STAT 535: Statistical Machine Learning **Autumn 2024**

Lecture 11: Monte Carlo Simulations and Sampling

Instructor: Yen-Chi Chen

11.1 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 Qor 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.

11.2 Monte Carlo approximations to a mean

11.2.1 Importance sampling

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

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

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

$$\widehat{I}_N = \frac{1}{N}\sum_{i=1}^{N} f(Y_i)\cdot\frac{p(Y_i)}{q(Y_i)}.$$

When $p = q$, this reduces to the simple estimator that uses sample means of $f(Y_i)$ to estimate its expectation.

Theorem 11.1 *The importance sampling is unbiased and has a variance*

$$\mathsf{Var}(\widehat{I}_N) = \frac{1}{N}\left(\int \frac{f^2(y)p^2(y)}{q(y)}dy - I^2\right).$$

Proof:

Bias.

$$\begin{aligned}\mathbb{E}(\widehat{I}_N) - I &= \mathbb{E}\left(f(Y_i)\cdot\frac{p(Y_i)}{q(Y_i)}\right) - I \\ &= \int f(y)\frac{p(y)}{q(y)}q(y)dy - I \\ &= \int f(y)p(y)dy - I = 0.\end{aligned}$$

Thus, it is an unbiased estimator.

Variance.

$$\begin{aligned}\mathsf{Var}(\widehat{I}_N) &= \frac{1}{N}\mathsf{Var}\left(f(Y_i)\cdot\frac{p(Y_i)}{q(Y_i)}\right) \\ &= \frac{1}{N}\left\{\mathbb{E}\left(f^2(Y_i)\cdot\frac{p^2(Y_i)}{q^2(Y_i)}\right) - \underbrace{\mathbb{E}^2\left(f(Y_i)\cdot\frac{p(Y_i)}{q(Y_i)}\right)}_{I^2}\right\} \\ &= \frac{1}{N}\left(\int\frac{f^2(y)p^2(y)}{q(y)}dy - I^2\right).\end{aligned}$$

■

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 $\int \frac{f^2(y)p^2(y)}{q(y)}dy$.

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 11.2 *The importance sampling has minimal variance when*

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

and the optimal variance is 0.

Proof:

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

$$\int A^2(y)dy \int B^2(y)dy \geq \left(\int A(y)B(y)dy\right)^2$$

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

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

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

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

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

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

implying

$$q_{\mathsf{opt}}(y) \propto f(y)p(y) \Longrightarrow p_{\mathsf{opt}}(y) = \frac{f(y)p(y)}{\int f(y)p(y)dy}. \tag{11.1}$$

■

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 $q_{\mathsf{opt}}(y)$, 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

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

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

$$\mu(x) = \int f(y)p(y|x)dy = \frac{\int f(y)p(x,y)dy}{\int p(x,y)dy} = \frac{\int f(y)r(x,y)dy}{\int r(x,y)dy} = \frac{\int f(y)g_x(y)dy}{\int g_x(y)dy},$$

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

$$\widehat{\mu}_N(x) = \frac{\frac{1}{N}\sum_{i=1}^N f(Z_i)\frac{r(x,Z_i)}{q(Z_i)}}{\frac{1}{N}\sum_{j=1}^N \frac{r(x,Z_j)}{q(Z_j)}} = \frac{\sum_{i=1}^N f(Z_i)\phi(Z_i)}{\sum_{j=1}^N \phi(Z_j)} = \sum_{i=1}^N \omega(Z_i)f(Z_i),$$

where $\omega(Z_i) = \frac{\phi(Z_i)}{\sum_{j=1}^N \phi(Z_i)}$ 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):

$$\mathbb{E}(\widehat{\mu}_N(x)) - \mu(x) = O(N^{-1}), \quad \mathsf{Var}(\widehat{\mu}_N(x)) = O(N^{-1}).$$

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 $\theta(V) = I(V \geq v_0)$ 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

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

where $p(v|x)$ 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

$$\widehat{\mathcal{E}}_n = \frac{1}{n}\sum_{i=1}^n \frac{q(X_i)}{w(X_i)}\theta(V_i).$$

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:

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

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 $\mathbb{E}(\theta(V)^2|X=x)$ 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 $\mathbb{E}(\theta(V)^2|X=x)$ and sample the second part from the estimate optimal $w^*(x)$.

11.3 Monte Carlo sampling

11.3.1 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 $\mathsf{Uni}[0,1]$.

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

Proof:

Now we consider the CDF of X.

$$\begin{aligned} P(X \leq x) &= P(Y \leq x|\mathsf{accept}Y) \\ &= P\left(Y \leq x|U < \frac{f(Y)}{M\cdot p(Y)}\right) \\ &= \frac{P\left(Y \leq x, U < \frac{f(Y)}{M\cdot p(Y)}\right)}{P\left(U < \frac{f(Y)}{M\cdot p(Y)}\right)}. \end{aligned} \tag{11.2}$$

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

For the numerator, using the feature of conditional probability,

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

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,

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

Thus, putting altogether into equation (11.2), we obtain

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

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 $P\left(U < \frac{f(Y)}{M\cdot p(Y)}\right) = P(\mathsf{Accept}Y) = \frac{1}{M}$. 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 $\pi(\theta|X_1,\cdots,X_n)$ be the posterior distribution, $L(\theta|X_1,\cdots,X_n)$ be the likelihood function, and $\pi(\theta)$ be the prior distribution. Also, let $\widehat{\theta}_{MLE} = \mathsf{argmax}_\theta L(\theta|X_1,\cdots,X_n)$ be the MLE.

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

1. Generate θ from prior distribution π and U from $\mathsf{Uni}(0,1)$ independently.
2. Accept θ if $U < \frac{L(\theta|X_1,\cdots,X_n)}{L(\widehat{\theta}_{MLE}|X_1,\cdots,X_n)}$.

The θ's that are accepted from the above two steps are IID from the posterior distribution $\pi(\theta|X_1,\cdots,X_n)$.

Why this approach works? Well here is how each quantity corresponds to the ones in rejection sampling (after rescaling):

$$\begin{aligned}\pi(\theta|X_1,\cdots,X_n) &\sim f(x)\\ L(\theta|X_1,\cdots,X_n) &\sim \frac{f(x)}{p(x)}\\ \pi(\theta) &\sim p(x)\\ L(\widehat{\theta}_{MLE}|X_1,\cdots,X_n) &\sim M\end{aligned}$$

Note that the acceptance probability is

$$p_a = \frac{\int L(\theta|X_1,\cdots,X_n)\pi(\theta)d\theta}{L(\widehat{\theta}_{MLE}|X_1,\cdots,X_n)} = \frac{p(X_1,\cdots,X_n)}{L(\widehat{\theta}_{MLE}|X_1,\cdots,X_n)}.$$

So the normalization constant of the posterior distribution $p(X_1,\cdots,X_n)$ can be estimated using

$$\widehat{p}(X_1,\cdots,X_n) = \widehat{p}_a \cdot L(\widehat{\theta}_{MLE}|X_1,\cdots,X_n),$$

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 (m, p_a).

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 (n_0, p_a).

11.3.2 MCMC: Metropolis-Hastings

Although rejection sampling is very powerful, it has a limitation that we need to know the MLE $\widehat{\theta}_{MLE}$, 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

$$\pi(\theta|X_1,\cdots,X_n) \propto p(X_1,\cdots,X_n|\theta)\cdot\pi(\theta).$$

Evaluating the likelihood value and the prior are simple but computing the integral over θ is hard. It is desirable to have a method that allows us to sample from $\pi(\theta|X_1,\cdots,X_n)$ with only access to $p(X_1,\cdots,X_n|\theta)$ 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 $\{X_n\}$ 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 θ. It proceeds as follows:

- Input: an initial value $x_0 \in S$ and a proposal function $q(x|y)$ 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 $Y_n \sim q(\cdot|X_n)$.
  2. Compute the Metropolis-Hastings *acceptance probability*:
  $$a(x,y) = \min\left\{\frac{\pi(y)\times q(x|y)}{\pi(x)\times q(y|x)}, 1\right\}$$
  3. Accept the candidate Y_n with a probability $a(X_n, Y_n)$. If we do not accept, we keep $X_{n+1} = X_n$. Namely,
  $$X_{n+1} = \begin{cases} Y_n & \text{with a probability of } a(X_n,Y_n) \\ X_n & \text{with a probability of } 1-a(X_n,Y_n) \end{cases}$$

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

Proof: It is easy to see that the sequence $\{X_n\}$ forms a homogeneous Markov chain. Let $\mathbf{P} = \{p_{ij}\}$ 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 $i \neq j$,

$$p_{ij} = P(X_{n+1} = j|X_n = i) = P(X_1 = j|X_0 = i) = a_{ij}q(j|i).$$

Therefore,

$$\pi_i p_{ij} = \pi_i a_{ij} q(j|i) = \begin{cases} \pi_i q(j|i) \times \frac{\pi_j q(i|j)}{\pi_i q(j|i)} & \text{if } \frac{\pi_j q(i|j)}{\pi_i q(j|i)} \leq 1 \\ \pi_i q(j|i) \times 1 & \text{otherwise} \end{cases}$$
$$= \begin{cases} \pi_j q(i|j) & \text{if } \pi_j q(i|j) \leq \pi_i q(j|i) \\ \pi_i q(j|i) & \text{if } \pi_j q(i|j) > \pi_i q(j|i) \end{cases}$$

Noe we consider $\pi_j p_{ji}$:

$$\pi_j p_{ji} = \pi_j a_{ji} q(i|j) = \begin{cases} \pi_j q(i|j) \times \frac{\pi_i q(j|i)}{\pi_j q(i|j)} & \text{if } \frac{\pi_i q(j|i)}{\pi_j q(i|j)} \leq 1 \\ \pi_j q(i|j) \times 1 & \text{otherwise} \end{cases}$$
$$= \begin{cases} \pi_i q(j|i) & \text{if } \pi_i q(j|i) \leq \pi_j q(i|j) \\ \pi_j q(i|j) & \text{if } \pi_i q(j|i) > \pi_j q(i|j) \end{cases},$$

which is the same as $\pi_i p_{ij}$.

So π 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 $q(y|x) > 0$ 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 ϵ_n is some perturbation independent of X_n with $\mathbb{E}(\epsilon_n) = 0$. A concrete example is to choose $\epsilon \sim N(0, \sigma^2)$ for some pre-specified σ^2. In this case,

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

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 $q(y|x)$ such that there exists $\delta, \epsilon > 0$ so that $q(y|x) > \epsilon$ if $|x - y| < \delta$. This allows the perturbation ϵ_n to have a finite support. For instance, we may choose

$$\epsilon_n \sim \mathsf{Uni}(-\delta, \delta).$$

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

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

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.,

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

We use the proposal density with a uniform perturbation such that

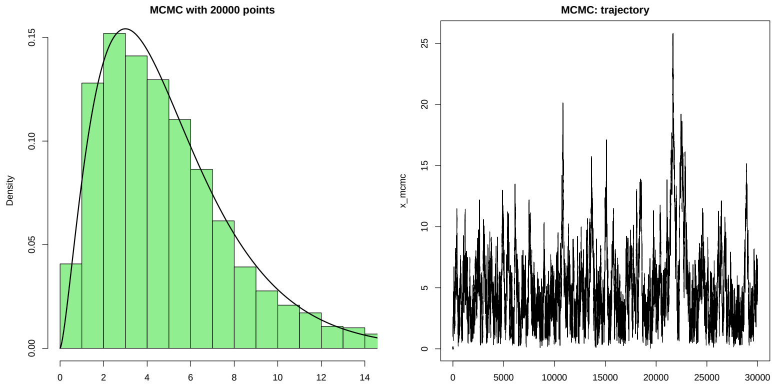
$$q(y|x) = \frac{1}{2\delta} I(|x-y| \leq \delta).$$

Let $U_1 \sim \mathsf{Uni}(-\delta, \delta)$. Then given a current state $x^{(t)}$, the next state will be

$$x^{(t+1)} = \begin{cases} x^{(t)} + U_1 & \text{with a probability } p_t = \min\{\exp\left[\left((x^{(t)})^2 - (x^{(t)} + U_1)^2\right)/2\right], 1\} \\ x^{(t)} & \text{with a probability } 1 - p_t \end{cases}$$

Here, as you can see, if δ 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 δ is large, the chance we accept the proposal is low but we can quickly explore state space.

Example: MCMC for a χ^2 distribution. In the following, we use the MCMC to generate points from a χ^2_5:

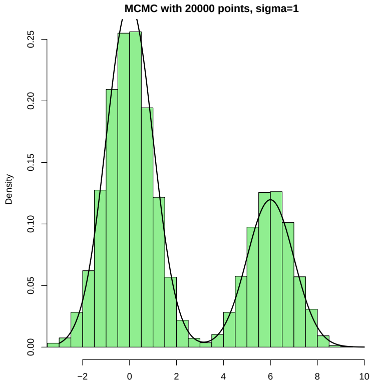


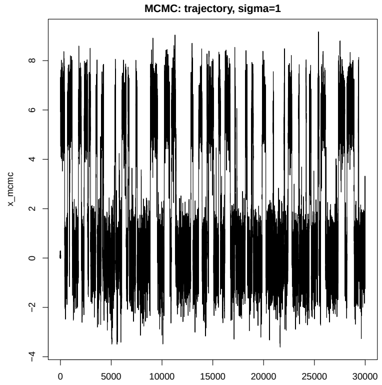
We use $q(y|x) \sim N(x-y, 0.5^2)$ and an initial point 0.5 and run the MCMC to generate 30000 points. Note that generally, the initial point is not important and we will remove the first few points in the MCMC chain (this is also called a burn out). The left panel shows the histogram of the MCMC points 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

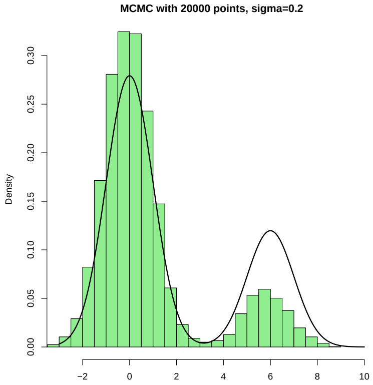
$$p(x) = 0.7\phi(x; 0, 1) + 0.3\phi(x; 5, 1),$$

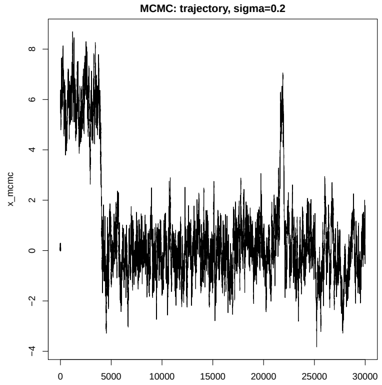
where $\phi(x; \mu, \sigma^2)$ is the PDF of a normal distribution with mean μ and variance σ^2. We first consider the case where the proposal $q(y|x) \sim N(x-y, 1)$:





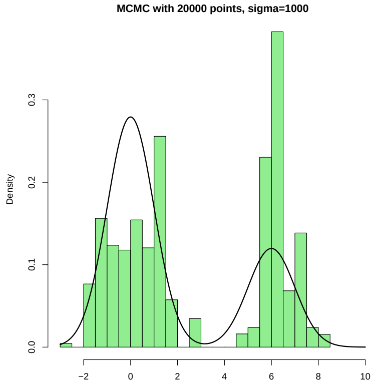
The MCMC did a good job in generating points from the Gaussian mixture. Now we consider an interesting case where we decrease the variance of the proposal $q(y|x)$ to $N(x-y, 0.2^2)$. Here is the result of MCMC after running it 30000 times:

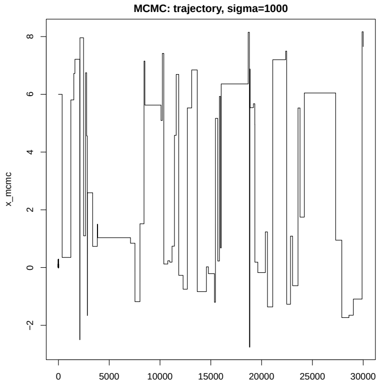




We see a biased result in our MCMC! In the left panel, the smaller bump was underestimated. The trajectory plot in the right panel explains what was happening – at the beginning, the MCMC was moving around the second mode (small bump centered at 5) and then it switches to the big bump and stuck there. 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.

11.3.3 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 $\pi(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$. 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. $P(X_1 = x_1^{(t+1)}|X_2 = x_2^{(t)}) = \pi(X_1 = x_1^{(t+1)}|X_2 = x_2^{(t)})$.

2. $P(X_2 = x_2^{(t+1)}|X_1 = x_1^{(t+1)}) = \pi(X_2 = x_2^{(t+1)}|X_1 = x_1^{(t+1)})$.

Thus, the transition probability matrix becomes

$$\mathbf{P}(x_1^{(t+1)}, x_2^{(t+1)}|x_1^{(t)}, x_2^{(t)}) = \pi(X_2 = x_2^{(t+1)}|X_1 = x_1^{(t+1)})\pi(X_1 = x_1^{(t+1)}|X_2 = x_2^{(t)}).$$

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

$$\pi(x_1, x_2) = \sum_{y_1 \in S_1, y_2 \in S_2} \pi(y_1, y_2)\mathbf{P}(x_1, x_2|y_1, y_2).$$

The right-hand sided (RHS) equals

$$\begin{aligned}
\mathsf{RHS} &= \sum_{y_1\in S_1, y_2\in S_2} \pi(y_1,y_2)\mathbf{P}(x_1,x_2|y_1,y_2)\\
&= \sum_{y_1\in S_1, y_2\in S_2} \pi(y_1,y_2)\pi(X_2=x_2|X_1=x_1)\pi(X_1=x_1|X_2=y_2)\\
&= \sum_{y_1\in S_1, y_2\in S_2} \pi(y_1,y_2)\frac{\pi(x_1,x_2)}{\pi(X_1=x_1)}\frac{\pi(x_1,y_2)}{\pi(X_2=y_2)}\\
&= \sum_{y_2\in S_2} \pi(X_2=y_2)\frac{\pi(x_1,x_2)}{\pi(X_1=x_1)}\frac{\pi(x_1,y_2)}{\pi(X_2=y_2)}\\
&= \frac{\pi(x_1,x_2)}{\pi(X_1=x_1)}\sum_{y_2\in S_2}\pi(x_1,y_2)\\
&= \pi(x_1,x_2)
\end{aligned}$$

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 $x_{1:j}=(x_1,\cdots,x_j)$. Given $x^{(0)}=(x_1^{(0)},\cdots,x_d^{(0)})$, we update using the following way

1. $P(X_1=x_1^{(t+1)}|X_{2:d}=x_{2:d}^{(t)})=\pi(X_1=x_1^{(t+1)}|X_{2:d}=x_{2:d}^{(t)})$.
2. $P(X_2=x_2^{(t+1)}|X_1=x_1^{(t+1)},X_{3:d}=x_{3:d}^{(t)})=\pi(X_2=x_2^{(t+1)}|X_1=x_1^{(t+1)},X_{3:d}=x_{3:d}^{(t)})$.
3. $\cdots$
4. $P(X_d=x_d^{(t+1)}|X_{1:(d-1)}=x_{1:(d-1)}^{(t+1)})=\pi(X_d=x_d^{(t+1)}|X_{1:(d-1)}=x_{1:(d-1)}^{(t+1)})$.

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 $x_{-i}=(x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_d)$. 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 $x_i^{(t+1)}\sim\pi(x_i|x_{-i}^{(t)})$ 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 $\pi(X_1=x_1|X_2=x_2)$ 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 $\pi(x_1|x_2)$ and $\pi(x_2|x_1)$ 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 $\pi(x_i|x_{-i})$. This quantity determines the transition probability. So the property of the corresponding Markov chain relies on $\pi(x_i|x_{-i})$ for each i. In Markov chain theory, the irreducibility is an important property and it turns out that if π satisfies the *positivity condition*, then the Gibbs sampler creates a Markov chain that is irreducible. For a density $\pi(x_1, \cdots, x_d)$, it satisfies positivity condition if every marginal density $\pi_i(x_i) > 0$ implies that $\pi(x_1, \cdots, x_d) > 0$.

Proposition 11.5 *If a density $\pi(x_1, \cdots, x_d)$ satisfies the positivity condition, then $\pi(x_i|x_{-i}) > 0$ for any x_i, x_{-i} such that $\pi_i(x_i) > 0$ and $\pi(x_{-i}) > 0$.*

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.

11.4 Hamiltonian Monte Carlo

The Hamiltonian Monte Carlo (HMC) is a new MCMC approach that has been shown to work better than the usual MH algorithm. It is based on the idea of Hamiltonian dynamics.

The high-level idea of HMC is to generate a proposal from a *better proposal distribution'* and modify the acceptance part so the it has a *higher acceptance rate*. In the usual MH algorithm, we are directly sample from a proposal density $q(y|x)$. The HMC modifies this process using two components: a random momentum (velocity) vector ω and the Hamiltonian dynamics. The momentum is required for every coordinate of the position x. Thus, if $x \in \mathbb{R}^d$, then we also need a vector of d elements for the momentum. As the name suggests, the momentum vector determines how we move x during the dynamics. The randomness is due to the random momentum vector (and the later acceptance part).

The rough idea of one-run HMC is as follows. Starting at the location x_0:

1. (Proposal step 1) We draw a random momentum vector $\omega_0 \sim p(\omega) \propto e^{-V(\omega)}$, where $V(\omega)$ is called the kinematic energy. Often $p(\omega)$ is taken to be a multivariate Gaussian.
2. (Proposal step 2) Then we apply the Hamiltonian dynamics at location x_0 and velocity ω_0 with the Hamiltonian (energy) $H(x, \omega) = -\log \pi(x) + V(\omega)$ and let the dynamics run for time T. This changes (x_0, ω_0) to (x_T, ω_T). Note that the pair (x, ω) is called the state.
3. (Acceptance step) We accept the new location x_T with a probability of

$$a(x_0, \omega_0, x_T, \omega_T) = \min\left\{1, \frac{\exp(-H(x_T, \omega_T))}{\exp(-H(x_0, \omega_0))}\right\}.$$

To approximate the distribution of π, we will iterate the HMC several times.

Note that in the second step (Hamiltonian dynamics), the dynamics is deterministic. Namely, if we start with the same location and the same momentum, we always end up being in the same destination. So for HMC the proposal density $q(x_T|x_0)$ is determined by the density $p(\omega) \propto e^{-V(\omega)}$ and the initial location x_0

To understand what happen in the HMC, we first note that the Hamiltonian contains two parts.

Potential energy. The targeted density $\pi(x)$ is incorporated into the HMC through the Hamiltonian

$$H(x, \omega) = \underbrace{-\log \pi(x)}_{=U(x)} + V(\omega).$$

The quantity $U(x) = -\log \pi(x)$ is also known as the potential energy.

Kinematic energy. The momentum is drawn from the Kinematic energy. The density $p(\omega) \propto e^{-V(\omega)}$ is crucial in the performance of an HMC algorithm. The function V has to be coordinatewise symmetric to ensure the detailed balance equation. In general, we will choose $V(\omega) = \sum_{j=1}^{d} \frac{\omega_j^2}{2m_j}$, where d is the dimension of x and m_j is called the mass of the j-th coordinate. Although this looks fancy, but it implied an extremely sample distribution of $p(\omega)$:

$$p(\omega) \propto e^{-\frac{1}{2}\omega^T M^{-1}\omega} \sim N(0, M),$$

where $M = \mathsf{diag}(m_1, \cdots, m_d)$. So in fact, we are generating the momentum from a multivariate Gaussian (and all coordinates are independent).

Hamiltonian dynamics. The Hamiltonian dynamics governs the usual motion of an object under a specified potential energy and the kinematic energy. It provides excellent description on many physical phenomena such as how planets orbiting around a star. When $H(x, \omega)$ is given, the Hamiltonian dynamics is a deterministic equation of motion. Suppose we start with a location $x(0)$ and a momentum $\omega(0)$, the trajectory of the state $\{x(t), \omega(t) : t \in [0, \infty)\}$ is determined by

$$\frac{dx_j(t)}{dt} \equiv x_j'(t) = \frac{\partial H(x(t), \omega(t))}{\partial \omega_j(t)}, \quad \frac{d\omega_j(t)}{dt} \equiv \omega_j'(t) = -\frac{\partial H(x(t), \omega(t))}{\partial x_j(t)}$$

for $j = 1, \cdots, d$. A powerful feature of Hamiltonian dynamics is that

> *Even if we only have access to $r(x) \propto \pi(x)$, we can still compute the dynamics since the potential energy $U(x) = -\log \pi(x) = -\log r(x) + C_0$ for some constant C_0. We can totally ignore the constant C_0 in practice.*

Because kinematic energy is $V(\omega) = \sum_{j=1}^{d} \frac{\omega_j^2}{m_j}$, the change in location (in j-th coordinate) is simply

$$x_j'(t) = \frac{\omega_j(t)}{m_j}.$$

Thus, given an initial state $(x(0), \omega(0)) = (x_0, \omega_0)$, after running the Hamiltonian dynamics for time T, we will move to the state $(x(T), \omega(T)) = (x_T, \omega_T)$. The new state (x_T, ω_T) is the new proposal. Then in the HMC, we will make a acceptance decision to see if we will accept this proposal.

Here is one caveat in the HMC:

> *No matter we accept or reject the proposal, we will draw a new momentum in the next iteration.*

Namely, only the location x_T will be kept after this iteration. The momentum will be deleted and we will draw a new momentum (from the kinematic energy) without using any information from the previous iteration.

Potential informs momentum: better proposal. The Hamiltonian dynamics allow the target density $\pi(x)$ changes the momentum vector via the equation

$$\omega_j'(t) = -\frac{\partial H(x(t), \omega(t))}{\partial x_j(t)} = \omega_j'(t) = -\frac{\partial \log \pi(x(t))}{\partial x_j(t)}.$$

Thus, even if the original momentum $\omega(0)$ may be pointing toward a bad direction (with least density), the dynamics will adjust its orientation so that it tends to point toward a higher density area.

11.4.1 Features of the HMC

- **(P1): Energy conservation of the Hamiltonian dynamics: high acceptance rate.** The Hamiltonian dynamics has a powerful property called *energy conservation*, which implies that the *acceptance probability is very high.* It is not hard to see that the change of Hamiltonian energy with respect to time is

  $$\begin{aligned}\frac{dH(x(t),\omega(t))}{dt} &= \sum_{j=1}^{d}\left\{\frac{\partial H(x(t),\omega(t))}{\partial \omega_j(t)}\frac{d\omega_j(t)}{dt}+\frac{\partial H(x(t),\omega(t))}{\partial x_j(t)}\frac{dx_j(t)}{dt}\right\}\\ &= 0.\end{aligned}$$

  Namely, the Hamiltonian energy will always stay the same during the dynamics. This is a powerful property! Now we examine the acceptance probability:

  $$a(x_0,\omega_0,x_T,\omega_T)=\max\left\{1,\frac{\exp(-H(x_T,\omega_T))}{\exp(-H(x_0,\omega_0))}\right\}.$$

  The acceptance probability uses the ratio between the initial Hamiltonian energy and the final Hamiltonian energy after applying the dynamics. Because the Hamiltonian energy is conserved during the dynamics, this ratio will always be 1! Namely, *the acceptance probability is* 1 if we apply a real Hamiltonian dynamics. In fact, we need this acceptance step because in practice, we are using a numerical approximation to the Hamiltonian dynamics so there could be some energy loss due to the approximation. So the final acceptance step is to account for this numerical error.

- **(P2): Unique trajectory.** Given initial conditions $x(0)=x_0,\omega(0)=\omega_0$, the Hamiltonian dynamics creates a unique trajectory $(x(t),\omega(t))$. So the only randomness in the HMC is the initial velocity ω_0.

- **(P3): Time-reversal.** Suppose the Hamiltonian dynamics starts at $x(0)=a,\omega(0)=u$ and at time T we obtain $x(T)=b,\omega(T)=w$. Then we have a reversed-time result that if we start the dynamics at $x(0)=b,\omega(0)=-w$, we will obtain $x(T)=a,\omega(T)=u$.

11.4.2 HMC and detailed balance.

Here we have seen that the HMC tends to give a better proposal and have a high acceptance rate. But to make sure we are indeed sampling from the desired density, we need to show that the generated points converge to a stationary distribution that is the desired density π. First, it is easy to see that the generated points form a Markov chain since in each iteration, we only use the information from the previous location. So we only need to show that π satisfies the detailed balance equation of the transition under HMC. In the HMC (that we indeed perform the Hamiltonian dynamics), since the dynamics is deterministic (property (P2)), given the time T being fixed, the mapping

$$(x_0,\omega_0)\rightarrow(x_T,\omega_T)$$

is deterministic. Namely, there exists ϕ_1,ϕ_2 such that $x_T=\phi_1(x_0,\omega_0)$ and $\omega_T=\phi_2(x_0,\omega_0)$. An interesting fact about Hamiltonian dynamics is that if we reverse the time, the trajectory will remain the same (property (P3)). Namely, if we start the dynamics with initial location x_T and momentum $-\omega_T$, after time T we will come back to x_0 and ω_0. Namely, $x_0=\phi_1(x_T,-\omega_T),\omega_0=\phi_2(x_T,-\omega_T)$. Thus, there is a one-one correspondence between $(x_0,\omega_0)\leftrightarrow(x_T,-\omega_T)$.

To show the detail balanced, we need to show that

$$\pi(x)p(x\rightarrow y)=\pi(y)p(x\rightarrow x),$$

where $p(x \to y)$ is the transition density.

Here is an intuitive explanation about the detailed balanced. Suppose that there is only one ω such that $\phi_1(x,\omega) = y$ and let $\tilde{\omega} = \phi_2(x,\omega)$ be the corresponding velocity. Then we also have $\phi_1(y,-\tilde{\omega}) = x$ and $\phi_2(y,-\tilde{\omega}) = \omega$. Thus, there is also only one η such that the dynamics moves (y,η) to (x,ω) and the choice is $\eta = -\tilde{\omega}$.

In this case, $p(x \to y) = p(\omega)$ because ω is the only choice that moves x into y. Similarly, we have $p(y \to x) = p(-\tilde{\omega})$. Then

$$\begin{aligned}
\pi(x)p(x \to y) &= \pi(x)p(\omega) \\
&= \frac{1}{Z_0}\exp\{-U(x) - V(\omega)\} \\
&= \frac{1}{Z_0}\exp\{-H(x,\omega)\} \\
&= \frac{1}{Z_0}\exp\{-H(y,\tilde{\omega})\} \qquad \text{(Energy conservation)} \\
&= \frac{1}{Z_0}\exp\{-U(y) - V(\tilde{\omega})\} \\
&= \pi(y)p(\tilde{\omega}) \\
&= \pi(y)p(-\tilde{\omega}) \qquad (\omega \text{ is coordinatewise symmetric}) \\
&= \pi(y)p(y \to x)
\end{aligned}$$

so the detailed balance is satisfied. Actually, this idea can be generalized to the case where we have more than one momentum leading to y; there is always a one-one correspondence between ω and $\tilde{\omega}$ and the detailed balance is always satisfied.

11.4.3 The HMC algorithm and the leapfrog method

The practical usage of the HMC involves a discretized step of the dynamics. This discretization is called the leapfrog method. Suppose that x_0 is the input location and we are only able to evaluate $r(x) \propto \pi(x)$. Also, let ϵ be the step size in the discretization and L is the number of updates in the dynamics. Namely, $\epsilon \cdot L = T$ is the time that we apply the dynamics.

1. Generate the initial momentum $\omega_0 \sim N(0, M^{-1})$.
2. Set $x^{(0)} = x_0$.
3. For the momentum, make a half-step update:
$$\omega^{(0)} = \omega_0 - \frac{\epsilon}{2}\nabla \log r(x^{(0)}).$$
4. For $\ell = 1, \cdots, L-1$, do the followings:
   (a) Update position: $x^{(\ell)} = x^{(\ell-1)} + \epsilon \cdot \omega^{(\ell-1)}$.
   (b) Update momentum: $\omega^{(\ell)} = \omega^{(\ell-)} - \epsilon \cdot \nabla \log r(x^{(\ell)})$.
5. Make one last update on the position: $x^{(L)} = x^{(L-1)} + \epsilon \cdot \omega^{(L-1)}$.
6. Make another half-step update of the momentum:
$$\omega^{(L)} = \omega^{(L-1)} - \frac{\epsilon}{2}\nabla \log r(x^{(L)}).$$

7. Compute the acceptance probability:

$$a(x_0, \omega_0, x^{(L)}, \omega^{(L)}) = \min\left\{1, \frac{\exp(-H(x^{(L)}, \omega^{(L)}))}{\exp(-H(x_0, \omega_0))}\right\}.$$

8. Accept $x_{\mathsf{new}} = x^{(L)}$ with a probability of $a(x_0, \omega_0, x^{(L)}, \omega^{(L)})$. If we reject, then $x_{\mathsf{new}} = x_0$.

9. Return x_{new}.

A practical challenge is how do we numerically approximate the dynamics part in the HMC algorithm. In the dynamics, the momentum and the location are updated simultaneously. But in practice, we have to make a choice on which one to update first. This leads to a problem that the algorithm is non-symmetric with respect to time. To see this, suppose that we start at x_1 with a momentum ω_1 and move to x_2 with a momentum ω_2. The actual Hamiltonian dynamics is time-reversible, meaning that if we apply the algorithm to $(x_2, -\omega_2)$, we will get back (x_1, ω_1). However, if we only use the leapfrog procedure (step 4), we will not move $(x_2, -\omega_2)$ back to (x_1, ω_1). So the `half step update` (step 3) before and after the `for` loop (step 6) is to resolve this problem and make the algorithm symmetric with respect to time.

Here is an `R` code for the HMC from https://arxiv.org/pdf/1206.1901.pdf, and excellent introduction to the HMC.

```
HMC = function (U, grad_U, epsilon, L, current_q)
{
  q = current_q
  p = rnorm(length(q),0,1)  # independent standard normal variates
  current_p = p
  # Make a half step for momentum at the beginning
  p = p - epsilon * grad_U(q) / 2
  # Alternate full steps for position and momentum
  for (i in 1:L)
  {
    # Make a full step for the position
    q = q + epsilon * p
    # Make a full step for the momentum, except at end of trajectory
    if (i!=L) p = p - epsilon * grad_U(q)
  }
  # Make a half step for momentum at the end.
  p = p - epsilon * grad_U(q) / 2
  # Negate momentum at end of trajectory to make the proposal symmetric
  p = -p
  # Evaluate potential and kinetic energies at start and end of trajectory
  current_U = U(current_q)
  current_K = sum(current_p^2) / 2
  proposed_U = U(q)
  proposed_K = sum(p^2) / 2
  # Accept or reject the state at end of trajectory, returning either
  # the position at the end of the trajectory or the initial position
  if (runif(1) < exp(current_U-proposed_U+current_K-proposed_K))
  {
    return (q)  # accept
  }
  else {
    return (current_q)  # reject
  }
}
```

Although the HMC works well theory, it requires tuning a couple of parameters for practical use. The choice of step size ϵ and the number of updates L and the mass M will all affect the final performance. Detailed discussion about the tuning can be found in Section 5.4 of https://arxiv.org/pdf/1206.1901.pdf.

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 $\{X_n : n = 0, 1, 2, \cdots\}$ is called a *Markov chain* if for every $x_0, x_1, \cdots, x_{n-2}, i, j \in S$ and $n \geq 0$,

$$P(X_n = i|X_{n-1} = j, \cdots, X_0 = x_0) = P(X_n = i|X_{n-1} = j)$$

whenever both sides are well-defined. The Markov chain has a one-step memory.

If the distribution function $P(X_n = x_n|X_{n-1} = x_{n-1}) = p_{ij}$ is independent of n, we called $\{X_n\}$ a **homogeneous Markov chain**. Otherwise we called it an inhomogeneous Markov chain. For a homogeneous Markov chain,

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

for every i, j. Note that sometimes people write $p_{ij} = p_{i\to j}$, where $p_{i\to j}$ 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 $\{p_{ij} : i, j = 1, \cdots, s\}$ forms an $s \times s$ matrix $\mathbf{P} = \{p_{ij}\}$. $\mathbf{P}$ is called the **transition (probability) matrix (t.p.m)**. The property of homogeneous Markov chain implies that

$$\mathbf{P} \geq 0, \quad \mathbf{P1}_s = \mathbf{1}_s, \tag{11.3}$$

where $\mathbf{1}_s = (1, 1, 1, 1, \cdots, 1)^T$ is the vector of 1's. Note that any matrix satisfying equation (11.3) 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

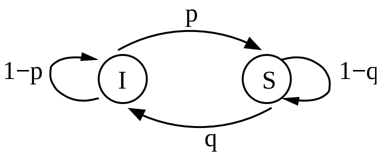
$$X_n = \begin{cases} I & \text{if infected} \\ S & \text{if susceptible.} \end{cases}$$

We would like to construct a stochastic model for the sequence $\{X_n : n = 1, 2, ...\}$. One possibility is to assume that the X_n's are independent with $P(X_n = I) = 1 - P(X_n = S) = \alpha$. 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:

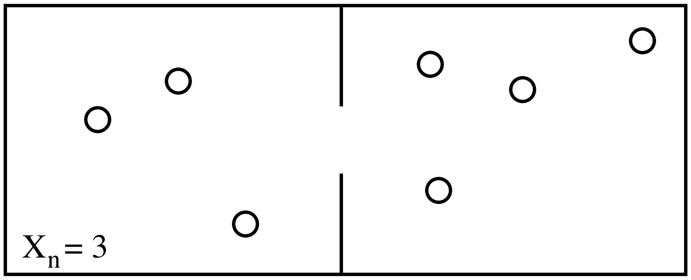
$$\mathbf{P} = \begin{array}{c c} \nearrow & \begin{array}{cc} I & S \end{array} \\ \begin{array}{c} I \\ S \end{array} & \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \end{array}$$

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.



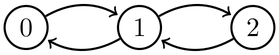
Transition probabilities of the Markov chain are:

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

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.

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{bmatrix}$$

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:

$$\begin{aligned}
&P(X_n = i_n, X_{n-1} = i_{n-1}, \cdots, X_1 = i_1, X_0 = i_0) \\
&= P(X_n = i_n | X_{n-1} = i_{n-1}, \cdots, X_0 = i_0)\, P(X_{n-1} = i_{n-1}, \cdots, X_0 = i_0) \\
&= p_{i_{n-1},i_n} \times P(X_{n-1} = i_{n-1} | X_{n-2} = i_{n-2}, \ldots, X_0 = i_0) \times P(X_{n-2} = i_{n-2}, \ldots, X_0 = i_0) \\
&= \cdots \\
&= p_{0i_0} p_{i_0,i_1} p_{i_1,i_2} \cdots p_{i_{n-2},i_{n-1}} p_{i_{n-1},i_n}
\end{aligned}$$

where $p_0 = (p_{01}, p_{02}, \dots)^T$ is the distribution of X_0, called the *initial distribution* of $\{X_n\}$. 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

$$P(X_{m+n} = j | X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = P(X_{m+n} = j | X_n = i). \tag{11.4}$$

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

$$\begin{aligned}
P(X_3 = i_3 | X_1 = i_1, X_0 = i_0) &= \sum_{i_2} P(X_3 = i_3, X_2 = i_2 | X_1 = i_1, X_0 = i_0) \\
&= \sum_{i_2} P(X_3 = i_3 | X_2 = i_2, X_1 = i_1, X_0 = i_0) P(X_2 = i_2 | X_1 = i_1, X_0 = i_0) \\
&\overset{!}{=} \sum_{i_2} P(X_3 = i_3 | X_2 = i_2, X_1 = i_1) P(X_2 = i_2 | X_1 = i_1) \\
&= \sum_{i_2} P(X_3 = i_3, X_2 = i_2 | X_1 = i_1) \\
&= P(X_3 = i_3 | X_1 = i_1).
\end{aligned}$$

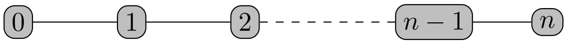
To argue the equality '$\overset{!}{=}$', observe that

$$P(X_3 = i_3 | X_2 = i_2, X_1 = i_1, X_0 = i_0) = P(X_3 = i_3 | X_2 = i_2).$$

But we also have that

$$\begin{aligned}
P(X_3 = i_3 | X_2 = i_1, X_1 = i_1) &= \sum_{i_0} P(X_3 = i_3, X_0 = i_0 | X_2 = i_2, X_1 = i_1) \\
&= \sum_{i_0} P(X_3 = i_3 | X_2 = i_2, X_1 = i_1, X_0 = i_0) P(X_0 = i_0 | X_2 = i_2, X_1 = i_1) \\
&= P(X_3 = i_3 | X_2 = i_2) \sum_{i_0} P(X_0 = i_0 | X_2 = i_2, X_1 = i_1) \\
&= P(X_3 = i_3 | X_2 = i_2).
\end{aligned}$$

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

$$\begin{aligned}
P(X_0 = i_0, X_2 = i_2|X_1 = i_1) &= \frac{P(X_0 = i_0, X_1 = i_1, X_2 = i_2)}{P(X_1 = i_1)} \\
&= \frac{P(X_2 = i_2|X_1 = i_1, X_0 = i_0)P(X_1 = i_1, X_0 = i_0)}{P(X_1 = i_1)} \\
&= P(X_2 = i_2|X_1 = i_1)\frac{P(X_1 = i_1, X_0 = i_0)}{P(X_1 = i_1)} \\
&= P(X_2 = i_2|X_1 = i_1)P(X_0 = i_0|X_1 = i_1)
\end{aligned}$$

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

$$p_{ij}^{(n)} = P(X_n = j|X_0 = i).$$

The n-step transition probability is time invariant.

Lemma 11.6 *Let $\{X_n\}$ be a homogeneous Markov chain and let $p_{ij}^{(n)}$ be the n-step transition probability. Then for any $k = 0, 1, 2, \cdots$,*

$$P(X_{n+k} = j|X_k = i) = p_{ij}^{(n)}.$$

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

Lemma 11.7 *Let $\{X_n\}$ be a homogeneous Markov chain and let $p_{ij}^{(n)}$ be the n-step transition probability. Then for any $n, m = 0, 1, 2, \cdots$*

$$p_{ij}^{(n+m)} = \sum_{k\in S} p_{ik}^{(n)} p_{kj}^{(m)}. \tag{11.5}$$

The Chapman-Kolmogorov Equation (equation (11.5)) also implies

$$\begin{aligned}
\textit{Forward equation}: p_{ij}^{(n+1)} &= \sum_k p_{ik}^{(n)} p_{kj}, \text{ for } n = 1, 2, \ldots \text{ and} \\
\textit{Backward equation}: p_{ij}^{(n+1)} &= \sum_k p_{ik} p_{kj}^{(n)}, \text{ for } n = 1, 2, \ldots.
\end{aligned}$$

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 $p_{ij}^{(n)}$'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 $p_{ij}^{(n)}$'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 ($p_{jj} = 1$).

If we collect the n-step transition probabilities into the matrix $\mathbf{P}^{(n)} = \{p_{ij}^{(n)}\}$, 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 $p_0 = (p_{01}, p_{02}, \cdots, p_{0s})^T$. Let $p_n = (p_{n1}, \cdots, p_{ns})^T$ be the marginal distribution of X_n, i.e., $p_{nj} = P(X_n = j)$. Then

$$\begin{aligned} p_{nj} = P(X_n = j) &= \sum_i P(X_n = j, X_0 = i) \\ &= \sum_i P(X_n = j | X_0 = i) P(X_0 = i) \\ &= \sum_i p_{0i} p_{ij}^{(n)}. \end{aligned}$$

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

$$\mathbf{P} = \begin{array}{c c} \nearrow & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \end{array}$$

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., $P(X_0 = 0) = 1-\alpha$. Moreover, assume that $\beta = 1-\alpha$ so the distribution of X_1 will be

$$\begin{aligned} p_1^T &= p_0^T \mathbf{P} = (1-\alpha, \alpha) \begin{bmatrix} 1-\alpha & \alpha \\ 1-\alpha & \alpha \end{bmatrix} \\ &= [(1-\alpha)^2 + \alpha(1-\alpha), \alpha(1-\alpha) + \alpha^2] = (1-\alpha, \alpha) = p_0^T. \end{aligned}$$

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, $P(X_n = 0) = 1-\alpha$ and $P(X_n = 1) = \alpha$ for all $n = 1, 2, 3, \cdots$.

A more interesting fact is the joint distribution of $X_0, X_1, \cdots, X_n$:

$$\begin{aligned} P(X_0 = i_0, X_1 = i_1, \cdots X_n = i_n) &= p_{0i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n} \\ &= \alpha^{i_0}(1-\alpha)^{1-i_0} \alpha^{i_1} (1-\alpha)^{1-i_\alpha^{i_0}(1-\alpha)^{1-i_0}} \cdots \alpha^{i_n}(1-\alpha)^{1-i_n}, \end{aligned}$$

which is the joint PMF of IID random Bernoulli variables with parameter α. 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: $i \sim j$, $j \sim k \Rightarrow i \sim k$.

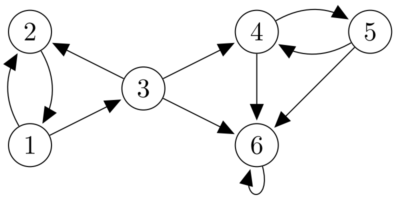
For a set $\mathcal{S}$ and $a \in \mathcal{S}$, $\{s \in \mathcal{S} : s \sim a\}$ is called an *equivalence class*. Equivalence relations will allow us to split Markov chain state spaces into equivalence classes.

State j is *accessible* from state i ($i \to j$) if there exists $m \geq 0$ such that $p_{ij}^{(m)} > 0$. 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 11.8 *Communication of states is an equivalence relation.*

A set of states $\mathcal{C}$ is *closed* if $\sum_{j \in \mathcal{C}} p_{ij} = 1$ 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 $\{X_n\}$ is called **irreducible** if it has only one communication class, i.e., for all i and j , $i \leftrightarrow j$. For state i, $d_i = \gcd\{n \geq 1 : p_{ii}^{(n)} > 0\}$ is called its **period**, where gcd = greatest common divisor and $d_i = +\infty$ if $p_{ii}^{(n)} = 0$ for all $n \geq 1$.

Example 7. Consider the example with state space $S = \{0, 1, 2, ...\}$ and X_n such that

$$P(X_{n+1} = i | X_n = 0) = \begin{cases} p & \text{if } i = 1, \\ 1 - p & \text{if } i = 0, \\ 0 & \text{otherwise.} \end{cases}$$

and for $j \neq 0$

$$P(X_{n+1} = i | X_n = j) = \begin{cases} p & \text{if } i = j + 1, \\ 1 - p & \text{if } i = j - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then, $d_2 = \gcd\{2, 4, 5, 6, ...\} = 1$ 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

$$P(X_{n+1} = i - 1|X_n = i) = q \text{ and } P(X_{n+1} = i + 1|X_n = i) = p$$

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 $p_{00}^{(2n+1)} = 0$ for all n, i.e.

$$d_0 = \gcd\{n \geq 1 : p_{00}^{(n)} > 0\} = \gcd\{2, 4, 6, \dots\} = 2.$$

Proposition 11.9 *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 $m < n < k$, $X_m \perp X_k|X_n$. Let $\{X_n\}$ be a Markov chain with state space $S = \{1, 2, 3\}$ and consider a random time

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

Then the probability

$$P(X_k = 3|X_n = 1, X_m = 2) = P(X_{T+1} = 3|X_T = 1, X_{T-1} = 2) = 1 \neq P(X_k = 3|X_n = 1)$$

because when $X_m = X_{n-1} = 3$, $P(X_k = 2|X_n = 1, X_m = 3) = P(X_{T+1} = 2|X_T = 1, X_{T-1} = 3) = 1$ so $P(X_k = 3|X_n = 1, X_m = 3) = 0$. 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 $T_i = \inf\{n \geq 1 : X_n = i\}$ is a stopping time because $\{T_i = m\} = \{X_1 \neq i, \cdots, X_{m-1} \neq i, X_m = i\}$. T_i is interpreted as the first time the chain returns to state i.
- Successive Returns. Let τ_k be the time of the k-th return to state i (note that $\tau_1 = T_i$). Then τ_k is a stopping time because
$$\{\tau_k = m\} = \left\{\sum_{n=1}^{m} I(X_n = i) = k, X_m = i\right\}.$$

- Counterexample – non-stopping time: Let $\tau = \inf\{n \geq 1 : X_{n+1} = i\}$ is not a stopping time because when the time arrives at m, $\{\tau = m\} = \{X_1 \neq i, \cdots, X_m \neq i, X_{m+1} = i\}$ 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 11.10 (Strong Markov Property) *Let $\{X_n\}$ be a homogeneous Markov chain with a transition probability matrix $\mathbf{P} = \{p_{ij}\}$ and let τ be a stopping time with respect to $\{X_n\}$. Then for any integer k,*

$$P(X_{\tau+k} = j | X_\tau = i, 0 \leq \ell < \tau, X_\ell = i_\ell) = P(X_k = j | X_0 = i) = p_{ij}^{(k)}$$

and

$$P(X_{\tau+k} = j | X_\tau = i) = P(X_k = j | X_0 = i) = p_{ij}^{(k)}.$$

A.5 Stationary distribution

It is often of great interest to study the limiting behavior of a Markov chain X_n when $n \to \infty$. 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.

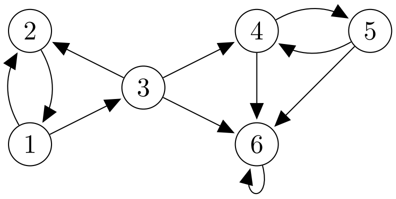
A.5.1 Return Times

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

$$P(\cdot | X_0 = i) = P_i(\cdot), \quad \mathbb{E}(\cdot | X_0 = i) = \mathbb{E}_i(\cdot).$$

Note that the quantity N_i may equal to ∞. 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 $T_i = \inf\{n \geq 1 : X_n = i\}$ be the return time. Then the following events can be defined using either T_i or N_i:

$$\{T_i = \infty\} = \{N_i = 0\}, \quad \{T_i < \infty\} = \{N_i > 0\}.$$

These are useful later.

We then define $f_{ji} = P_j(T_i < \infty) = P_j(N_i > 0)$ to be the probability of reaching state i in a finite number of time when the chain starts at state j. Note that because $P_j(T_i = \infty) + P_j(T_i < \infty) = 1$, we have $f_{ii} = P_i(T_i < \infty)$ and $P_i(T_i = \infty) = 1 - f_{ii}$.

Proposition 11.11

$$P_j(N_i = r) = \begin{cases} f_{ji} f_{ii}^{r-1}(1 - f_{ii}) & \textit{if } r \geq 1 \\ 1 - f_{ji} & \textit{if } r = 0 \end{cases}.$$

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

$$P_i(N_i = r) = f_{ii}(1 - f_{ii}), \quad P_i(N_r > r) = f_{ii}^{r+1},$$

where $f_{ii} = P_i(T_i < \infty)$.

We have seen many situations that T_i and N_i are closely related. Here is another result about their relationship.

Corollary 11.12

$$P_i(N_i = \infty) = 1 \Leftrightarrow P_i(T_i < \infty) = 1$$

and

$$P_i(T_i < \infty) < 1 \Leftrightarrow P_i(N_i = \infty) = 0 \Leftrightarrow \mathbb{E}_i(N_i) < \infty.$$

Corollary 11.12 links the finiteness of T_i and N_i and also relates it to the expectation. With the following formula of expectation, Corollary 11.12 will be very useful:

$$\mathbb{E}(X) = \sum_{t=1}^{\infty} P(X \geq t), \tag{11.6}$$

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

$$\begin{cases} \textbf{recurrent/persistent}, & \text{if } P_i(T_i < \infty) = f_{ii} = 1 \\ \textbf{transient}, & \text{otherwise.} \end{cases}$$

Furthermore, a recurrent state is called

$$\begin{cases} \textbf{positive recurrent}, & \text{if } \mathbb{E}_i(T_i) < \infty \\ \textbf{null recurrent}, & \text{otherwise.} \end{cases}$$

Note that: either $P_i(N_i = \infty) = 0$ or $P_i(N_i = \infty) = 1$, with nothing in between (if $f_{ii} < 1$, then $P_i(N_i = \infty) = 0$; if $f_{ii} = 1$, then $P_i(N_i = \infty) = 1$). This, together with Corollary 11.12, implies that $\mathbb{E}_i(N_i) = \infty \iff P_i(N_i = \infty) = 1$.

Note that:

$$f_{ii} = P_i(T_i < \infty) = 1 \iff P_i(N_i = \infty) = 1.$$

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

Proposition 11.13 *State i is recurrent* $\iff \sum_{n=1}^{\infty} p_{ii}^{(n)} = \infty$.

Proposition 11.14 *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:

$$P_1(T_1 < \infty) = 1 - P_1(T_1 = \infty) \leq 1 - q < 1 \text{ if } q \in (0, 1).$$

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

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

Let's study recurrence of state 0. We know that $p_{00}^{(2n+1)} = 0$ for all $n \geq 0$ and that, conditional on $X_0 = 0$, $X_{2n} =_d \xi_1 + \cdots + \xi_{2n}$, where $\xi_1, \ldots, \xi_n$ are i.i.d. with $P(\xi_i = 1) = 1 - P(\xi = -1) = p$. Hence,

$$p_{00}^{(2n)} = P(X_{2n} = 0 | X_0 = 0) = \binom{2n}{n} p^n q^n.$$

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

$$\lim_{n \to \infty} \frac{n!}{n^{n+\frac{1}{2}} e^{-n} \sqrt{2\pi}} = 1.$$

Therefore,

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

We deduce that

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

(Ratio Test: Let $\sum_{n=1}^{\infty} a_n$ be a series which satisfies $\lim_{n\to\infty} |\frac{a_{n+1}}{a_n}| = k$. 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., $x_i = \sum_j x_j p_{ji}$ for each i.

A probability vector π on a Markov chain state space is called a **stationary distribution** of a stochastic matrix $\mathbf{P}$ if $\pi^T\mathbf{P} = \pi^T$, i.e., $\pi_i = \sum_j \pi_j p_{ji}$ 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

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix} \qquad 0 < p < 1, \quad 0 < q < 1.$$

The global balance equations:

$$\begin{bmatrix}\pi_0, \pi_1\end{bmatrix} \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix} = \begin{bmatrix}\pi_0, \pi_1\end{bmatrix} \text{ or}$$

$$\begin{cases} (1-p)\pi_0 + q\pi_1 &= \pi_0 \\ p\pi_0 + (1-q)\pi_1 &= \pi_1 \end{cases} \quad \Rightarrow \quad p\pi_0 = q\pi_1 \quad \Rightarrow \quad \pi_0 = \frac{q}{p}\pi_1.$$

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

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 \\ 0 & q & 0 & p & 0 \\ 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

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 11.15 (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* $\pi^T = [\pi_1, \pi_2, \dots]$ *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 $\mathbf{y}^T = [1, 1, \dots]$ for any p and q since

$$\mathbf{P} = \begin{bmatrix} \ddots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots \\ \cdots & q & 0 & p & \cdots & \cdots & \cdots \\ \cdots & \cdots & q & 0 & p & \cdots & \cdots \\ \cdots & \cdots & \cdots & q & 0 & p & \cdots \\ \cdots & \cdots & \cdots & \cdots & \ddots & \ddots & \ddots \end{bmatrix}$$

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 11.16 *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 11.17 *Let* $\{X_n\}$ *be an irreducible homogeneous positive recurrent Markov chain. Then*

$$\pi_i = \frac{1}{\mathbb{E}_i(T_i)},$$

where $\pi = (\pi_1, \cdots, \pi_s)$ *is the stationary distribution of* $\{X_n\}$ *and* $T_i = \inf\{n \geq 1 : X_n = i\}$ *is the return time to state* i*.*