# Lecture 9: Monte Carlo Simulations and Sampling 

Instructor: Yen-Chi Chen

### 9.1 Introduction

Monte Carlo simulation is a common approach to numerically approximate a quantity of interest. It uses random evaluations to approximate a deterministic result.

Monte Carlo simulation may involves a sampling procedure, known as the Monte Carlo sampling. Monte Carlo sampling concerns the problem of sampling from a distribution $Q$ or a density $q$ with only limited amount of knowledge about $Q$ or $q$. When $Q$ is Gaussian or some parametric family that is well-known, sampling from $Q$ is an easy task. However, when $Q$ does not have a simple form or we may not even be able to exactly compute $Q$ or $q$, sampling from this distribution is challenging. For instance, suppose I want to sample from a PDF $q(x)$ but all I know is that $q(x) \propto e^{-x^{4}}$. How can I sample from $q$ ?

Depending on the final goal of the analysis, our sampling strategy may be different. There are two common problems that Monte Carlo sampling will be used:

- Evaluating the mean of a function. We want to compute

$$
\mathbb{E}(f(Z)), \quad Z \sim Q
$$

where $f(x)$ is a function we can evaluate at every $x$ and $Q$ is a distribution of interest. In this case, the famous importance sampling technique can be applied when we can only evaluate the PDF $q(x)$ at each $x$ but not able to sample from it. In the case that we can only evaluate $r(x) \propto q(x)$, the self-normalizing importance sampling can be applied.

- Sampling from a distribution. Sampling from a distribution is a more general problem that we want to generate points from $Q$ or $q$. When $q(x)$ can be evaluate at every point $x$, a famous approach is the rejection sampling. However, in many statistical problems (in particular, Bayesian inference) where we only have access to evaluating $r(x) \approx q(x)$, the rejection sampling is not applicable. An alternative approach is the MCMC (Markov Chain Monte Carlo), which only requires being able to evaluate $r(x)$, not $q(x)$.

Here is an interesting fact. The problem of Monte Carlo sampling and the problem of density estimation are like inverse problems to each other. In Monte Carlo sampling, we often are able to evaluate the PDF $q(x)$ or its surrogate $r(x)$ but we do not know how to sample from it. In density estimation, we obtain a sample from an unknown density function and the goal is to approximate and establish a way to evaluate the PDF. If we are able to generate many points from a distribution (solved the problem of Monte Carlo sampling), even if we do not know how to evaluate the PDF, we can approximate the underlying PDF at every $x$ using density estimator.

### 9.2 Monte Carlo approximations to a mean

### 9.2.1 Importance sampling

Let $X$ be a random variable with PDF $p$. Consider evaluating the following quantity:

$$
I=\mathbb{E}(f(X))=\int f(x) p(x) d x
$$

where $f$ is a known function.
One famous approach to approximate $I$ is the importance sampling. We first pick a proposal density (also called sampling density) $q$ and generate random numbers $Y_{1}, \cdots, Y_{N}$ IID from $q$. Then the importance sampling estimator is

$$
\widehat{I}_{N}=\frac{1}{N} \sum_{i=1}^{N} f\left(Y_{i}\right) \cdot \frac{p\left(Y_{i}\right)}{q\left(Y_{i}\right)}
$$

When $p=q$, this reduces to the simple estimator that uses sample means of $f\left(Y_{i}\right)$ to estimate its expectation.

Theorem 9.1 The importance sampling is unbiased and has a variance

$$
\operatorname{Var}\left(\widehat{I}_{N}\right)=\frac{1}{N}\left(\int \frac{f^{2}(y) p^{2}(y)}{q(y)} d y-I^{2}\right)
$$

## Proof:

## Bias.

$$
\begin{aligned}
\mathbb{E}\left(\widehat{I}_{N}\right)-I & =\mathbb{E}\left(f\left(Y_{i}\right) \cdot \frac{p\left(Y_{i}\right)}{q\left(Y_{i}\right)}\right)-I \\
& =\int f(y) \frac{p(y)}{q(y)} q(y) d y-I \\
& =\int f(y) p(y) d y-I=0
\end{aligned}
$$

Thus, it is an unbiased estimator.

## Variance.

$$
\begin{aligned}
\operatorname{Var}\left(\widehat{I}_{N}\right) & =\frac{1}{N} \operatorname{Var}\left(f\left(Y_{i}\right) \cdot \frac{p\left(Y_{i}\right)}{q\left(Y_{i}\right)}\right) \\
& =\frac{1}{N}\{\mathbb{E}\left(f^{2}\left(Y_{i}\right) \cdot \frac{p^{2}\left(Y_{i}\right)}{q^{2}\left(Y_{i}\right)}\right)-\underbrace{\mathbb{E}^{2}\left(f\left(Y_{i}\right) \cdot \frac{p\left(Y_{i}\right)}{q\left(Y_{i}\right)}\right)}_{I^{2}}\} \\
& =\frac{1}{N}\left(\int \frac{f^{2}(y) p^{2}(y)}{q(y)} d y-I^{2}\right)
\end{aligned}
$$

So only the first quantity depends on the choice of proposal density $q$. Thus, if we have multiple proposal density, say $q_{1}, q_{2}, q_{3}$, the best proposal will be the one that minimizes the integration $\int \frac{f^{2}(y) p^{2}(y)}{q(y)} d y$.

You may be curious about the optimal proposal density (the $q$ that minimizes the variance). And here is a striking result about this optimal proposal density.

Theorem 9.2 The importance sampling has minimal variance when

$$
q_{\mathrm{opt}}(y) \propto f(y) p(y) \Longrightarrow p_{\mathrm{opt}}(y)=\frac{f(y) p(y)}{\int f(y) p(y) d y}
$$

and the optimal variance is 0 .

## Proof:

First, we recall the Cauchy-Scharwz inequality-for any two functions $A(y)$ and $B(y)$,

$$
\int A^{2}(y) d y \int B^{2}(y) d y \geq\left(\int A(y) B(y) d y\right)^{2}
$$

and the $=$ holds whenever $A(y) \propto \cdot B(y)$ for some constant. One way to think about this is to view them as vectors-for any two vectors $u, v,\|u\|^{2}\|v\|^{2} \geq\|u \cdot v\|^{2}$ and the equality holds whenever $u$ and $v$ are parallel to each other. Identifying $A^{2}(y)=\frac{f^{2}(y) p^{2}(y)}{q(y)}$ and $B^{2}(y)=q(y)$, we have

$$
\int \frac{f^{2}(y) p^{2}(y)}{q(y)} d y \underbrace{\int q(y) d y}_{=1} \geq\left(\int \frac{f^{2}(y) p^{2}(y)}{q(y)} q(y) d y\right)^{2}=I^{2}
$$

Namely, this tells us that the optimal choice $q_{\text {opt }}(y)$ leads to

$$
\operatorname{Var}\left(\widehat{I}_{N, \mathrm{opt}}\right)=\frac{1}{N}\left(I^{2}-I^{2}\right)=0
$$

a zero-variance estimator! Moreover, the optimal $q$ satisfies

$$
\sqrt{\frac{f^{2}(y) p^{2}(y)}{q_{\mathrm{opt}}(y)}}=A(y) \propto B(y)=\sqrt{q_{\mathrm{opt}}(y)}
$$

implying

$$
\begin{equation*}
q_{\mathrm{opt}}(y) \propto f(y) p(y) \Longrightarrow p_{\mathrm{opt}}(y)=\frac{f(y) p(y)}{\int f(y) p(y) d y} \tag{9.1}
\end{equation*}
$$

This gives us a good news-the optimal proposal density has 0 variance and it is unbiased. Thus, we only need to sample it once and we can obtain the actual value of $I$. However, even if we know the closed form of $q_{\text {opt }}(y)$, how to sample from this density is still unclear. In the next section, we will talk about a method called Rejection Sampling, which is an approach that can tackle this problem.

Self-normalized importance sampling. Let $X, Y$ be two random variables and we are interested in evaluating the conditional expectation

$$
\mu(x)=\mathbb{E}(f(Y) \mid X=x)=\int f(y) p(y \mid x) d y
$$

for some function $f(x)$. Suppose that we do not know the conditional density but only have access to evaluate the joint density $p(x, y)$ up to some constant. Namely, for every $(x, y)$ we know the number $r(x, y)$ where
$r(x, y) \propto p(x, y)$. This scenario occurs in Bayesian inference when the joint PDF $p(x, y)$ are the posterior distribution of two parameters and there is no clear form of the posterior distribution- we only have access to the likelihood function and the prior density value; the function $r(x, y)$ is the product between the likelihood function and prior density value.

We can approximate $\mu(x)$ using a modification of importance sampling called the self-normalized importance sampling. Note that $\mu(x)$ can be written as

$$
\mu(x)=\int f(y) p(y \mid x) d y=\frac{\int f(y) p(x, y) d y}{\int p(x, y) d y}=\frac{\int f(y) r(x, y) d y}{\int r(x, y) d y}=\frac{\int f(y) g_{x}(y) d y}{\int g_{x}(y) d y}
$$

where $g_{x}(y)=r(x, y)$ is something we can evaluate at every $y$. This motivates us to use importance sampling for both numerator and denominator. Suppose that we generate $Z_{1}, \cdots, Z_{N} \sim q(z)$ such that the the support of $q$ covers the support of $g_{x}(y)$. Using $\phi(y)=\frac{r(x, y)}{q(y)}$, an approximation to $\mu(x)$ is

$$
\widehat{\mu}_{N}(x)=\frac{\frac{1}{N} \sum_{i=1}^{N} f\left(Z_{i}\right) \frac{r\left(x, Z_{i}\right)}{q\left(Z_{i}\right)}}{\frac{1}{N} \sum_{j=1}^{N} \frac{r\left(x, Z_{j}\right)}{q\left(Z_{j}\right)}}=\frac{\sum_{i=1}^{N} f\left(Z_{i}\right) \phi\left(Z_{i}\right)}{\sum_{j=1}^{N} \phi\left(Z_{j}\right)}=\sum_{i=1}^{N} \omega\left(Z_{i}\right) f\left(Z_{i}\right),
$$

where $\omega\left(Z_{i}\right)=\frac{\phi\left(Z_{i}\right)}{\sum_{j=1}^{N} \phi\left(Z_{i}\right)}$ is the weight on $Z_{i}$-this quantity is generated by a self-normalizing process and it is why the procedure is called the self-normalizing importance sampling. Because of the denominator is also random in the weight, the estimator is not necessarily unbiased (but is asymptotically unbiased):

$$
\mathbb{E}\left(\widehat{\mu}_{N}(x)\right)-\mu(x)=O\left(N^{-1}\right), \quad \operatorname{Var}\left(\widehat{\mu}_{N}(x)\right)=O\left(N^{-1}\right)
$$

Stochastic simulation model. The stochastic simulation model occurs from a problem in the industrial engineering problem. Suppose that we want to run a simulation to estimate the reliability of a nuclear power plant under natural environments (distribution of wind speed, temperature, ...etc). The simulation model requires a set of configurations $X$ of the power plant and then it will generate a reliability index $V$, which is a random quantity whose distribution depends on $X$. We are interested in estimating the quantity

$$
\mathcal{E}=\mathbb{E}(\theta(V) ; X \sim q)
$$

where $\theta(\cdot)$ is a known function and $q$ is a PDF we can sample from. You can think of $\theta(V)$ as some transformed quantity of the reliability index such as $\theta(V)=I\left(V \geq v_{0}\right)$ that we are interested in the chance that the reliability index is greater than a threshold $v_{0}$ (so that we can claim that this nuclear power plant is safe). The PDF $q$ should be viewed as how the configurations will be under the 'natural condition'-which is often generated by a climate model. Note that the parameter of interest can be written as

$$
\mathcal{E}=\mathbb{E}(\theta(V) ; X \sim q)=\int \theta(v) p(v \mid x) q(x) d x
$$

where $p(v \mid x)$ is the conditional density of $V$ given $X=x$. We can estimate $\mathcal{E}$ easily using the importance sampling. We generate $X_{1}, \cdots, X_{n} \sim w$ such that $w$ is not necessarily the same as $q$ and perform simulations to obtain $V_{1}, \cdots, V_{n}$ (note that generating $V$ could be very expansive since it involves running a complex simulation model). Then the final estimate of $\mathcal{E}$ is

$$
\widehat{\mathcal{E}}_{n}=\frac{1}{n} \sum_{i=1}^{n} \frac{q\left(X_{i}\right)}{w\left(X_{i}\right)} \theta\left(V_{i}\right)
$$

It is easy to show that this is an unbiased estimator of $\mathcal{E}$ so one may be wondering which choice of $w$ that has the smallest variance. Interestingly, you can show that the following $w$ has the minimal variance:

$$
w^{*}(x) \propto \sqrt{\mathbb{E}\left(\theta(V)^{2} \mid X=x\right)} \cdot q(x) .
$$

This result is from

- Choe, Y., Byon, E., \& Chen, N. (2015). Importance sampling for reliability evaluation with stochastic simulation models. Technometrics, 57(3), 351-361.
- Chen, Y. C., \& Choe, Y. (2019). Importance sampling and its optimality for stochastic simulation models. Electronic Journal of Statistics, 13(2), 3386-3423.

Although we know the form of $w^{*}(x)$, it depends on unknown quantity $\mathbb{E}\left(\theta(V)^{2} \mid X=x\right)$ so we cannot directly apply it. In Chen\& Choe (2019), the authors proposed to use a two-stage procedure that we split the data into two parts: we use the first part to estimate $\mathbb{E}\left(\theta(V)^{2} \mid X=x\right)$ and sample the second part from the estimate optimal $w^{*}(x)$.

### 9.3 Monte Carlo sampling

### 9.3.1 Rejection sampling

Given a density function $f(x)$, the rejection sampling is a method that can generate data points from this density function $f$.

Here is how one can generate a random variable from $f$.

1. We first choose a number $M \geq \sup _{x} \frac{f(x)}{p(x)}$ and a proposal density $p$ where we know how to draw sample from ( $p$ can be the density of a standard normal distribution).
2. Generate a random number $Y$ from $p$ and another random number $U$ from Uni[0,1].
3. If $U<\frac{f(Y)}{M \cdot p(Y)}$, we set $X=Y$. Otherwise go back to the previous step to draw another new pair of $Y$ and $U$.

The above procedure is called rejection sampling (or rejection-acceptance sampling). If we want to generate $X_{1}, \cdots, X_{n}$ from $f$, we can apply the above procedure multiple times until we accept $n$ points.

Theorem 9.3 The observations generated by the rejection sampling are IID from a density $f(x)$.

## Proof:

Now we consider the CDF of $X$.

$$
\begin{align*}
P(X \leq x) & =P(Y \leq x \mid \operatorname{accept} Y) \\
& =P\left(Y \leq x \left\lvert\, U<\frac{f(Y)}{M \cdot p(Y)}\right.\right) \\
& =\frac{P\left(Y \leq x, U<\frac{f(Y)}{M \cdot p(Y)}\right)}{P\left(U<\frac{f(Y)}{M \cdot p(Y)}\right)} . \tag{9.2}
\end{align*}
$$

Note that in the last equality, we used the definition of conditional probability.

For the numerator, using the feature of conditional probability,

$$
\begin{aligned}
P\left(Y \leq x, U<\frac{f(Y)}{M \cdot p(Y)}\right) & =\int P\left(Y \leq x, \left.U<\frac{f(Y)}{M \cdot p(Y)} \right\rvert\, Y=y\right) p(y) d y \\
& =\int P\left(y \leq x, U<\frac{f(y)}{M \cdot p(y)}\right) p(y) d y \\
& =\int I(y \leq x) P\left(U<\frac{f(y)}{M \cdot p(y)}\right) p(y) d y \\
& =\int_{-\infty}^{x} \frac{f(y)}{M \cdot p(y)} p(y) d y \\
& =\frac{1}{M} \int_{-\infty}^{x} f(y) d y
\end{aligned}
$$

Note that in the fourth equality, we use the fact that the choice of $M: M \geq \sup _{x} \frac{f(x)}{p(x)}$ ensures

$$
\frac{f(y)}{M \cdot p(y)} \leq 1 \quad \forall y
$$

For the denominator, using the similar trick,

$$
\begin{aligned}
P\left(U<\frac{f(Y)}{M \cdot p(Y)}\right) & =\int P\left(\left.U<\frac{f(Y)}{M \cdot p(Y)} \right\rvert\, Y=y\right) p(y) d y \\
& =\int P\left(U<\frac{f(y)}{M \cdot p(y)}\right) p(y) d y \\
& =\int \frac{f(y)}{M \cdot p(y)} p(y) d y \\
& =\frac{1}{M} \int f(y) d y=\frac{1}{M} .
\end{aligned}
$$

Thus, putting altogether into equation (9.2), we obtain

$$
P(X \leq x)=\frac{P\left(Y \leq x, U<\frac{f(Y)}{M \cdot p(Y)}\right)}{P\left(U<\frac{f(Y)}{M \cdot p(Y)}\right)}=\frac{\frac{1}{M} \int_{-\infty}^{x} f(y) d y}{\frac{1}{M}}=\int_{-\infty}^{x} f(y) d y
$$

which means that the random variable $X$ does have the density $f$.

Note that the rejection sample can be applied to scenarios where we only have access to a function $q(x) \propto$ $f(x)$. We just need to replace $f(x)$ by $q(x)$ and modify the upper bound as $M \geq \sup _{x} \frac{q(x)}{p(x)}$.
Here are some features about the rejection sampling:

- Using the rejection sampling, we can generate sample from any density $f$ as long as we know the closed form of $f$.
- If we do not choose $M$ well, we may reject many realizations of $Y, U$ to obtain a single realization of $X$.
- There is an upper on $M$ at the first step: $M \geq \sup _{x} \frac{f(x)}{p(x)}$.
- In practice, we want to choose $M$ as small as possible because a small $M$ leads to a higher chance of accepting $Y$. To see this, note that the denominator $P\left(U<\frac{f(Y)}{M \cdot p(Y)}\right)=P(\operatorname{Accept} Y)=\frac{1}{M}$. Thus, a small $M$ leads to a large accepting probability.
- If you want to learn more about rejection sampling, I would recommend http://www.columbia.edu/ ~ks20/4703-Sigman/4703-07-Notes-ARM.pdf.

An application in Bayesian inference. Here we explain how to use the rejection sampling to sample from a posterior distribution. Let $\pi\left(\theta \mid X_{1}, \cdots, X_{n}\right)$ be the posterior distribution, $L\left(\theta \mid X_{1}, \cdots, X_{n}\right)$ be the likelihood function, and $\pi(\theta)$ be the prior distribution. Also, let $\widehat{\theta}_{M L E}=\operatorname{argmax}_{\theta} L\left(\theta \mid X_{1}, \cdots, X_{n}\right)$ be the MLE.

Then we can generate points from the posterior distribution with the followings:

1. Generate $\theta$ from prior distribution $\pi$ and $U$ from $\operatorname{Uni}(0,1)$ independently.
2. Accept $\theta$ if $U<\frac{L\left(\theta \mid X_{1}, \cdots, X_{n}\right)}{L\left(\widehat{\theta}_{M L E} \mid X_{1}, \cdots, X_{n}\right)}$.

The $\theta$ 's that are accepted from the above two steps are IID from the posterior distribution $\pi\left(\theta \mid X_{1}, \cdots, X_{n}\right)$.
Why this approach works? Well here is how each quantity corresponds to the ones in rejection sampling (after rescaling):

$$
\begin{aligned}
\pi\left(\theta \mid X_{1}, \cdots, X_{n}\right) & \sim f(x) \\
L\left(\theta \mid X_{1}, \cdots, X_{n}\right) & \sim \frac{f(x)}{p(x)} \\
\pi(\theta) & \sim p(x) \\
L\left(\widehat{\theta}_{M L E} \mid X_{1}, \cdots, X_{n}\right) & \sim M
\end{aligned}
$$

Note that the acceptance probability is

$$
p_{a}=\frac{\int L\left(\theta \mid X_{1}, \cdots, X_{n}\right) \pi(\theta) d \theta}{L\left(\widehat{\theta}_{M L E} \mid X_{1}, \cdots, X_{n}\right)}=\frac{p\left(X_{1}, \cdots, X_{n}\right)}{L\left(\widehat{\theta}_{M L E} \mid X_{1}, \cdots, X_{n}\right)} .
$$

So the normalization constant of the posterior distribution $p\left(X_{1}, \cdots, X_{n}\right)$ can be estimated using

$$
\widehat{p}\left(X_{1}, \cdots, X_{n}\right)=\widehat{p}_{a} \cdot L\left(\widehat{\theta}_{M L E} \mid X_{1}, \cdots, X_{n}\right)
$$

where $\widehat{p}_{a}$ is the empirical acceptance probability.
In the analysis of rejection sampling, there are two random quantities that are often of great interest. The first quantity occurs in the case where we are given a fixed amount of target sample, say $m$, and we are interested in the number of points we generate to accept $m$ points. Let $M$ be such a value. The behavior of $M$ is something we would like to analyze because the difference $M-m$ informs us the amount of additional points we need to generate to obtain $m$ points. Note that in ideal case, $M$ is a random variable following a negative binomial distribution with parameter $\left(m, p_{a}\right)$.

The other quantity occurs in situations where we have a fixed budget $n_{0}$ to generate points. Let $N$ be the number of accepted points. This quantity $N$ is also of research interest because we can use the ratio $\frac{N}{n_{0}}$ as an estimate of the acceptance probability $p_{a}$. Note that in the ideal case, the quantity $N$ follows from a binomial distribution with parameter $\left(n_{0}, p_{a}\right)$.

### 9.3.2 MCMC: Metropolis-Hastings

Although rejection sampling is very powerful, it has a limitation that we need to know the MLE $\widehat{\theta}_{M L E}$, which is often a challenge problem when the likelihood function is non-convex. Moreover, in many cases such as Bayesian analysis, we may not know the exact value of the density function $f(x)$ but only know the value up to a constant. This is because the posterior

$$
\pi\left(\theta \mid X_{1}, \cdots, X_{n}\right) \propto p\left(X_{1}, \cdots, X_{n} \mid \theta\right) \cdot \pi(\theta)
$$

Evaluating the likelihood value and the prior are simple but computing the integral over $\theta$ is hard. It is desirable to have a method that allows us to sample from $\pi\left(\theta \mid X_{1}, \cdots, X_{n}\right)$ with only access to $p\left(X_{1}, \cdots, X_{n} \mid \theta\right)$ and $\pi(\theta)$.

The Markov Chain Monte Carlo (MCMC) is a tool that allows us to do this. The idea of MCMC is to generate an ergodic Markov chain $\left\{X_{n}\right\}$ from a stationary distribution that is the same as the distribution we want to generate from. In the Bayesian setting, the stationary distribution will be the posterior distribution.

Let $\pi(\theta)$ be the PDF that we want to generate from with $\theta \in S$ that is univariate (the case of discrete random variable is the same by replacing PDF with PMF). The Metropolis-Hastings algorithm is a simple approach to generate from $\pi(\theta)$ for a univariate $\theta$. It proceeds as follows:

- Input: an initial value $x_{0} \in S$ and a proposal function $q(x \mid y)$ with $x, y \in S$.
- Start with an initial value $X_{0}=x_{0}$.
- For $n=0, \cdots, N$, do the following:

1. Simulate a candidate value $Y_{n} \sim q\left(\cdot \mid X_{n}\right)$.
2. Compute the Metropolis-Hastings acceptance probability:

$$
a(x, y)=\min \left\{\frac{\pi(y) \times q(x \mid y)}{\pi(x) \times q(y \mid x)}, 1\right\}
$$

3. Accept the candidate $Y_{n}$ with a probability $a\left(X_{n}, Y_{n}\right)$. If we do not accept, we keep $X_{n+1}=X_{n}$. Namely,

$$
X_{n+1}= \begin{cases}Y_{n} & \text { with a probability of } a\left(X_{n}, Y_{n}\right) \\ X_{n} & \text { with a probability of } 1-a\left(X_{n}, Y_{n}\right)\end{cases}
$$

The MH algorithm works because it generates a Markov chain with a stationary distribution being $\pi(x)$. The following result shows the case of discrete space Markov chain but the similar idea applies to the continuous space as well.

Proposition 9.4 Assume that $\pi(i)>0$ for all $i \in S$ and that $q(i \mid j)>0 \Leftrightarrow q(j \mid i)>0$ for all $i, j \in S$. Then the Metropolis-Hastings algorithm generates a Markov chain with stationary distribution $\pi$.

Proof: It is easy to see that the sequence $\left\{X_{n}\right\}$ forms a homogeneous Markov chain. Let $\mathbf{P}=\left\{p_{i j}\right\}$ be the transition probability matrix of $X_{n}$. We will prove this by showing that the detailed balance is statisfied. For the case of $i=j$, it is trivial so we assume $i \neq j$.

For $i \neq j$,

$$
p_{i j}=P\left(X_{n+1}=j \mid X_{n}=i\right)=P\left(X_{1}=j \mid X_{0}=i\right)=a_{i j} q(j \mid i) .
$$

Therefore,

$$
\begin{aligned}
\pi_{i} p_{i j}=\pi_{i} a_{i j} q(j \mid i) & = \begin{cases}\pi_{i} q(j \mid i) \times \frac{\pi_{j} q(i \mid j)}{\pi_{i} q(j \mid i)} & \text { if } \frac{\pi_{j} q(i \mid j)}{\pi_{i} q(j \mid i)} \leq 1 \\
\pi_{i} q(j \mid i) \times 1 & \text { otherwise }\end{cases} \\
& = \begin{cases}\pi_{j} q(i \mid j) & \text { if } \pi_{j} q(i \mid j) \leq \pi_{i} q(j \mid i) \\
\pi_{i} q(j \mid i) & \text { if } \pi_{j} q(i \mid j)>\pi_{i} q(j \mid i)\end{cases}
\end{aligned}
$$

Noe we consider $\pi_{j} p_{j i}$ :

$$
\begin{aligned}
\pi_{j} p_{j i}=\pi_{j} a_{j i} q(i \mid j) & = \begin{cases}\pi_{j} q(i \mid j) \times \frac{\pi_{i} q(j \mid i)}{\pi_{j} q(i \mid j)} & \text { if } \frac{\pi_{i} q(j \mid i)}{\pi_{j} q(i \mid j)} \leq 1 \\
\pi_{j} q(i \mid j) \times 1 & \text { otherwise }\end{cases} \\
& = \begin{cases}\pi_{i} q(j \mid i) & \text { if } \pi_{i} q(j \mid i) \leq \pi_{j} q(i \mid j) \\
\pi_{j} q(i \mid j) & \text { if } \pi_{i} q(j \mid i)>\pi_{j} q(i \mid j)\end{cases}
\end{aligned}
$$

which is the same as $\pi_{i} p_{i j}$.
So $\pi$ satisfies the detailed balance, it is a stationary distribution.
To make sure that this is a Markov chain with irreducible state space, we choose the proposal density $q(y \mid x)>0$ for all $x, y \in S$. To simplify the problem, here we assume that $S \subset \mathbb{R}$ but you can easily generalize it to higher dimensions. A common example is to use the random walk proposal:

$$
Y_{n}=X_{n}+\epsilon_{n}
$$

where $\epsilon_{n}$ is some perturbation independent of $X_{n}$ with $\mathbb{E}\left(\epsilon_{n}\right)=0$. A concrete example is to choose $\epsilon \sim$ $N\left(0, \sigma^{2}\right)$ for some pre-specified $\sigma^{2}$. In this case,

$$
q(y \mid x)=\frac{1}{\sqrt{2 \pi \sigma^{2}}} e^{\frac{1}{2 \sigma^{2}}\|x-y\|^{2}}
$$

When $S$ is a connected set and $\pi(x)>0$ for all $x \in S$, there is no need to choose a proposal that has infinite support. Instead, we can choose $q(y \mid x)$ such that there exists $\delta, \epsilon>0$ so that $q(y \mid x)>\epsilon$ if $|x-y|<\delta$. This allows the perturbation $\epsilon_{n}$ to have a finite support. For instance, we may choose

$$
\epsilon_{n} \sim \operatorname{Uni}(-\delta, \delta)
$$

Note that often people choose the proposal $q$ to be isotropic, namely,

$$
q(y \mid x)=q(\|y-x\|)
$$

Both normal perturbation and uniform perturbation are examples leading to an isotropic proposal. Isotropic proposals have an advantage that the update probability becomes very simple. If the value $\frac{\pi(y)}{\pi(x)}$ is greater than or equal to 1 (i.e., the proposed point has a higher value compared to the current point), we move to the proposed point. Otherwise with a probability of the density ratio, we move to the proposed point.

The choice of perturbation is often a challenging question. There are two quantities we want to optimize the MCMC algorithm - the acceptance rate and the exploration rate (speed of exploring the state space $S$ ). We want high acceptance rate as well as a high exploration rate but they are often inversely related to each other. To see this, consider the following example.

Example. Support that $S=\mathbb{R}$ and our target is a univariate standard normal distribution, i.e.,

$$
\pi(x)=\frac{1}{\sqrt{2 \pi}} \exp \left(-x^{2} / 2\right)
$$

We use the proposal density with a uniform perturbation such that

$$
q(y \mid x)=\frac{1}{2 \delta} I(|x-y| \leq \delta)
$$

Let $U_{1} \sim \operatorname{Uni}(-\delta, \delta)$. Then given a current state $x^{(t)}$, the next state will be

$$
x^{(t+1)}= \begin{cases}x^{(t)}+U_{1} & \text { with a probability } p_{t}=\min \left\{\exp \left[\left(\left(x^{(t)}\right)^{2}-\left(x^{(t)}+U_{1}\right)^{2}\right) / 2\right], 1\right\} \\ x^{(t)} & \text { with a probability } 1-p_{t}\end{cases}
$$

Here, as you can see, if $\delta$ is small, the chance that we accept the proposal is generally higher but our exploration of $S$ will be slow (leading to a Markov chain with a high dependence). On the other hand, if $\delta$ is large, the chance we accept the proposal is low but we can quickly explore state space.

Example: MCMC for a $\chi^{2}$ distribution. In the following, we use the MCMC to generate points from a $\chi_{5}^{2}$ :


We use $q(y \mid x) \sim N\left(x-y, 0.5^{2}\right)$ and an initial point 0.5 and run the MCMC to generate 30000 points. Note that generally, the initial point is not important and we will remove the first few points in the MCMC chain (this is also called a burn out). The left panel shows the histogram of the MCMC points and the black curve denotes the true density curve. In the right panel, we display the trajectory of the MCMC; this plot is also called the trace plot. The trace plot will be useful in examining the behavior of the MCMC (we will talk about it later).

Example: MCMC for a Gaussian mixture. To show that MCMC can be applied a wide class of problem, we implement it to generate points from a Gaussian mixture model with the density

$$
p(x)=0.7 \phi(x ; 0,1)+0.3 \phi(x ; 5,1)
$$

where $\phi\left(x ; \mu, \sigma^{2}\right)$ is the PDF of a normal distribution with mean $\mu$ and variance $\sigma^{2}$. We first consider the case where the proposal $q(y \mid x) \sim N(x-y, 1)$ :


The MCMC did a good job in generating points from the Gaussian mixture. Now we consider an interesting case where we decrease the variance of the proposal $q(y \mid x)$ to $N\left(x-y, 0.2^{2}\right)$. Here is the result of MCMC after running it 30000 times:



We see a biased result in our MCMC! In the left panel, the smaller bump was underestimated. The trajectory plot in the right panel explains what was happening - at the beginning, the MCMC was moving around the second mode (small bump centered at 5) and then it switches to the big bump and stuck there. If we compare this to the trajectory where we have a proposal with a higher variance, we see that when the proposal has a higher variance, MCMC switches between the two bumps very frequently but when the proposal has a low variance, it does not switch that frequently. In other word, the variance of the proposal determines the speed of mixing in the MCMC. When MCMC has a slow speed of mixing (i.e., variance of the proposal is low), we need to run it a lot longer to obtain a stable result.

So should we always choose a huge variance in our proposal? Not really. Remembered that in the previous example, we have demonstrated that a high variance proposal may lead to an MCMC with a low acceptance rate. Here is what will happen if we increase the variance of the proposal to 1000:


Again, we see a biased result and the trace plot shows several flat line, indicating that the chain was staying in the same value for a long time, which is what we expect when the acceptance probability is low.

### 9.3.3 MCMC: Gibbs Sampling

Gibbs sampling is an alternative approach to sample from a target PDF/PMF based on the idea of MCMC when the target PDF/PMF is multivariate. The appealing feature of Gibbs sampling is that we only need to know the conditional PDF/PMF of the target rather than the joint PDF/PMF. But we do need to know how to sample from the conditional PDF/PMF of each variable given the others.

To illustrate the idea, we start with discrete state space with two variables $x_{1}, x_{2}$ such that $x_{1} \in S_{1}$ and $x_{2} \in S_{2}$. Our goal is to generate from the target PDF/PMF $\pi\left(x_{1}, x_{2}\right)=P\left(X_{1}=x_{1}, X_{2}=x_{2}\right)$. Let $X_{1}=x_{1}^{(0)}$ and $X_{2}=x_{2}^{(0)}$ be the initial starting point. The Gibbs sampling uses the following iterative updates:

1. $P\left(X_{1}=x_{1}^{(t+1)} \mid X_{2}=x_{2}^{(t)}\right)=\pi\left(X_{1}=x_{1}^{(t+1)} \mid X_{2}=x_{2}^{(t)}\right)$.
2. $P\left(X_{2}=x_{2}^{(t+1)} \mid X_{1}=x_{1}^{(t+1)}\right)=\pi\left(X_{2}=x_{2}^{(t+1)} \mid X_{1}=x_{1}^{(t+1)}\right)$.

Thus, the transition probability matrix becomes

$$
\mathbf{P}\left(x_{1}^{(t+1)}, x_{2}^{(t+1)} \mid x_{1}^{(t)}, x_{2}^{(t)}\right)=\pi\left(X_{2}=x_{2}^{(t+1)} \mid X_{1}=x_{1}^{(t+1)}\right) \pi\left(X_{1}=x_{1}^{(t+1)} \mid X_{2}=x_{2}^{(t)}\right)
$$

An interesting fact is that the resulting Markov chain does not satisfy the detailed balance (so the chain is not reversible) but it does satisfy the global balance. To derive the global balance, we need to show that

$$
\pi\left(x_{1}, x_{2}\right)=\sum_{y_{1} \in S_{1}, y_{2} \in S_{2}} \pi\left(y_{1}, y_{2}\right) \mathbf{P}\left(x_{1}, x_{2} \mid y_{1}, y_{2}\right)
$$

The right-hand sided (RHS) equals

$$
\begin{aligned}
\mathrm{RHS} & =\sum_{y_{1} \in S_{1}, y_{2} \in S_{2}} \pi\left(y_{1}, y_{2}\right) \mathbf{P}\left(x_{1}, x_{2} \mid y_{1}, y_{2}\right) \\
& =\sum_{y_{1} \in S_{1}, y_{2} \in S_{2}} \pi\left(y_{1}, y_{2}\right) \pi\left(X_{2}=x_{2} \mid X_{1}=x_{1}\right) \pi\left(X_{1}=x_{1} \mid X_{2}=y_{2}\right) \\
& =\sum_{y_{1} \in S_{1}, y_{2} \in S_{2}} \pi\left(y_{1}, y_{2}\right) \frac{\pi\left(x_{1}, x_{2}\right)}{\pi\left(X_{1}=x_{1}\right)} \frac{\pi\left(x_{1}, y_{2}\right)}{\pi\left(X_{2}=y_{2}\right)} \\
& =\sum_{y_{2} \in S_{2}} \pi\left(X_{2}=y_{2}\right) \frac{\pi\left(x_{1}, x_{2}\right)}{\pi\left(X_{1}=x_{1}\right)} \frac{\pi\left(x_{1}, y_{2}\right)}{\pi\left(X_{2}=y_{2}\right)} \\
& =\frac{\pi\left(x_{1}, x_{2}\right)}{\pi\left(X_{1}=x_{1}\right)} \sum_{y_{2} \in S_{2}} \pi\left(x_{1}, y_{2}\right) \\
& =\pi\left(x_{1}, x_{2}\right)
\end{aligned}
$$

so it satisfies the global balance.
When we have $d$ variables $x_{1}, \cdots, x_{d}$, the Gibbs sampler is often done by using a sequential scan. Let $x_{1: j}=\left(x_{1}, \cdots, x_{j}\right)$. Given $x^{(0)}=\left(x_{1}^{(0)}, \cdots, x_{d}^{(0)}\right)$, we update using the following way

1. $P\left(X_{1}=x_{1}^{(t+1)} \mid X_{2: d}=x_{2: d}^{(t)}\right)=\pi\left(X_{1}=x_{1}^{(t+1)} \mid X_{2: d}=x_{2: d}^{(t)}\right)$.
2. $P\left(X_{2}=x_{2}^{(t+1)} \mid X_{1}=x_{1}^{(t+1)}, X_{3: d}=x_{3: d}^{(t)}\right)=\pi\left(X_{2}=x_{2}^{(t+1)} \mid X_{1}=x_{1}^{(t+1)}, X_{3: d}=x_{3: d}^{(t)}\right)$.
3. ...
4. $P\left(X_{d}=x_{d}^{(t+1)} \mid X_{1:(d-1)}=x_{1:(d-1)}^{(t+1)}\right)=\pi\left(X_{d}=x_{d}^{(t+1)} \mid X_{1:(d-1)}=x_{1:(d-1)}^{(t+1)}\right)$.

Namely, we keep updating from $x_{1}$, then $x_{2}$, then all the way to $x_{d}$ and we always use the latest value of other variables.

In addition to the sequential scan, the random scan is also a popular approach. Let $x_{-i}=\left(x_{1}, \cdots, x_{i-1}, x_{i+1}, \cdots, x_{d}\right)$. The random scan updates as follows:

1. Randomly select an index $i \in\{1, \cdots, d\}$ from a multinomial distribution.
2. For the selected index $i$, we update it by $x_{i}^{(t+1)} \sim \pi\left(x_{i} \mid x_{-i}^{(t)}\right)$ and set $x_{-i}^{(t+1)}=x_{-i}^{(t)}$.

It is very simple to generalize Gibbs sampler to continuous state space. We just replace the PMF in the above by a PDF. Then the sequential scan and random scan can be defined easily.

## Remark.

- Sometimes, we only know the value of $\pi\left(X_{1}=x_{1} \mid X_{2}=x_{2}\right)$ up to some constant rather than able to directly sample from it. This occurs in the case of computing the posterior distribution of multiple parameters. In this case, we can combine the Metropolis-Hastings algorithm and Gibbs sampler - we use the Metropolis-Hastings algorithm for sampling from $\pi\left(x_{1} \mid x_{2}\right)$ and $\pi\left(x_{2} \mid x_{1}\right)$ and use the Gibbs sampling to obtain the joint result.
- In cases we know how to sample from a conditional PDF/PMF with multiple variables, we do need to update each variable once at a time but we can just update multiple variable in one shot. This is called the block Gibbs sampling.

Gibbs sampling relies on the conditional PDF/PMF $\pi\left(x_{i} \mid x_{-i}\right)$. This quantity determines the transition probability. So the property of the corresponding Markov chain relies on $\pi\left(x_{i} \mid x_{-i}\right)$ for each $i$. In Markov chain theory, the irreducibility is an important property and it turns out that if $\pi$ satisfies the positivity condition, then the Gibbs sampler creates a Markov chain that is irreducible. For a density $\pi\left(x_{1}, \cdots, x_{d}\right)$, it satisfies positivity condition if every marginal density $\pi_{i}\left(x_{i}\right)>0$ implies that $\pi\left(x_{1}, \cdots, x_{d}\right)>0$.

Proposition 9.5 If a density $\pi\left(x_{1}, \cdots, x_{d}\right)$ satisfies the positivity condition, then $\pi\left(x_{i} \mid x_{-i}\right)>0$ for any $x_{i}, x_{-i}$ such that $\pi_{i}\left(x_{i}\right)>0$ and $\pi\left(x_{-i}\right)>0$.

The positivity condition implies that the support of the joint density is the Cartesian product of the support of the marginals. This proposition further shows that if the target density satisfies positivity condition, then the Gibbs sampler is irreducible. Note that positivity is sufficient for irreducibility but not necessary.

## A Theory of Markov Chain

Note that this appendix is from STAT 516. We exclude the proofs and some details. If you are interested in the details, please check http://faculty.washington.edu/yenchic/18A_stat516/Lec3_DTMC_p1.pdf.

A discrete time stochastic process $\left\{X_{n}: n=0,1,2, \cdots\right\}$ is called a Markov chain if for every $x_{0}, x_{1}, \cdots, x_{n-2}, i, j \in$ $S$ and $n \geq 0$,

$$
P\left(X_{n}=i \mid X_{n-1}=j, \cdots, X_{0}=x_{0}\right)=P\left(X_{n}=i \mid X_{n-1}=j\right)
$$

whenever both sides are well-defined. The Markov chain has a one-step memory.
If the distribution function $P\left(X_{n}=x_{n} \mid X_{n-1}=x_{n-1}\right)=p_{i j}$ is independent of $n$, we called $\left\{X_{n}\right\}$ a homogeneous Markov chain. Otherwise we called it an inhomogeneous Markov chain. For a homogeneous Markov chain,

$$
\sum_{j \in S} p_{i j}=1, \quad p_{i j} \geq 0
$$

for every $i, j$. Note that sometimes people write $p_{i j}=p_{i \rightarrow j}$, where $p_{i \rightarrow j}$ stands for that the probability moving from state $i$ to state $j$.

Because $S$ is a discrete set, we often label it as $S=\{1,2,3, \cdots, s\}$ and the elements $\left\{p_{i j}: i, j=1, \cdots, s\right\}$ forms an $s \times s$ matrix $\mathbf{P}=\left\{p_{i j}\right\}$. $\mathbf{P}$ is called the transition (probability) matrix (t.p.m). The property of homogeneous Markov chain implies that

$$
\begin{equation*}
\mathbf{P} \geq 0, \quad \mathbf{P} \mathbf{1}_{s}=\mathbf{1}_{s} \tag{9.3}
\end{equation*}
$$

where $\mathbf{1}_{s}=(1,1,1,1, \cdots, 1)^{T}$ is the vector of 1 's. Note that any matrix satisfying equation (9.3) is called a stochastic matrix.

## Example 3: SIS (Susceptible-Infected-Susceptible) model.

Suppose we observe an individual over a sequence of days $n=1,2, \ldots$ and classify this individual each day as

$$
X_{n}= \begin{cases}I & \text { if infected } \\ S & \text { if susceptible }\end{cases}
$$

We would like to construct a stochastic model for the sequence $\left\{X_{n}: n=1,2, \ldots\right\}$. One possibility is to assume that the $X_{n}$ 's are independent with $P\left(X_{n}=I\right)=1-P\left(X_{n}=S\right)=\alpha$. However, this model is not very realistic since we know from experience that the individual is more likely to stay infected if he or she is already infected.

Since Markov chains are the simplest models that allow us to relax independence, we proceed by defining a transition probability matrix:

$$
\left.\mathbf{P}=\begin{array}{c}
\nearrow \\
I \\
S
\end{array} \begin{array}{cc}
I & S \\
{[1-\alpha} & \alpha \\
\beta & 1-\beta
\end{array}\right]
$$

It can be helpful to visualize the transitions that are possible (have positive probability) by a transition diagram:


## Example 4: Example: Ehrenfest Model of Diffusion.

We start with $N$ particles in a closed box, divided into two compartments that are in contact with each other so that particles may move between compartments. At each time epoch, one particle is chosen uniformly at random and moved from its current compartment to the other compartment. Let $X_{n}$ be the number of particles in compartment 1 (say) at step $n$. This stochastic process is Markov by construction.


Transition probabilities of the Markov chain are:

$$
p_{i j}= \begin{cases}\frac{i}{N} . & \text { for } j=i-1 \\ 1-\frac{i}{N}, & \text { for } j=i+1 \\ 0, & \text { otherwise }\end{cases}
$$

The probability of transfer depends on the number of particles in each compartment. For $N=2$ we have states $0,1,2$ and t.p.m.

$$
\mathbf{P}=\left[\begin{array}{ccc}
0 & 1 & 0 \\
1 / 2 & 0 & 1 / 2 \\
0 & 1 & 0
\end{array}\right]
$$

and the transition diagrm


## A. 1 Property of Markov chain

Suppose we observe a finite realization of the discrete Markov chain and want to compute the probability of this random event:

$$
\begin{aligned}
& P\left(X_{n}=i_{n}, X_{n-1}=i_{n-1}, \cdots, X_{1}=i_{1}, X_{0}=i_{0}\right) \\
& =P\left(X_{n}=i_{n} \mid X_{n-1}=i_{n-1}, \cdots, X_{0}=i_{0}\right) P\left(X_{n-1}=i_{n-1}, \cdots, X_{0}=i_{0}\right) \\
& =p_{i_{n-1}, i_{n}} \times P\left(X_{n-1}=i_{n-1} \mid X_{n-2}=i_{n-2}, \ldots, X_{0}=i_{0}\right) \times P\left(X_{n-2}=i_{n-2}, \ldots, X_{0}=i_{0}\right) \\
& =\cdots \\
& =p_{0 i_{0}} p_{i_{0}, i_{1}} p_{i_{1}, i_{2}} \cdots p_{i_{n-2}, i_{n-1}} p_{i_{n-1}, i_{n}}
\end{aligned}
$$

where $p_{0}=\left(p_{01}, p_{02}, \ldots\right)^{T}$ is the distribution of $X_{0}$, called the initial distribution of $\left\{X_{n}\right\}$. Thus, every Markov chain is fully specified by its transition probability matrix $\mathbf{P}$ and initial distribution $p_{0}$.

The Markov chain has a powerful property called the Markov property - the distribution of $X_{m+n}$ given a set of previous states depends only on the latest available state. Assume that we observe a Markov chain from $n=0,1, \cdots, n$ and we are analyzing the distribution of $X_{m+n}$. Then

$$
\begin{equation*}
P\left(X_{m+n}=j \mid X_{n}=i, X_{n-1}=i_{n-1}, \ldots, X_{0}=i_{0}\right)=P\left(X_{m+n}=j \mid X_{n}=i\right) \tag{9.4}
\end{equation*}
$$

To give an intuition about how we obtain the Markov property, consider a simple case where $n=1$ and $m=2$.

$$
\begin{aligned}
P\left(X_{3}=i_{3} \mid X_{1}=i_{1}, X_{0}=i_{0}\right) & =\sum_{i_{2}} P\left(X_{3}=i_{3}, X_{2}=i_{2} \mid X_{1}=i_{1}, X_{0}=i_{0}\right) \\
& =\sum_{i_{2}} P\left(X_{3}=i_{3} \mid X_{2}=i_{2}, X_{1}=i_{1}, X_{0}=i_{0}\right) P\left(X_{2}=i_{2} \mid X_{1}=i_{1}, X_{0}=i_{0}\right) \\
& \stackrel{!}{=} \sum_{i_{2}} P\left(X_{3}=i_{3} \mid X_{2}=i_{2}, X_{1}=i_{1}\right) P\left(X_{2}=i_{2} \mid X_{1}=i_{1}\right) \\
& =\sum_{i_{2}} P\left(X_{3}=i_{3}, X_{2}=i_{2} \mid X_{1}=i_{1}\right) \\
& =P\left(X_{3}=i_{3} \mid X_{1}=i_{1}\right)
\end{aligned}
$$

To argue the equality ' $\stackrel{!}{=}$, observe that

$$
P\left(X_{3}=i_{3} \mid X_{2}=i_{2}, X_{1}=i_{1}, X_{0}=i_{0}\right)=P\left(X_{3}=i_{3} \mid X_{2}=i_{2}\right)
$$

But we also have that

$$
\begin{aligned}
P\left(X_{3}=i_{3} \mid X_{2}=i_{1}, X_{1}=i_{1}\right) & =\sum_{i_{0}} P\left(X_{3}=i_{3}, X_{0}=i_{0} \mid X_{2}=i_{2}, X_{1}=i_{1}\right) \\
& =\sum_{i_{0}} P\left(X_{3}=i_{3} \mid X_{2}=i_{2}, X_{1}=i_{1}, X_{0}=i_{0}\right) P\left(X_{0}=i_{0} \mid X_{2}=i_{2}, X_{1}=i_{1}\right) \\
& =P\left(X_{3}=i_{3} \mid X_{2}=i_{2}\right) \sum_{i_{0}} P\left(X_{0}=i_{0} \mid X_{2}=i_{2}, X_{1}=i_{1}\right) \\
& =P\left(X_{3}=i_{3} \mid X_{2}=i_{2}\right)
\end{aligned}
$$

We can represent the Markov chain using a simple graphical model:


The claim of the Markov property is now obvious from the theorem on conditional independence and graphical factorization. Indeed, the latest available state serves as a separating set.

Using the graph representation, we obtain an interesting property about a Markov chain: the past and the future are independent given the present.

To see this, again we consider a simple case where $n=2$ and we have $X_{0}, X_{1}, X_{2}$. Here $X_{0}$ denotes the past, $X_{1}$ denotes the present, and $X_{2}$ denotes the future. Then

$$
\begin{aligned}
P\left(X_{0}=i_{0}, X_{2}=i_{2} \mid X_{1}=i_{1}\right) & =\frac{P\left(X_{0}=i_{0}, X_{1}=i_{1}, X_{2}=i_{2}\right)}{P\left(X_{1}=i_{1}\right)} \\
& =\frac{P\left(X_{2}=i_{2} \mid X_{1}=i_{1}, X_{0}=i_{0}\right) P\left(X_{1}=i_{1}, X_{0}=i_{0}\right)}{P\left(X_{1}=i_{1}\right)} \\
& =P\left(X_{2}=i_{2} \mid X_{1}=i_{1}\right) \frac{P\left(X_{1}=i_{1}, X_{0}=i_{0}\right)}{P\left(X_{1}=i_{1}\right)} \\
& =P\left(X_{2}=i_{2} \mid X_{1}=i_{1}\right) P\left(X_{0}=i_{0} \mid X_{1}=i_{1}\right)
\end{aligned}
$$

for any $i_{0}, i_{1}, i_{2} \in S$. Namely, $X_{0}$ and $X_{2}$ are conditional independent given $X_{1}$.

## A. 2 n-step Transition Probability and Chapman-Kolmogorov Equation

For a Markov chain, we define the $n$-step transition probability as

$$
p_{i j}^{(n)}=P\left(X_{n}=j \mid X_{0}=i\right) .
$$

The $n$-step transition probability is time invariant.
Lemma 9.6 Let $\left\{X_{n}\right\}$ be a homogeneous Markov chain and let $p_{i j}^{(n)}$ be the $n$-step transition probability. Then for any $k=0,1,2, \cdots$,

$$
P\left(X_{n+k}=j \mid X_{k}=i\right)=p_{i j}^{(n)} .
$$

The $n$-step transition probabilities are related to each other via the famous Chapman-Kolmogorov Equation.

Lemma 9.7 Let $\left\{X_{n}\right\}$ be a homogeneous Markov chain and let $p_{i j}^{(n)}$ be the $n$-step transition probability. Then for any $n, m=0,1,2, \cdots$

$$
\begin{equation*}
p_{i j}^{(n+m)}=\sum_{k \in S} p_{i k}^{(n)} p_{k j}^{(m)} \tag{9.5}
\end{equation*}
$$

The Chapman-Kolmogorov Equation (equation (9.5)) also implies

$$
\begin{gathered}
\text { Forward equation: } p_{i j}^{(n+1)}=\sum_{k} p_{i k}^{(n)} p_{k j}, \text { for } n=1,2, \ldots \text { and } \\
\text { Backward equation : } p_{i j}^{(n+1)}=\sum_{k} p_{i k} p_{k j}^{(n)}, \text { for } n=1,2, \ldots
\end{gathered}
$$

The forward equation singles out the final step and has the initial state $i$ fixed. The equation is most useful when interest centers on the $p_{i j}^{(n)}$,s for a particular $i$ but all values of $j$. Conversely, the backward equation
singles out the change from the initial state $i$ and has the final state $j$ fixed. This equation is useful when interest is in the $p_{i j}^{(n)}$ 's for a particular $j$ but all values of $i$. The backward equation can be interesting, in particular, when there is an absorbing state $j$ from which there is no escape $\left(p_{j j}=1\right)$.

If we collect the $n$-step transition probabilities into the matrix $\mathbf{P}^{(n)}=\left\{p_{i j}^{(n)}\right\}$, then Kolmogorov's forward and backward equations can be rewritten in matrix form as

$$
\mathbf{P}^{(n+1)}=\mathbf{P}^{(n)} \mathbf{P}=\mathbf{P} \mathbf{P}^{(n)},
$$

where $\mathbf{P}^{(1)}=\mathbf{P}$. Therefore, $\mathbf{P}^{(n)}=\mathbf{P}^{n}$.
This matrix form also implies a cool property about the marginal distribution of $X_{n}$. Assume that $X_{0}$ has a distribution $p_{0}=\left(p_{01}, p_{02}, \cdots, p_{0 s}\right)^{T}$. Let $p_{n}=\left(p_{n 1}, \cdots, p_{n s}\right)^{T}$ be the marginal distribution of $X_{n}$, i.e., $p_{n j}=P\left(X_{n}=j\right)$. Then

$$
\begin{aligned}
p_{n j}=P\left(X_{n}=j\right) & =\sum_{i} P\left(X_{n}=j, X_{0}=i\right) \\
& =\sum_{i} P\left(X_{n}=j \mid X_{0}=i\right) P\left(X_{0}=i\right) \\
& =\sum_{i} p_{0 i} p_{i j}^{(n)}
\end{aligned}
$$

Using the matrix form, we obtain

$$
p_{n}^{T}=p_{0}^{T} \mathbf{P}^{n}
$$

## Example 3: SIS model (revisited).

Recalled that SIS model has a transition probability

$$
\left.\mathbf{P}=\begin{array}{c}
\nearrow \\
0 \\
1
\end{array} \begin{array}{cc}
0 & 1 \\
{\left[\begin{array}{c}
1-\alpha \\
\beta
\end{array}\right.} & 1-\beta
\end{array}\right]
$$

Note that we use $\{0,1\}$ to denote the state $I$ and $S$ in the SIS model.
Assume that the initial distribution $p_{0}=(1-\alpha, \alpha)$, i.e., $P\left(X_{0}=0\right)=1-\alpha$. Moreover, assume that $\beta=1-\alpha$ so the distribution of $X_{1}$ will be

$$
\begin{aligned}
p_{1}^{T} & =p_{0}^{T} \mathbf{P}=(1-\alpha, \alpha)\left[\begin{array}{ll}
1-\alpha & \alpha \\
1-\alpha & \alpha
\end{array}\right] \\
& =\left[(1-\alpha)^{2}+\alpha(1-\alpha), \alpha(1-\alpha)+\alpha^{2}\right]=(1-\alpha, \alpha)=p_{0}^{T}
\end{aligned}
$$

What will be the distribution of $X_{n}$ ? Using the matrix form, we know that

$$
p_{n}^{T}=p_{0}^{T} \mathbf{P}^{n}=p_{1}^{T} \mathbf{P}^{n-1}=p_{0}^{T} \mathbf{P}^{n-1}=\cdots=p_{0}^{T}
$$

Therefore, $P\left(X_{n}=0\right)=1-\alpha$ and $P\left(X_{n}=1\right)=\alpha$ for all $n=1,2,3, \cdots$.
A more interesting fact is the joint distribution of $X_{0}, X_{1}, \cdots, X_{n}$ :

$$
\begin{aligned}
P\left(X_{0}=i_{0}, X_{1}=i_{1}, \cdots X_{n}=i_{n}\right) & =p_{0 i_{0}} p_{i_{0} i_{1}} p_{i_{1} i_{2}} \cdots p_{i_{n-1} i_{n}} \\
& =\alpha^{i_{0}}(1-\alpha)^{1-i_{0}} \alpha^{i_{1}}(1-\alpha)^{1-i_{\alpha}^{i_{0}}(1-\alpha)^{1-i_{0}}} \cdots \alpha^{i_{n}}(1-\alpha)^{1-i_{n}}
\end{aligned}
$$

which is the joint PMF of IID random Bernoulli variables with parameter $\alpha$. Therefore, under this special case, the Markov chain reduces to IID Bernoulli RVs.

Note that in general, when the rows of t.p.m are the same, the corresponding Markov chain is a sequence of IID RVs whose distribution is given by the first/any row of the t.p.m.

## A. 3 Classification of States

We now turn to a classification of the states of a Markov chain that is crucial to understanding the behavior of Markov chains.

An equivalence relation " $\sim$ " is a binary relation between elements of a set satisfying

1. Reflexivity: $i \sim i$ for all $i$
2. Symmetry: $i \sim j \Rightarrow j \sim i$
3. Transitivity: $i \sim j, j \sim k \Rightarrow i \sim k$.

For a set $\mathcal{S}$ and $a \in \mathcal{S},\{s \in \mathcal{S}: s \sim a\}$ is called an equivalence class. Equivalence relations will allow us to split Markov chain state spaces into equivalence classes.
State $j$ is accessible from state $i(i \rightarrow j)$ if there exists $m \geq 0$ such that $p_{i j}^{(m)}>0$. We say that $i$ communicates with $j(i \leftrightarrow j)$ if $j$ is accessible from $i$ and $i$ is accessible from $j$. A set of states $\mathcal{C}$ is a communicating class if every pair of states in $\mathcal{C}$ communicates with each other, and no state in $\mathcal{C}$ communicates with any state not in $\mathcal{C}$.

Proposition 9.8 Communication of states is an equivalence relation.

A set of states $\mathcal{C}$ is closed if $\sum_{j \in \mathcal{C}} p_{i j}=1$ for all $i \in \mathcal{C}$.
Example 6. Consider a Markov chain with the following transition diagram:


Then

- Communication classes: $\{1,2,3\},\{4,5\}$, and $\{6\}$.
- Closed sets: $\{6\},\{4,5,6\},\{1,2,3,4,5,6\}$.

Note that a single state (node in a transition diagram) may belong to multiple closed set. However, a single state can only belong to a communication class.

A Markov chain $\left\{X_{n}\right\}$ is called irreducible if it has only one communication class, i.e., for all $i$ and $j$, $i \leftrightarrow j$. For state $i, d_{i}=\operatorname{gcd}\left\{n \geq 1: p_{i i}^{(n)}>0\right\}$ is called its period, where gcd $=$ greatest common divisor and $d_{i}=+\infty$ if $p_{i i}^{(n)}=0$ for all $n \geq 1$.
Example 7. Consider the example with state space $S=\{0,1,2, \ldots\}$ and $X_{n}$ such that

$$
P\left(X_{n+1}=i \mid X_{n}=0\right)= \begin{cases}p & \text { if } i=1 \\ 1-p & \text { if } i=0 \\ 0 & \text { otherwise }\end{cases}
$$

and for $j \neq 0$

$$
P\left(X_{n+1}=i \mid X_{n}=j\right)= \begin{cases}p & \text { if } i=j+1 \\ 1-p & \text { if } i=j-1 \\ 0 & \text { otherwise }\end{cases}
$$

Then, $d_{2}=\operatorname{gcd}\{2,4,5,6, \ldots\}=1$ though 1 is not in the list (think about why $p_{22}^{(5)}>0$ ).
Example 8: Simple (1-D) Random Walk on the Integers. Consider another example with state space $\mathbb{Z}$. Let $X_{n}$ be the position at time $n$. Then

$$
P\left(X_{n+1}=i-1 \mid X_{n}=i\right)=q \text { and } P\left(X_{n+1}=i+1 \mid X_{n}=i\right)=p
$$

with $p=1-q$. Suppose we start at 0 , then it is clear that we cannot return to 0 after an odd number of steps, so $p_{00}^{(2 n+1)}=0$ for all $n$, i.e.

$$
d_{0}=\operatorname{gcd}\left\{n \geq 1: p_{00}^{(n)}>0\right\}=\operatorname{gcd}\{2,4,6, \ldots\}=2
$$

Proposition 9.9 Period is a communication class property. Namely, $i \leftrightarrow j \Rightarrow d_{i}=d_{j}$.

In a communication class, all states have the same period. Since all states communicate in an irreducible Markov chains, it makes sense to define the period of such a Markov chain. If $d_{i}=1$, state $i$ is called aperiodic. An irreducible Markov chain with period 1 is also called aperiodic.

## A. 4 Strong Markov Property

The Markov property states that the random variable at time $n+m$ conditional on its behavior at time $n$ is independent of the at time prior to $n$. However, what if the time $n$ is random?

Say we are interested in the behavior of $X_{T+m}$ given $X_{T}$, where $T$ is the first time that the Markov chain hits the state 0. Do we still have the Markov property?

Some random time does not have the Markov property. Recall that the Markov property states that for any $m<n<k, X_{m} \perp X_{k} \mid X_{n}$. Let $\left\{X_{n}\right\}$ be a Markov chain with state space $S=\{1,2,3\}$ and consider a random time

$$
T=\inf \left\{n \geq 1:\left(X_{n-1}, X_{n}, X_{n+1}\right)=(2,1,3) \text { or }(3,1,2)\right\}
$$

Then the probability

$$
P\left(X_{k}=3 \mid X_{n}=1, X_{m}=2\right)=P\left(X_{T+1}=3 \mid X_{T}=1, X_{T-1}=2\right)=1 \neq P\left(X_{k}=3 \mid X_{n}=1\right)
$$

because when $X_{m}=X_{n-1}=3, P\left(X_{k}=2 \mid X_{n}=1, X_{m}=3\right)=P\left(X_{T+1}=2 \mid X_{T}=1, X_{T-1}=3\right)=1$ so $P\left(X_{k}=3 \mid X_{n}=1, X_{m}=3\right)=0$. Thus, the conditional probability of $X_{k}$ given $X_{n}$ depends on $X_{m}$, which is a violation of Markov property.

Therefore, it is crucial to identify a class of random time such that the Markov property holds. It turns out that there is a simple class of random times that has the Markov property. This class is called the stopping time.

A random variable $\tau \in\{1,2,3, \cdots\} \cup\{\infty\}$ is called a stopping time if the event $\{\tau=m\}$ can be expressed in terms of $X_{0}, X_{1}, \cdots, X_{m}$. Intuitively, a stopping time is a random time such that we can observe it when the time arrives.

Examples 9: Stopping times.

- Return time. Let $T_{i}=\inf \left\{n \geq 1: X_{n}=i\right\}$ is a stopping time because $\left\{T_{i}=m\right\}=\left\{X_{1} \neq\right.$ $\left.i, \cdots, X_{m-1} \neq i, X_{m}=i\right\} . T_{i}$ is interpreted as the first time the chain returns to state $i$.
- Successive Returns. Let $\tau_{k}$ be the time of the $k$-th return to state $i$ (note that $\tau_{1}=T_{i}$ ). Then $\tau_{k}$ is a stopping time because

$$
\left\{\tau_{k}=m\right\}=\left\{\sum_{n=1}^{m} I\left(X_{n}=i\right)=k, X_{m}=i\right\}
$$

- Counterexample - non-stopping time: Let $\tau=\inf \left\{n \geq 1: X_{n+1}=i\right\}$ is not a stopping time because when the time arrives at $m,\{\tau=m\}=\left\{X_{1} \neq i, \cdots, X_{m} \neq i, X_{m+1}=i\right\}$ depends on $X_{m+1}$.

Stopping time is a very important class of random variable in statistics. Many statistical procedure involves a stopping time. For instance, if we are performing a sequence of experiments and we will stop when we observe certain behavior such as a high signal or enough anomaly. Then the time (of related to the number of sample) is a stopping time. If we want to use data from this sequence of experiments, then we need to use theorems of stopping time (such as optional sampling theorem).

Theorem 9.10 (Strong Markov Property) Let $\left\{X_{n}\right\}$ be a homogeneous Markov chain with a transition probability matrix $\mathbf{P}=\left\{p_{i j}\right\}$ and let $\tau$ be a stopping time with respect to $\left\{X_{n}\right\}$. Then for any integer $k$,

$$
P\left(X_{\tau+k}=j \mid X_{\tau}=i, 0 \leq \ell<\tau, X_{\ell}=i_{\ell}\right)=P\left(X_{k}=j \mid X_{0}=i\right)=p_{i j}^{(k)}
$$

and

$$
P\left(X_{\tau+k}=j \mid X_{\tau}=i\right)=P\left(X_{k}=j \mid X_{0}=i\right)=p_{i j}^{(k)}
$$

## A. 5 Stationary distribution

It is often of great interest to study the limiting behavior of a Markov chain $X_{n}$ when $n \rightarrow \infty$. Here, for simplicity, we assume that our Markov chain is homogeneous. A property of limiting behavior is that $X_{n}$ and $X_{n+1}$ should have the same distribution when $n$ is large. So we are interested in understanding if a Markov chain will eventually converge to a 'stable' distribution (formally, we will call it a stationary distribution). In particular, we would like to know given a Markov chain,

- does this chain has a stationary distribution?
- if so, what is the stationary distribution?
- and does this stationary distribution unique?

It turns out that to answer these questions, we will use concepts related to return time. Thus, we start with understanding properties about return time.

## A.5.1 Return Times

Let $N_{i}=\sum_{i=1}^{\infty} I\left(X_{n}=i\right)$ denotes the number of visits of $\left\{X_{n}\right\}$ to state $i$ not counting the initial state. We also define the following notations:

$$
P\left(\cdot \mid X_{0}=i\right)=P_{i}(\cdot), \quad \mathbb{E}\left(\cdot \mid X_{0}=i\right)=\mathbb{E}_{i}(\cdot)
$$

Note that the quantity $N_{i}$ may equal to $\infty$. It is a finite number with a non-zero probability if there are some states such that when the chain enters one of them, the chain never go back to state $i$. Later we will describe this phenomena using the concept of transient states and recurrent states.

Example 6 (revisited). Consider a Markov chain with the following transition diagram:


As can be seen easily, when the Markov chain enters states $\{4,5,6\}$, it never comes back to any of $\{1,2,3\}$. Thus, $N_{1}$ takes a non-trivial probability to be a finite number.

Let $T_{i}=\inf \left\{n \geq 1: X_{n}=i\right\}$ be the return time. Then the following events can be defined using either $T_{i}$ or $N_{i}$ :

$$
\left\{T_{i}=\infty\right\}=\left\{N_{i}=0\right\}, \quad\left\{T_{i}<\infty\right\}=\left\{N_{i}>0\right\}
$$

These are useful later.
We then define $f_{j i}=P_{j}\left(T_{i}<\infty\right)=P_{j}\left(N_{i}>0\right)$ to be the probability of reaching state $i$ in a finite number of time when the chain starts at state $j$. Note that because $P_{j}\left(T_{i}=\infty\right)+P_{j}\left(T_{i}<\infty\right)=1$, we have $f_{i i}=P_{i}\left(T_{i}<\infty\right)$ and $P_{i}\left(T_{i}=\infty\right)=1-f_{i i}$.

## Proposition 9.11

$$
P_{j}\left(N_{i}=r\right)=\left\{\begin{array}{ll}
f_{j i} f_{i i}^{r-1}\left(1-f_{i i}\right) & \text { if } r \geq 1 \\
1-f_{j i} & \text { if } r=0
\end{array} .\right.
$$

The above formula also gives an interesting result on the case of 'starting from state $i$, returning to state $i$ ' when we set $j=i$ :

$$
P_{i}\left(N_{i}=r\right)=f_{i i}\left(1-f_{i i}\right), \quad P_{i}\left(N_{r}>r\right)=f_{i i}^{r+1}
$$

where $f_{i i}=P_{i}\left(T_{i}<\infty\right)$.
We have seen many situations that $T_{i}$ and $N_{i}$ are closely related. Here is another result about their relationship.

## Corollary 9.12

$$
P_{i}\left(N_{i}=\infty\right)=1 \Leftrightarrow P_{i}\left(T_{i}<\infty\right)=1
$$

and

$$
P_{i}\left(T_{i}<\infty\right)<1 \Leftrightarrow P_{i}\left(N_{i}=\infty\right)=0 \Leftrightarrow \mathbb{E}_{i}\left(N_{i}\right)<\infty .
$$

Corollary 9.12 links the finiteness of $T_{i}$ and $N_{i}$ and also relates it to the expectation. With the following formula of expectation, Corollary 9.12 will be very useful:

$$
\begin{equation*}
\mathbb{E}(X)=\sum_{t=1}^{\infty} P(X \geq t) \tag{9.6}
\end{equation*}
$$

when $X$ is a random variable taking integer values.

## A.5.2 Recurrence and Transience

Based on the return time property, we classify a state $i$ as

$$
\begin{cases}\text { recurrent/persistent, } & \text { if } P_{i}\left(T_{i}<\infty\right)=f_{i i}=1 \\ \text { transient, } & \text { otherwise. }\end{cases}
$$

Furthermore, a recurrent state is called

$$
\begin{cases}\text { positive recurrent, } & \text { if } \mathbb{E}_{i}\left(T_{i}\right)<\infty \\ \text { null recurrent, } & \text { otherwise }\end{cases}
$$

Note that: either $P_{i}\left(N_{i}=\infty\right)=0$ or $P_{i}\left(N_{i}=\infty\right)=1$, with nothing in between (if $f_{i i}<1$, then $P_{i}\left(N_{i}=\right.$ $\infty)=0$; if $f_{i i}=1$, then $P_{i}\left(N_{i}=\infty\right)=1$ ). This, together with Corollary 9.12, implies that $\mathbb{E}_{i}\left(N_{i}\right)=\infty \Longleftrightarrow$ $P_{i}\left(N_{i}=\infty\right)=1$.
Note that:

$$
f_{i i}=P_{i}\left(T_{i}<\infty\right)=1 \Longleftrightarrow P_{i}\left(N_{i}=\infty\right)=1
$$

In other words, if a Markov chain returns to state $i$ in finite time, then the chain visits this state infinitely often.

Proposition 9.13 State $i$ is recurrent $\Longleftrightarrow \sum_{n=1}^{\infty} p_{i i}^{(n)}=\infty$.
Proposition 9.14 Recurrence is a communication class property, i.e. if $i \leftrightarrow j$ and $i$ is recurrent, then $j$ is recurrent.

Example: Gambler's Ruin. Recall that in Gambler's ruin, 0 and $a+b$ states are absorbing. herefore, $\sum_{n=1}^{\infty} p_{00}^{(n)}=\sum_{n=1}^{\infty} p_{a+b, a+b}^{(n)}=\sum_{n=1}^{\infty} 1=\infty$. Hence, 0 and $a+b$ are recurrent states. Once they are reached we stay there forever. Consider state 1 :

$$
P_{1}\left(T_{1}<\infty\right)=1-P_{1}\left(T_{1}=\infty\right) \leq 1-q<1 \text { if } q \in(0,1)
$$

Therefore, by definition, 1 is a transient state. Since states $\{1, \ldots, a+b-1\}$ form a communication class, all states in this class are also transient. These states are transient because they occur a finite number of times before absorption into states 0 or $a+b$.

Example 8: 1-D Random Walk (revisited). Let $X_{n}$ be a random walk on the set of all integers $\mathbb{Z}$ such that

$$
p_{i j}= \begin{cases}p & \text { if } j=i+1 \\ q:=1-p & \text { if } j=i-1\end{cases}
$$

Let's study recurrence of state 0 . We know that $p_{00}^{(2 n+1)}=0$ for all $n \geq 0$ and that, conditional on $X_{0}=0$, $X_{2 n}={ }_{d} \xi_{1}+\cdots+\xi_{2 n}$, where $\xi_{1}, \ldots, \xi_{n}$ are i.i.d. with $P\left(\xi_{i}=1\right)=1-P(\xi=-1)=p$. Hence,

$$
p_{00}^{(2 n)}=P\left(X_{2 n}=0 \mid X_{0}=0\right)=\binom{2 n}{n} p^{n} q^{n}
$$

Recall that Stirlings formula says that $n!\sim n^{n+\frac{1}{2}} e^{-n} \sqrt{2 \pi}$, meaning that

$$
\lim _{n \rightarrow \infty} \frac{n!}{n^{n+\frac{1}{2}} e^{-n} \sqrt{2 \pi}}=1
$$

Therefore,

$$
\begin{aligned}
p_{00}^{(2 n)} & =\frac{(2 n)!}{n!n!} p^{n} q^{n} \\
& \sim \frac{(2 n)^{2 n+\frac{1}{2}} e^{-2 n} \sqrt{2 \pi}}{n^{2 n+1} e^{-2 n} 2 \pi}(p q)^{n} \\
& =\frac{2^{2 n+\frac{1}{2}} n^{2 n+\frac{1}{2}}}{n^{2 n+1} 2^{\frac{1}{2}} \sqrt{\pi}}(p q)^{n}=\frac{(p q)^{n} 2^{2 n}}{\sqrt{\pi n}}=\frac{(4 p q)^{n}}{\sqrt{\pi n}} .
\end{aligned}
$$

We deduce that

$$
\sum_{n=1}^{\infty} p_{00}^{(n)}=\sum_{n=1}^{\infty} p_{00}^{(2 n)}=\infty \quad \Leftrightarrow \quad 4 p q \geq 1 \quad \Leftrightarrow \quad p=q=\frac{1}{2}
$$

(Ratio Test: Let $\sum_{n=1}^{\infty} a_{n}$ be a series which satisfies $\lim _{n \rightarrow \infty}\left|\frac{a_{n+1}}{a_{n}}\right|=k$. If $k>1$ the series diverges, if $k<1$ the series converges.) Conclusion: Only the symmetric random walk is recurrent on $\mathbb{Z}$. Interestingly, the symmetric random walk on $\mathbb{Z}^{2}$ is also recurrent, but it is transient on $\mathbb{Z}^{n}$ for $n \geq 3$, See Brémaud (1999, p. 98).

## A.5.3 Invariant Measures

With the knowledge about recurrence, we are able to talk about the invariant measures and stationary distribution of a stochastic matrix.

A vector $x \neq 0$ is called an invariant measure of a stochastic matrix $\mathbf{P}$ if

- $x_{i} \geq 0$ for each $i$, and
- $x^{T} \mathbf{P}=x^{T}$, i.e., $x_{i}=\sum_{j} x_{j} p_{j i}$ for each $i$.

A probability vector $\pi$ on a Markov chain state space is called a stationary distribution of a stochastic $\operatorname{matrix} \mathbf{P}$ if $\pi^{T} \mathbf{P}=\pi^{T}$, i.e., $\pi_{i}=\sum_{j} \pi_{j} p_{j i}$ for each $i$.

The equation $x^{T} \mathbf{P}=x^{T}$ or $\pi^{T} \mathbf{P}=\pi^{T}$ is also called the global balance equaitons - the probability flow in equals the flow out. Note that for an invariance measure $x$ such that $c=\sum_{i} x_{i}<\infty, c^{-1} x$ is a stationary distribution. But it may happen that $c=\infty$ for some invariant measure so one cannot always normalize it.

Example 9: Two-State Markov Chain. Consider a Markov chain with two states and a transition probability matrix

$$
\mathbf{P}=\left[\begin{array}{cc}
1-p & p \\
q & 1-q
\end{array}\right] \quad 0<p<1, \quad 0<q<1
$$

The global balance equations:

$$
\begin{aligned}
& {\left[\pi_{0}, \pi_{1}\right]\left[\begin{array}{cc}
1-p & p \\
q & 1-q
\end{array}\right]=\left[\pi_{0}, \pi_{1}\right] \text { or }} \\
& \left\{\begin{array}{ll}
(1-p) \pi_{0}+q \pi_{1} & =\pi_{0} \\
p \pi_{0}+(1-q) \pi_{1} & =\pi_{1}
\end{array} \quad \Rightarrow \quad p \pi_{0}=q \pi_{1} \quad \Rightarrow \quad \pi_{0}=\frac{q}{p} \pi_{1} .\right.
\end{aligned}
$$

Using that $\pi_{0}+\pi_{1}=1$, we obtain

$$
\frac{q}{p} \pi_{1}+\pi_{1}=1 \quad \Rightarrow \quad \pi_{1}=\frac{p}{p+q}
$$

and deduce that the global balance equations have the unique solution

$$
\pi^{T}=\left[\frac{q}{p+q}, \frac{p}{p+q}\right]
$$

which is the stationary distribution.
Example: Gambler's Ruin (simple version). Let the total fortune of both players be $a+b=4$. Then

$$
\mathbf{P}=\left[\begin{array}{lllll}
1 & 0 & 0 & 0 & 0 \\
q & 0 & p & 0 & 0 \\
0 & q & 0 & p & 0 \\
0 & 0 & q & 0 & p \\
0 & 0 & 0 & 0 & 1
\end{array}\right]
$$

By inspection, vectors $\pi_{\alpha}^{T}=[\alpha, 0,0,0,1-\alpha]$ satisfy global balance equations: $\pi_{\alpha}^{T} \mathbf{P}=\pi_{\alpha}^{T}$ for any $\alpha \in(0,1)$. So the Gambler's ruin chain has an uncountable number of stationary distributions.

Here, we see the case where a Markov chain may have infinite number of stationary distribution. And in some cases it may not even have a stationary distribution! So returning to our original questions, we would like to know (i) when will a Markov chain has a stationary distribution? and (ii) how to find a stationary distribution? and (ii) when the stationary distribution will be unique?

Theorem 9.15 (Stationary Distribution Criterion) An irreducible homogeneous Markov chain is positive recurrent if and only if it has a stationary distribution. Moreover, if the stationary distribution $\pi^{T}=\left[\pi_{1}, \pi_{2}, \ldots\right]$ exists, it is unique and $\pi_{i}>0$ for all $i \in S$.

To see why positive recurrent is important, consider the following example about a $1-D$ random walk on all integers $\mathbb{Z}$ with $p \neq q$ is transient and recurrent if $p=q=0.5$. This Markov chain has an invariant measure $\mathbf{y}^{T}=[1,1, \ldots]$ for any $p$ and $q$ since

$$
\mathbf{P}=\left[\begin{array}{ccccccc}
\ddots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots \\
\cdots & q & 0 & p & \cdots & \cdots & \cdots \\
\cdots & \cdots & q & 0 & p & \cdots & \cdots \\
\cdots & \cdots & \cdots & q & 0 & p & \cdots \\
\cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \ddots
\end{array}\right]
$$

Since this measure is not normalizable (the state space is $\mathbb{Z}$ ), the 1-D random walk can not be positive recurrent. Thus, we see that an irreducible homogeneous Markov chain can have an invariant measure and still be transient or null recurrent.

In the above case, we are working on a state space $S$ that may possibly contain infinite number of states. In many realistic scenarios the number of states is finite. Does the finiteness of state number gives us any benefits? The answer is yes - and it gives us a huge benefit.

Theorem 9.16 An irreducible homogeneous Markov chain on a finite state space is positive recurrent. Therefore, it always has a stationary distribution.

Finally, we end this lecture on the relation between the return time and the stationary distribution.

Theorem 9.17 Let $\left\{X_{n}\right\}$ be an irreducible homogeneous positive recurrent Markov chain. Then

$$
\pi_{i}=\frac{1}{\mathbb{E}_{i}\left(T_{i}\right)}
$$

where $\pi=\left(\pi_{1}, \cdots, \pi_{s}\right)$ is the stationary distribution of $\left\{X_{n}\right\}$ and $T_{i}=\inf \left\{n \geq 1: X_{n}=i\right\}$ is the return time to state $i$.

