# 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 MCMC example .

## Contents



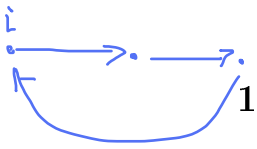
Ni := sum I{Xn=i} the
number of state i in the chain

Ti=Ti(1): hitting time from i to i
Ti(k): k-th hitting time
Pi := P(.|x0=i)

---

Lemma.
condi. mf: P(Ni=k|X0=i) = P(Ti<∞|X0=i) ^ {k}P(Ti=∞|X0=i)
condi. sf: P(Ni≥k|X0=i) = P(Ti<∞|X0=i) ^ k

Fact.
Ni≥k = Ti(k)<∞
Ni=k = Ti(k)<∞ and Ti(k+1)=∞
in finite cases, exists i, Ni=∞




# 1 Introduction

E(Ni) = ∑pii(n) = P(Ti<∞|X0=i)/P(Ti=∞|X0=i)
E(Ni) = ∞ iff Ni=∞ a.s. iff Ti<∞ a.s.

Pi(Ti(k+1)=Ti(k)+n|Ti(k)<∞)
=Pi(Ti(k)=n)

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 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.

**Definition of Markov Property** Informally it is the condition that given a state, the past and future states are independent of it. Formally

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

= P(1)=P

## 2.1 Theorems and lemmas

pij(n) := P(Xn=j|X0=i)
P(n) = P ^ n, homeogeneous

*notation*: $P$ represents a transition matrix, $p_{ij}$ represents an element of it.

**Definition** State $i$ is recurrent if

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

Otherwise it is transient.

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

**Fact.**
A state is recurrent iff $\sum_n p_{ii}(n) = \infty$; it is transient iff $\sum_n p_{ii}(n) < \infty$.

∑n ∑ pji(n)=∞, for all j-->i

**Lemma 2.2**

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

j--> i def. pji(n)>0 for some n

Lemma.
if j--> i, then pii >> pij; pji >> pjj >> pij
if i <--> j, then pii(l) ~ pjj(l)

Fact. i: recurrent iff
Xn = i, for some n, i-a.s.;
Pi(∪n{xn=i})=1;
Ti<∞, i-a.s.
Ni = ∞, i-a.s. (i: Harris rec.)

=E Ti

**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. Pi(Ti = n) = fii(n)

the distr. of hitting time Ti

non-int. positive / int.

*fact* 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.

exists n, n+1, pii(n) > 0

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

**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

$$p_{ii}(m+n) \geq p_{ij}(n)\, p_{ji}(m).$$

Now by definition $p_{ii}(n+m) > 0$ and $d(i) \,|\, (n+m)$.

$$p_{ii}(m+l+n) \geq p_{ij}(n)\, p_{jj}(l)\, p_{ji}(m)$$

So if we have that $p_{jj}(l) > 0$ then $d(j) \,|\, l$ implying as desired that $p_{ii}(m+l+n) > 0$ and so $d(i) \,|\, (n+m+l)$ but combining this with $d(i) \,|\, (m+n)$ we get that $d(i) \,|\, l$ and so since $d(j) = \gcd\{l | p_{jj}(l) > 0\}$ we get that $d(j) \geq d(i)$. /

==> $d(i) = d(j)$

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

**Definition** Let $\pi$ be a pmf, then we say that $\pi$ is a stationary/invariant (probability) distribution if $\pi = \pi P$ ($\pi$ is an eighenvector of the $P$)

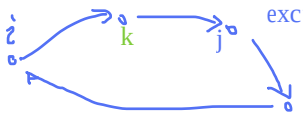
**Definition** A limiting distribution exists if $P^n \to \begin{bmatrix} \pi \\ \vdots \\ \pi \end{bmatrix}$ for some $\pi$

non null rec. iff exists inv. proba. π

Theorem. Irreducible & Recurrent ==> E! unique stationary mea.

$\pi j \sim E\,\#\{xt=j,\ 0<t\le Ti\}$

$\pi i = 1$




**Theorem 2.4 (Fundamental theorem for Markov Chains)** *An irreducible, non null rec. ergodic Markov chain has a unique stationary distribution $\pi$= The limiting distribution.*

**Proof** Since the chain is ergodic, it is non-null recurrent which implies from above that $\pi_j = \lim_{n\to\infty} p_{ij}(n) > 0 \ \forall i$ and $\sum \pi_j = 1$.

by Fatou lemma

==> $\pi_i \ge \sum_k \pi_k p_{ki}$.

Now we assume the inequality is strict for some $i$ which leads to the following contradiction:

$$\sum_{i=0}^{\infty} \pi_i \quad \geq \quad \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} \pi_k p_{ki} \quad = \quad \sum_{k=0}^{\infty} \pi_k.$$

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.

Now to show uniqueness :

by Fatou lemma lim_n

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

Now we know that $\tilde{\pi}$ is a stationary distribution so it sums up to 1, we get $\tilde{\pi}_i \le \sum_j \pi_i \tilde{\pi}_j = \pi_i$

==> that the stationary distribution is unique.

*the existence of a limiting distribution and so we now know that an ergodic chain converges to its stationary distribution.*

take any bounded function $g$ and say with probability 1 that

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

**Definition** $\pi$ satisfies detailed balance if $\pi_i p_{ij} = p_{ji} \pi_j$ DBC

**Lemma 2.5** *If $\pi$ satisfies detailed balance then it is a stationary distribution.*

# 3 Markov Chain Monte Carlo

McMC integration is a method for integrating function that might not have a c losed 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. Importancesamplingallowsustoestimatethedistributionofarvusingadifferent rv. Theideabehindtheprocessisthatduringthesimulation,duetoweighingofthe rvfromwhichwehavetheobservations,wegetabetter,lessbiasedideaoftheparameterweareestimating. Thusthechoiceoftheweightisveryimportant. Wewillnot gothroughthederivationofthe"best"(interms of minimizingthevariance)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 $g(x) = \frac{|h(x)|f(x)}{\int |h(s)|f(s)ds}$.

## 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 pdf. 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 rv, and we note that they are all iid . Due to this we can consider just one of them for the following:

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

and so we can apply the law of large numbers

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

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:

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

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:

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

## 3.3 Further Discussion

We see above a very basic application of McMC methods which allows us to use a biased sampling distribution in order to estimate a rv 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.

[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.