to i) = \frac{f(i)}{f(j)} \frac{q(j|i)}{q(i|j)}$

Case II: $f(j)q(i|j) \leq f(i)q(j|i)$, in which case
$R(i \to j) = \frac{f(j)}{f(i)} \frac{q(i|j)}{q(j|i)} \quad$ and $\quad R(j \to i) = 1.$

M-H algorithm satisfies detailed balance (cont'd):

So, $P(i \to j)$ and $P(j \to i)$ in those cases are:

Case I:

$$P(i \to j) = q(j|i) \qquad P(j \to i) = q(i|j) \times \frac{f(i)}{f(j)} \frac{q(j|i)}{q(i|j)}$$

Case II:

$$P(i \to j) = q(j|i) \times \frac{f(j)}{f(i)} \frac{q(i|j)}{q(j|i)} \qquad P(j \to i) = q(i|j)$$

Both of which, with a little rearranging, yield:

$$f(i)P(i \to j) = f(j)P(j \to i)$$

which is the detailed balance equation.

M-H algorithm example—beta distribution[4]:

Let θ be a variable with probability density function

$$f(\theta) \equiv \text{Beta}(74, 128) = \frac{\Gamma(74 + 128)}{\Gamma(74)\Gamma(128)}\theta^{73}(1-\theta)^{127}$$

Note that this is the posterior distribution of the frequency θ of allele A at a locus, given a uniform prior, and a sample of 100 diploids in which are found 73 copies of allele A.

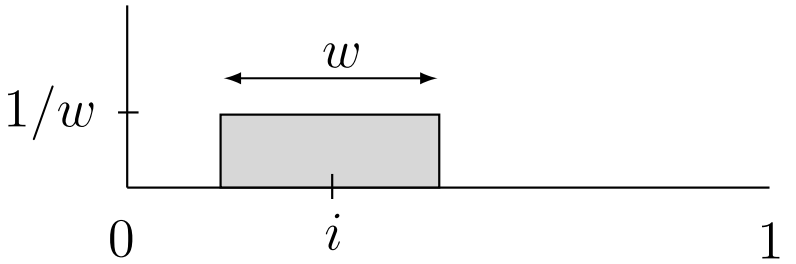
This distribution is known exactly, but for illustration, we will construct a Markov chain that has limiting distribution f, and sample from it.

[4]Up to this point we have been dealing with discrete Markov chains. The extension to continuous state spaces is straightforward. In such cases, π, P, q, and f may be taken as probability density functions, rather than as probability mass functions, and the general balance equation is expressed by a collection of integrals rather than in terms of matrix multiplication: $\int_x \pi(x)P(x \to y)dx = \pi(y)\ \forall y$

Step one: Choose a proposal distribution:

One is free to choose any proposal distribution q. The only property that it should satisfy is that if $q(i|j) > 0$, then $q(j|i)$ should also be positive[5]. Otherwise you end up wasting some time.

We will choose a uniform density for q with width w, centered on the current state i. It looks like:



Thus, $q(i|j) = 1/w$ for all i and j. (Note that if $j \geq 1$ or $j \leq 0$ then $f(j) = 0$ and we reject the proposal immediately.)

[5] And the proposal distributions used should fulfill irreducibility, *etc.*. Clearly, also, some proposal distributions will work better than others.

Step two: Compute the Hastings Ratio:

$$\frac{f(j)}{f(i)}\frac{q(i|j)}{q(j|i)} =$$

$$\frac{\Gamma(74+128)}{\Gamma(74)\Gamma(128)} \times \frac{\Gamma(74)\Gamma(128)}{\Gamma(74+128)} \times \frac{j^{73}(1-j)^{127}}{i^{73}(1-i)^{127}} \times \frac{1/w}{1/w}$$

$$= \left(\frac{j}{i}\right)^{73} \left(\frac{1-j}{1-i}\right)^{127}$$

Notice that the normalizing constant cancels out. (That *always* happens). And, in this case, q cancels out (*only because q is symmetrical*).

Computer Demo: `beta_sim`

Conclusions:

1. MCMC is just Monte Carlo with samples drawn from a Markov chain
2. The Markov chain in MCMC is constructed by concatenating moves together, each of which satisfies detailed balance
3. The rest of the module will explore methods for implementing MCMC in problems relevant to statistical genetics