Markov Chains and Applications

Alexander Volfovsky

August 17, 2007

Abstract

In this paper I provide a quick overview of Stochastic processes and then quickly delve into a discussion of Markov Chains. There is some assumed knowledge of basic calculus, probability, and matrix theory. I build up Markov Chain theory towards a limit theorem. I prove the Fundamental Theorem of Markov Chains relating the stationary distribution to the limiting distribution. I then employ this limiting theorem in a Markov Chain Monte Carlo example.[1]

Contents

[1] I would like to thank Prof. Lawler for all of his incites into stochastic processes during his course on random walks. Also, I would like to thank Sam Raskin for inspirational conversations when the paper seemed to be taking unfortunate turns for the worse. The book that is used in order to find statements of some of the theorems is the reference guide by Larry Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. Published by Springer Books in 2004.

1 Introduction

In a deterministic world, it is good to know that sometimes randomness can still occur. A stochastic process is the exact opposite of a deterministic one, and is a random process that can have multiple outcomes as time progresses. This means that if we know an initial condition for the process and the function by which it is defined, we can speak of likely outcomes of the process. One of the most commonly discussed stochastic processes is the Markov chain. Section 2 defines Markov chains and goes through their main properties as well as some interesting examples of the actions that can be performed with Markov chains. The conclusion of this section is the proof of a fundamental central limit theorem for Markov chains. We conclude the dicussion in this paper by drawing on an important aspect of Markov chains: the Markov chain Monte Carlo (MCMC) methods of integration. While we provide an overview of several commonly used algorithms that fall under the title of MCMC, Section 3 employs importance sampling in order to demonstrate the power of MCMC.

2 Markov Chains

Markov chains are stochastic processes that have the Markov Property, named after Russian mathematician Andrey Markov.

Definition of Markov Property Informally it is the condition that given a state, the past and future states are independent of it. Formally we can define it as follows:

$$\mathbb{P}\left(X_n = x | X_0, \ldots, X_{n-1}\right) = \mathbb{P}\left(X_n = x | X_{n-1}\right) \; \forall n \forall x.$$

We now present some important theorems and lemmas regarding Markov Chains.

2.1 Theorems and lemmas

First some notation: P always represents a transition matrix, while p_{ij} represents an element of it. X_j is always a random variable.

Definition State i is recurrent if

$$\mathbb{P}\left(X_n = i \, \text{for some} \, n \geq 1 | X_0 = i\right) = 1.$$

Otherwise it is transient.

Definition A chain is irreducible if every state can be reached from any other one. That is $p_{ij}\left(1\right) > 0 \, \forall i, j$

We state without proof the following results:

A state is recurrent if and only if $\sum_n p_{ii}(n) = \infty$ and it is transient if and only if $\sum_n p_{ii}(n) < \infty$.[2]

We delve right in with a lemma that connects the notions of reccurence and irreducibility.

Lemma 2.1 *In an irreducible chain, all the states are either transient or recurrent.*

Proof We take the shortest path from state i to state j (let it have n steps), and the shortest path from j to i (let it have m steps). Thus we have $p_{ij}(n) = a > 0$ and $p_{ji}(m) = b > 0$ and so we have

$$\begin{aligned} p_{ii}(l+n+m) &\geq p_{ij}(n)\, p_{jj}(l)\, p_{ji}(m) \\ &= abp_{jj}(l) \\ p_{jj}(l+n+m) &\geq p_{ji}(m)\, p_{ii}(l)\, p_{ij}(n) \\ &= abp_{ii}(l). \end{aligned}$$

So it is obvious that either $\sum_n p_{ii}(n)$ and $\sum_n p_{jj}(n)$ are both finite or are both infinite. Thus from the above results we note that all the states of an irreducible chain are either transient or recurrent, as desired.

Lemma 2.2 *Facts about reccurrence.*

1. *If state i is recurrent and $i \leftrightarrow j$, then j is reccurent.*
2. *If state i is transient and $i \leftrightarrow j$, then j is transient.*
3. *The states of a finite, irreducible Markov chain are all recurrent.*

Proof 1. We employ the definition of reccurrence. Thus $\exists n, m \geq 0$ st $p_{ij}(n) > 0$ and $p_{ji}(m) > 0$. Thus we have

$$\begin{aligned} \sum_{l=1}^{\infty} p_{jj}(l) &\geq \sum_{l=n+m+1}^{\infty} p_{jj}(l) \\ &\geq \sum_{k=1}^{\infty} p_{ji}(m)\, p_{ii}(k)\, p_{ij}(n) \\ &= p_{ji}(m) \left(\sum_{k=1}^{\infty} p_{ii}(k) \right) p_{ij}(n) \\ &= \infty \end{aligned}$$

where the last part follows from the recurrence of i.

[2] I hope that at this point the reader realizes a fundamental truth about transient states (as it becomes relevant soon). We have that given a nonzero probability p of returning to transient state i after starting from it, the distribution of the number of times that we return to that state is geometric. This requires only a little justification in one's mind.

2. We apply similar logic as above. Since $i \leftrightarrow j \, \exists n > 0$ st $p_{ij}(n) > 0$ so for $m > n$ we have $p_{ii}(m) \geq p_{ij}(n) p_{ji}(m-n)$ and thus we have:

$$\begin{aligned} \sum_{k=1}^{\infty} p_{ii}(k) &\geq \sum_{k=n+1}^{\infty} p_{ii}(k) \\ &\geq p_{ij}(n) \left(\sum_{k=n+1}^{\infty} p_{ji}(k-n) \right) \\ &= p_{ij}(n) \left(\sum_{l=1}^{\infty} p_{ji}(l) \right) \end{aligned}$$

which implies that $\sum_{l=1}^{\infty} p_{ji}(l) \leq \frac{1}{p_{ij}(n)} \left(\sum_{l=1}^{\infty} p_{ji}(l) \right) < \infty$ as desired.

3. We know from Lemma 2.1 that since the chain is irreducible, all the states are either recurrent or transient. First assume that all states are transient. Now, fix a state i and consider the number of times that we pass state j after starting at i. Since we only have finitely many states, the expectation for the number of times that we pass state j for some state j would be infinite. So this implies that the expected number of returns to state j after starting at state j would also be infinite. But that contradicts the geometric distribution of the number of returns, which has the expectation be at one over the probability of returning.

Definition Mean recurrence time for a recurrent state i is $m_i = \sum_n n f_{ii}(n)$ where $f_{ii}(n)$ is the probability of getting from i to i in exactly n steps. A state is null recurrent if $m_i = \infty$ and non-null otherwise.

We should note that we used exactly this value in the proof of (3) in lemma 2.1 so the above result can easily be extended to the fact that a finite state Markov chain has all its recurrent states be non-null.

Definition The period of state i, $d(i) = d$ if $p_{ii}(n) = 0$ for $d \nmid n$ and $d = \gcd \{n | p_{ii}(n) > 0\}$. Thus a state is periodic if $d(i) > 1$ and aperiodic otherwise.

Definition For $\pi_i = \lim_{n \to \infty} p_{ij}(n \times d(i))$, if greater than zero then non-null recurrent otherwise, null recurrent.

Lemma 2.3 *If $i \leftrightarrow j$ then $d(i) = d(j)$*

Proof We consider m, n st $p_{ij}(n) > 0$ and $p_{ji}(m) > 0$ thus we can note that from the Kolmogorov Chapman equations we have [3]

$$p_{ii}(m+n) = \sum_{k=1}^{N} p_{ik}(n) p_{ki}(m) \geq p_{ij}(n) p_{ji}(m).$$

[3] The proof of the Kolmogorov Chapman Equations is provided in the Appendix

Now by definition $p_{ii}(n+m) > 0$ and $d(i) \mid (n+m)$. Now we can consider $p_{ii}(m+l+n)$ and apply the same reasoning as above to arrive at:

$$\begin{aligned} p_{ii}(m+l+n) &= \sum_{k=1}^{N} p_{ir}(n) \sum_{t=1}^{N} p_{rk}(l) p_{ki}(m) \\ &\geq p_{ij}(n) p_{jj}(l) p_{ji}(m) \end{aligned}$$

So if we have that $p_{jj}(l) > 0$ then $d(j) \mid l$ implying as desired that $p_{ii}(m+l+n) > 0$ and so $d(i) \mid (n+m+l)$ but combining this with $d(i) \mid (m+n)$ we get that $d(i) \mid l$ and so since $d(j) = \gcd\{l | p_{jj}(l) > 0\}$ we get that $d(j) \geq d(i)$. We can apply the same logic going from j to i and so we arrive at the conclusion that $d(i) = d(j)$ as desired.

Definition A chain is ergodic if all of its states are non-null recurent and aperiodic.

Definition Let π be a probability mass function, then we say that π is a stationary probability distribution if $\pi = \pi P$ (that is, if $\pi$is an eighenvector of the transition probability matrix P)

Definition A limiting distribution exists if $P^n \rightarrow \begin{bmatrix} \pi \\ \vdots \\ \pi \end{bmatrix}$ for some π

Theorem 2.4 (Fundamental theorem for Markov Chains) *An irreducible, ergodic Markov chain has a unique stationary distribution π. The limiting distribution exists and is equal to π.*

Proof Since the chain is ergodic, it is non-null recurrent which implies from above that $\pi_j = \lim_{n\rightarrow\infty} p_{ij}(n) > 0 \forall i$ and $\sum \pi_j = 1$. Now we consident for any M $\sum_{i=0}^{M} p_{ij}(n) \leq \sum_{i=0}^{\infty} p_{ij}(n) = 1$. Now, letting $n \rightarrow \infty$ we get that $\sum_{i=0}^{M} \pi_i \leq 1 \forall M$ which implies from above that the same is true for the inifinte case $\sum_{i=0}^{\infty} \pi_i \leq 1$. Now we consider the probability of moving from i to j in $n+1$ steps so $p_{ij}(n+1) = \sum_{k=0}^{\infty} p_{ik}(n) p_{kj} \geq \sum_{k=0}^{M} p_{ik}(n) p_{kj} \forall M$. Now we again can let $n \rightarrow \infty$ which implies that $\pi_i \geq \sum_{k=0}^{M} \pi_k p_{kj} \forall M$ which implies that $\pi_i \geq \sum_{k=0}^{\infty} \pi_k p_{kj}$. Now we assume the inequality is strict for some i which leads to the following contradiction:

$$\begin{aligned} \sum_{i=0}^{\infty} \pi_i &\geq \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \pi_k p_{ki} \\ &= \sum_{k=0}^{\infty} \pi_k \sum_{i=0}^{\infty} p_{ki} \\ &= \sum_{k=0}^{\infty} \pi_k. \end{aligned}$$

Thus we come to the conclusion that $\pi_i = \sum_{k=0}^{\infty} \pi_k p_{ki} \, \forall i$. Now we can consider $\tilde{\pi}_i = \pi_i / \sum_{k=0}^{\infty} \pi_k$ to be a stationary distribution. So we have shown existence. Now to show uniqueness we consider the following:

$$\tilde{\pi}_i = \mathbb{P}(X_n = i) = \sum_{j=0}^{\infty} \mathbb{P}(X_n = i | X_0 = j) \, \mathbb{P}(X_0 = j) = \sum_{j=0}^{\infty} p_{ji}(n) \, \tilde{\pi}_j.$$

So we have that $\tilde{\pi}_i \geq \sum_{j=0}^{M} p_{ji}(n) \, \tilde{\pi}_j$and taking $M, n \to \infty$ we get that $\tilde{\pi}_i \geq \sum_{j=0}^{\infty} \tilde{\pi}_j \pi_i = \pi_i$ but we know that from the transition matrix that $p_{ji}(n) \leq 1$ and so $\tilde{\pi}_i \leq \sum_{j=0}^{M} p_{ji}(n) \, \tilde{\pi}_j + \sum_{j=M+1}^{\infty} \tilde{\pi}_j \, \forall M$ and so taking $n \to \infty$we get $\tilde{\pi}_i \leq \sum_{j=0}^{M} \pi_i \tilde{\pi}_j + \sum_{j=M+1}^{\infty} \tilde{\pi}_j \, \forall M$. Now we know that $\tilde{\pi}$is a stationary distribution so it sums up to 1, and so we let $M \to \infty$ and we get $\tilde{\pi}_i \leq \sum_{j=0}^{\infty} \pi_i \tilde{\pi}_j = \pi_i$ which implies that the stationary distribution is unique.

The above process thus shows the existence of a limiting distribution and so we now know that an ergodic chain converges to its stationary distribution.

This proof allows us to take any bounded function g and say with probability 1 that

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} g(X_n) \to E_{\pi}(g) \equiv \sum_{j} g(j) \, \pi_j$$

which is a very strong result that we use constantly in Markov chain Monte Carlo.

We conclude this section of lemmas and theorems with a useful definition and a small lemma.

Definition π satisfies detailed balance if $\pi_i p_{ij} = p_{ji} \pi_j$

Lemma 2.5 *If π satisfies detailed balance then it is a stationary distribution.*

Proof We consider the j^{th} element of πP which is $\sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j$ as desired.

2.2 Applications

In this section I provide a few basic examples of Markov chains to illustrate the points made above. All the examples come from Chapter 23 of Larry Wasserman's book (provided there as exercises rather than as solved examples).

Example Consider a two-state Markov chain with state $\chi = \{1, 2\}$ and transition matrix

$$P = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$

where $0 < a, b < 1$. Prove that

$$\lim_{n \to \infty} P^n = \begin{bmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{bmatrix}.$$

Solution To verify this we need to first show that the chain is irreducible and ergodic. Irreducible is easy since all states communicate with nonzero probabilities. To show that it is ergodic we need to show that all states are reccurent, non-null and aperiodic. Since $p_{ii}(1) > 0$ we have that this chain is aperiodic. Now at time n we must be at some location so $\sum_{j \in S} p_{ij}(n) = 1$ for any i. Since this is true for every n we can take the limit as $n \to \infty$ to get $\lim_{n\to\infty} \sum_{j \in S} p_{ij}(n) = 1$ but in our case S is finite so we can move the limit under the sum. Now, if every state in our MC is transient or null reccurent we would have that $\lim_{n\to\infty} p_{ij}(n) = 0$ thus contradicting the above statement, so at least one state must be positive reccurent. Now, since we are in a finite state MC we have that all states are positive reccurent (due to question (7) below) since all the states communicate. Thus it is clear that our state is irreducible and ergodic and so it has a unique stationary distribution. Now we can solve:

$$\begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix} \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix} = \begin{bmatrix} \pi_1(1-a) + \pi_2 b & \pi_1 a + \pi_2(1-b) \end{bmatrix}$$

and now we can solve this system of equations with the added restriction $\pi_1 + \pi_2 = 1$. So we get

$$\begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix} = \begin{bmatrix} \frac{b}{a+b} & \frac{a}{a+b} \end{bmatrix}$$

as desired. And now we can easily note that the limiting distribution is $\lim_n P^n = \begin{bmatrix} \pi \\ \pi \end{bmatrix}$ as desired.

Example Let

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Show that $\pi = (0.5, 0.5)$is a stationary distribution. Does this chain converge?

Solution All we are asked to show is that $\pi = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$ is a stationary distribution so:

$$\begin{aligned} \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} &= \begin{bmatrix} 0+0.5 & 0.5+0 \end{bmatrix} \\ &= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \end{aligned}$$

as desired. However, the chain does not converge as it is clear that if we start with π we will have an equal probability of being in either state. This means that the period of each of the states is 2 and so the chain does not converge.

3 Markov Chain Monte Carlo

Markov chain Monte Carlo integration is a method for integrating function that might not have a closed form using estimation.

We demonstrate a very basic example of MCMC processes through Importance Sampling. Importance sampling is used in statistics as a variance reduction method. While the standard method that we describe below does not

necessarily optimize the variance, we will state the condition for minimal variance. Importance sampling allows us to estimate the distribution of a random variable using a different random variable. The idea behind the process is that during the simulation, due to weighing of the random variable from which we have the observations, we get a better, less biased idea of the parameter we are estimating. Thus the choice of the weight is very important. We will not go through the derivation of the "best" (in terms of minimizing the variance) weight, but just state the result here.

In an importance sampling problem we are trying to estimate the distribution of I using $\hat{I} = \frac{1}{N}\sum \frac{h(X_i)f(X_i)}{g(X_i)}$.The optimal choise of g in this case is $g(x) = \frac{|h(x)|f(x)}{\int |h(s)|f(s)ds}$.

The problem is from Larry Wasserman's All of Statistics: A Concise Course in Statistical Inference.

3.1 Statement

(From Larry Wasserman's All of Statistics, 24.7.2)

Let $f_{X,Y}(x,y)$be a bivariate density and let $(X_1,Y_1),\ldots,(X_N,Y_N) \sim f_{X,Y}$.

1. Let $w(x)$ be an arbitrary probability density function. Let

$$\hat{f}_X(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{f_{X,Y}(x,Y_i)\,w(x)}{f_{X,Y}(X_i,Y_i)}.$$

   Show that, for each x, $\hat{f}_X(x) \xrightarrow{p} f_X(x)$. Find an expression for the variance.

2. Let $Y \sim N(0,1)$ and $X|Y=y \sim N\left(y, 1+y^2\right)$. Use the method in (1) to estimate $f_X(x)$.

3.2 Solution

1. We consider each part of the sum to be its own random variable, and we note that they are all identically and independently distributed. Due to this we can consider just one of them for the following:

$$\begin{aligned} E\left[\frac{f_{X,Y}(x,Y_i)\,w(x)}{f_{X,Y}(X_i,Y_i)}\right] &= \int \frac{f_{X,Y}(x,y)\,w(z)}{f_{X,Y}(z,y)} f_{X,Y}(z,y)\,dzdy \\ &= \int f_{X,Y}(x,y)\,w(z)\,dzdy \\ &= \int f_{X,Y}(x,y)\,dy \\ &= f_X(x) \end{aligned}$$

and so we can apply the law of large numbers to note that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{f_{X,Y}\left(x,Y_i\right)w\left(x\right)}{f_{X,Y}\left(X_i,Y_i\right)} \quad \xrightarrow{p} \quad E\left[\frac{f_{X,Y}\left(x,Y_i\right)w\left(x\right)}{f_{X,Y}\left(X_i,Y_i\right)}\right]$$

which is the same as $\hat{f}_X\left(x\right) \xrightarrow{p} f_X\left(x\right)$, as desired.

The variance calculation is fairly simple and we do not dwell on it, providing simply the easily verifiable answer:

$$\text{var}\hat{f}_X\left(x\right) = \frac{1}{N}\left[\int\frac{f_{X,Y}^2\left(x,y\right)w^2\left(z\right)}{f_{X,Y}\left(z,y\right)}dzdy - f_X^2\left(x\right)\right].$$

2. We note that the marginal density of X is hard to evaluate:

$$\begin{aligned} f_X\left(x\right) &= \int f_{X|Y}\left(x\right)f_Y\left(y\right)dy \\ &= \int\frac{1}{\sqrt{2\pi\left(1+y^2\right)}}e^{-\frac{(x-y)^2}{2\left(1+y^2\right)}}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy. \end{aligned}$$

Thus it makes sense to employ importance sampling as in (1) in order to estimate $f_X\left(x\right)$. So we take the distribution of $w\left(x\right)$ to be the Standard Normal as it seems like a reasonable one in this case. So we have:

$$\begin{aligned} \hat{f}_X\left(x\right) &= \frac{1}{N}\sum_{i=1}^{N}\frac{f_{X,Y}\left(x,Y_i\right)w\left(x\right)}{f_{X,Y}\left(X_i,Y_i\right)} \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{f_{X|Y}\left(x|Y_i\right)w\left(X_i\right)}{f_{X|Y}\left(X|Y_i\right)} \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2\left(1+y^2\right)}\left[\left(x-Y_i\right)^2-\left(X_i-Y_i\right)^2+\left(x\left(1+y^2\right)\right)^2\right]\right\} \end{aligned}$$

using the above knowledge.

3.3 Further Discussion

We see above a very basic application of Markov chain Monte Carlo methods which allows us to use a biased sampling distribution in order to estimate a random variable of interest. The above discussed method is a very basic and introductory one. We actually have multiple possible algorithms that we can apply in order to arrive at the best possible estimate, however they are a topic for another paper and will only be briefly mentioned here.

The most commonly used algorithms for MCMC are the Metropolis Hastings algorithms which use a conditional proposal distribution in order to construct a Markov chain with a stationary distribution f. It supposes that X_0 was chosen arbitrarily and then proceeds to use the proposal distribution in order to

generate candidates that are either added to the chain or are overlooked based on a specified probability distribution.[4] There are several different incarnations of this algorithm, with different suggested proposal distributions: the random-walk M-H algorithm is the one that was described above (as if we do not accept or reject the generated value, all we are doing is simulating a random walk on the real line). In independence M-H we change the proposal distribution to a fixed distribution which we believe to be an approximation of f.

Another method that gets a lot of use is the Gibbs sampling algorithm which is simply an emblishment of the M-H algorithm. What this method does is take a multi-dimensional problem and turns it into several one-dimensional problems that can be easily estimated using the above method. So instead of simply getting X_{i+1}, Gibbs sampling allows for the estimation of $\left(X_{i+1}^{(1)}, \ldots, X_{i+1}^{(n)}\right)$for an n-dimensional model.

This process works due to detailed balance of Markov Chains, which was briefly mentioned in the previos section. For further information consult Wasserman's book.

[4] For basic Metropolis-Hastings, we have the probability be $r(x,y) = \min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}$where q is the proposal distribution. In the case that we have, we can easily choose a q such that $q(x|y) = q(y|x)$ so obviously one part of the multiplication cancels out.

4 Appendix

Theorem 4.1 *Chapman Kolmogorov Equations. The n-step probabilities satisfy*

$$p_{ij}(n+m) = \sum_k p_{ik}(m)\, p_{kj}(n).$$

(We thus get $P_{m+n} = P_m P_n$ which is standard matrix multiplication)

Proof We employ the law of conditional probability and the law of total probability to arrive at the following:

$$\begin{aligned}
p_{ij}(m+n) &= P(X_{m+n} = j | X_0 = i) \\
&= \sum_k P(X_{m+n} = j, X_m = k | X_0 = i) \\
&= \sum_k P(X_{m+n} = j | X_m = k, X_0 = i)\, P(X_m = k | X_0 = i) \\
&= \sum_k P(X_{m+n} = j | X_m = k)\, P(X_m = k | X_0 = i) \\
&= \sum_k p_{ik}(m)\, p_{kj}(n)
\end{aligned}$$

by definition, as desired.