# Markov Processes and MCMC

**Problem:** What is a Markov process and how is it specified? Derive the expression for the distribution of the $t^{th}$ state.

**Problem:** What is a limiting distribution? What is a stationary distribution and the conditions for existence and uniqueness? Prove that a symmetric process has the uniform distribution as a stationary distribution. What is time-reversibility?

**Question:** How can MCs be used for simulation? What is the Metropolis-Hastings Algorithm?

**Question:** How can MCs be used in a Bayesian data analysis procedure?

**Student Groups Problem**: Let say we have two randomly chosen groups of students who are undergoing the same summer math course - one group sees a positive subliminal message while the other sees a neutral message. We have a before and after test score for each student. Our general question is whether the positive subliminal message has an effect compared to the neutral one. Cast this into a full Bayesian analysis utilizing Python's PyMC module for MCMC.

## Contents


**1 Markov Chains** **2**

**2 Limiting and Stationary Distributions** **3**

**3 Metropolis-Hastings Method of Simulation** **4**

3.1 Gibbs Sampler Algorithm . . . . . 5

**4 MCMC In Bayesian Data Analysis** **5**

**5 Student Groups Problem** **5**

# 1 Markov Chains

A Markov Process (or chain) is a sequence of random variables, $\{X_1, X_2...\}$, conveniently thought of as a time-series of observations of a state wherein $X_t$ is the current state. You must specify a **state space S** = {possible states}, an initial distribution, $\pi_0$, which gives the probability for the system starting in each state and a transition rule which is a matrix specifying the probability to transition to each state $j$ given your current state $i$, $\mathbf{P}(i,j) = \mathrm{P}(X_{t+1} = j|X_t = i)$. The fundamental property of an MC is that the distribution of $X_t$ conditional on the entire preceeding chain is identical the the distribution conditional on only $X_{t-1}$:

$$\mathrm{P}(X_{t+1} = x_{t+1}|X_t = x_t, ....X_0 = x_0) = \mathrm{P}(X_{t+1} = x_{t+1}|X_t = x_t), \qquad (1)$$

or equivalently - the likelihood to transition to any state in the next time step is a function only of your current state. Using this fact the probability of any specific path can be calculated easily:

$$\begin{aligned}\mathrm{P}(X_2 = 9, X_1 = 5, X_0 = 7) &= \mathrm{P}(X_2 = 9|X_1 = 5)\mathrm{P}(X_1 = 5|X_0 = 7)\mathrm{P}(X_0 = 7) \\ &= \mathbf{P}(5,9)\mathbf{P}(7,5)\pi_0(7)\end{aligned}$$

An example is a **random walk** on the integers, where the state space is all integers and the transition rule is that $P(i, i+1) = p$ and $P(i, i-1) = 1-p$. **Gamblers ruin** is a similar chain except the state space is $\{0, 1, 2..., n\}$ and the transition rule for state 0 and state $n$ are that you are stuck there forever - we call these **absorbing states**. The **Ehrenfest chain** is a good model for diffusion - there are two urns holding balls and each time a ball is chosen at random and moved to the other urn from where it was. There is a probabilistic restoring force which keeps the number of balls about equal between the two urns.

If we denote the distribution of the RV $X_t$ by $\pi_t$ then we could use the Law of Total Probability and condition on the previous state to calculate the values $\pi_t(j)$:

$$\pi_t(j) = \sum_i \mathrm{P}(X_t = j|X_{t-1} = i)\mathrm{P}(X_{t-1} = 1) = \sum_i \mathbf{P}(i,j)\pi_{t-1}(i) \qquad (2)$$

$$\pi_t = \pi_{t-1}\mathbf{P} \qquad (3)$$

Here $\mathbf{P}$ is a matrix indexed by $i, j$. This relationship can be applied iteratively to get the distribution of $\pi_t$ starting from the initial distribution $\pi_0$ - which illustrates the proper interpretation of the powers of $\mathbf{P}$ as holding the conditional probabilities for $X_t$'s given $X_0$:

$$\pi_t = \pi_0\mathbf{P}^t \qquad (4)$$

$$\mathbf{P}^t(i,j) = \mathrm{P}(X_t = j|X_0 = i) \qquad (5)$$

## 2 Limiting and Stationary Distributions

First some definitions.

For some $\mathbf{P}$ it happens that with increasing $t$ the distribution of $X_t$ approaches some unique fixed distribution, $\pi_t \to \pi$ as $t \to \infty$, regardless of the starting distributions. We call $\pi$ the **limiting distribution** of the process. In these cases as $t \to \infty$ we see $\mathbf{P}^t$ approaches a form in which all the rows are identical (this is what gives independence from the initial distribution).

Another object of interest for a markov process $\mathbf{P}$ is any PMF that satisfies the eigenvector equation $\pi = \pi\mathbf{P}$. Any such pmf is called a **stationary distribution** of $\mathbf{P}$ where term stationary refers to the fact that if $\pi_0 = \pi_{stationary}$ then $\pi_t = \pi_{stationary}$ for all $t$. We can show that a limiting distribution must be stationary:

$$\pi_t \to \pi \text{ as } t \to \infty \tag{6}$$

$$\text{but also } \pi_t\mathbf{P} = \pi_{t+1} \to \pi \text{ so we must have} \tag{7}$$

$$\pi = \pi\mathbf{P}. \tag{8}$$

Note that a process $\mathbf{P}$ need not have a limiting or any stationary distributions, and a stationary distribution doesn't need to be a limiting distribution - but all limiting distributions are stationary. The eigenvector equation along with the normality constraint are sufficient to solve for any stationary distributions if they exist.

If you can get from every state $i$ from every other state $j$ in a finite series of moves then we say that $\mathbf{P}$ is **irreducible**.

**Fact 1.** *If an MC has a finite number of states then it has at least one stationary distribution.*

**Fact 2.** *If an MC is irreducible the chain has at most one stationary distribution.*

**Fact 3.** *If $\boldsymbol{P}$ is symmetric than the uniform distribution is stationary (symmetry implies columns sum to 1).*

From the eigenvector equation we can see that stationary distributions must satisfy $\pi(j) = \sum_i \pi(j)\mathbf{P}(i,j)$. If additionally a pmf $\mu$ on the state space satisfies $\mu(i)\mathbf{P}(i,j) = \mu(j)\mathbf{P}(j,i)$ for all states $i$ and $j$ then $\mu$ is certainly a stationary distribution of $\mathbf{P}$ but we also say that the process is **time reversible**. Time-reversibility means that given a chain generated with $\mu$ initial (stationary) distribution and matrix $\mathbf{P}$ you can't tell in which direction it was generated. This means that any sequence must be equally likely whether it is run forward or backward. If you imagine a two-element chain $\{F, H\}$ on the state space of letters there are two considerations for how believable the forward direction is:

how likely it is that F was the initial state, and how likely it is to transition from F to H. The believability of the reverse direction lies in how likely H is to be the initial state and how likely it is to transition from H to F. What the time-reversibility condition says is that if the starting state is more believable in one direction, then the transition is less believable in that direction - the two considerations perfectly cancel so that both directions are equally believable. You can imagine zooming in on any two elements $X_t$ and $X_{t+1}$ of a long chain and applying the same logic: since the initial distribution is stationary it still describes the believability that $X_t = x_t$ or $X_{t+1} = x_{t+1}$.

**Ergodic Theorem for Markov Chains 1.** *Let $\boldsymbol{P}$ be an irreducible process with a (unique) stationary distribution $\pi$ and with any starting state you like. We are guaranteed that for long times the fraction of the different states tends increasingly toward the stationary distribution.*

$$\frac{1}{n}\sum_{t=1}^{n} I(X_t = j) \to \pi(j) \text{ as } t \to \infty. \tag{9}$$

*Equivalently the average value of $f(X)$ approaches the expectation of $f$ over the pmf $\pi$.*

# 3 Metropolis-Hastings Method of Simulation

We might desire to generate random objects from a certain underlying distribution. For example, consider the sample space of 4x4 grids of non-negative numbers where each row and column has a specified sum - we would have trouble even writing down all the elements of this sample space. But consider doing a random walk through this sample space (now thought of as a state space). If you can find a transition rule (from one table to another) that is symmetric, then we know that the uniform distribution with be stationary.

In general we want to simulate a representative random sample from a distribution $\pi$ on a set $\mathbf{S}$. By the ergodic theorem we can do this by running a Markov chain on $\mathbf{S}$ for a sufficiently long time, provided we can find a $\mathbf{P}$ that is irreducible and has $\pi$ as its stationary distribution!

The **Metropolis-Hastings Method** is an algorithmic way of accomplishing the above.

**Metropolis-Hastings Algorithm.** *Given a pmf $f$ on state space $\boldsymbol{S}$ we want to find a transition rule $\mathbf{P}$ such that $f\mathbf{P} = f$ i.e. find a $\boldsymbol{P}$ for which $f$ is stationary.*

*First come up with any transition rule you like on $\boldsymbol{S}$ and call it $\mathbf{Q}(x, y)$.*

*If you are at state $X_t = x$ then draw a **candidate state** $y$ using the transition probabilities given by $\boldsymbol{Q}$.*

$$\textit{Consider the ratio } r = \frac{f(y)\mathbf{Q}(y,x)}{f(x)\mathbf{Q}(x,y)}$$
$$\textit{If } r \geq 1 \textit{ then let } X_{t+1} = y$$
$$\textit{If } r \leq 1 \textit{ then } \mathrm{P}(X_{t+1} = y) = r \textit{ and } \mathrm{P}(X_{t+1} = x) = 1 - r$$

You can see that this algorithm does satisfy our requirements if you consider the quantity $f(x)\mathrm{P}(x,y)$ which can be shown to be symmetric in $x$ and $y$. This means that $f(x)\mathrm{P}(x,y) = f(y)\mathrm{P}(y,x)$ which is the time reversibility condition and guarantees that $f$ is stationary.

### 3.1 Gibbs Sampler Algorithm

The first step of an MH transition rule is to propose a movement from point A to point B in parameter space and accept or reject the move with probability given by a specific formula (this formula ensures time-reversibility). The Gibbs Sampling algorithm is an implementation of MH which rather than randomly picking a proposed next point, instead uses a well-designed "proposal" method which results in a move that is always "accepted". Gibbs Sampling implementation of MH thus explores the parameter space more intelligently and efficiently.

## 4 MCMC In Bayesian Data Analysis

According to the Ergodic Theorem, if a process $P$ on state space $\{x_i\}$ is irreducible (you can get from $x_i$ to $x_j$ in a finite number of steps for all $i$ and $j$) and it has one unique stationary distribution $f(x)$ then the long-run fractions of state values in the chain approaches $f(x)$. If we want a sample from $f(x)$ we could find an irreducible $P$ for which it is stationary and run the chain for a long time then sample from that chain. **Metropolis-Hastings gives a formula for constructing such a $P$ which relies only on a non-normalized version of $f(x)$!** In a Bayesian analysis we always can write the posterior up to a normalization factor - it is just prior times likelihood. Now we can use metropolis-hastings to simulate a Markov Chain which behaves, in the long-run, like a sample of the posterior. The chain walks around in the space of unknown parameters following the MH algorithm, and after a sufficiently long time we construct a histogram of the values it has sampled as an approximate numerical representation of our posterior distribution.

## 5 Student Groups Problem

First we will recast our data as being the set of *changes* in test score, and we need to specify some model for the generating mechanism of this data. If we

think that the performance is dictated by some fixed group effect of the message plus an effect of intelligence then a reasonable model will be $X \sim N(\mu, \sigma)$ since intelligence is normally distributed in the population. So our state space is $\{\mu_1, \sigma_1, \mu_2, \sigma_2\}$ to capture the distributions of both groups. Now our question is mathematically well specified - how likely is it that $\mu_1 > \mu_2$ and by how much?

The above data-generating mechanism allows us to calculate the likelihood for any set of $X$ value, now we need priors on all four of the parameters. From a quick glance at the data lets say we think a typical change in test score, $X$ is around 10 points and that the changes seem somewhat well clustered so we belive the $\sigma$s are smallish. Beyond that it is hard to say what is a reasonable prior (Joe says normal on $\mu$'s and exponential on $\sigma$'s?). Whatever we choose as the four distributions on these parameters, the complete prior we need to use will be the joint prior density on the 4-vector (since they are independent parameters we can just multiply to get this).

Now finally we have a full mathematical form for (likelihood) x (joint prior). In PyMC we would need to read in the data as a vector of test score changes. We would need to define a stochastic type object which was the Data and calculated a complete log probability by summing the individual log-probabilities of each $X$ value. Here $X$s from the test group would use the $\mu_1$ distribution while the control group likelihoods come from the $\mu_2$ distribution.

In PyMC we would further specify either four new stochastic variables or one stochastic vector which defined the prior distributions on the parameters. Then we would be off to the races! For convenience we could specify a deterministic variable which calculated the difference between the two means.