6 Monte Carlo Sampling Techniques

6.1 Motivations

Drawing independent identically distributed (i.i.d.) samples from a probability distribution:

1. Numerical approximation of integrals. Replace an integral by an expectation over a distribution. Then, estimate the expectation by sampling from the distribution and averaging the integrand.
2. Integrating over Bayesian posterior distributions.
3. Computational physics: e.g. sampling from the Boltzmann distribution.
4. Numerical optimization.

6.2 Transformation of Random Variables

6.2.1 Jacobian

Recall from calculus that the volume element should be invariant under coordinate transformations. Let $X_1, \ldots, X_n$ be continuous random variables and $Y_1 = Y_1(\mathbf{X}), \ldots, Y_n = Y_n(\mathbf{X})$ one-to-one differentiable transformations of $\mathbf{X} = (X_1, \ldots, X_n)$ with differentiable inverses. Because probability density computes local volume, imposing the invariance implies that

$$\begin{aligned} p_Y(y_1, \ldots, y_n) dy_1 \cdots dy_n &= p_X(x_1, \ldots, x_n) dx_1 \cdots dx_n \\ &= p_X(x_1(\mathbf{y}), \ldots, x_n(\mathbf{y})) |J(y_1, \ldots, y_n; x_1, \ldots, x_n)|^{-1} dy_1 \cdots dy_n \end{aligned}$$

where the Jacobian J is defined as

$$J(y_1, \ldots, y_n; x_1, \ldots, x_n) = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{vmatrix}.$$

Hence,

$$p_Y(y_1, \ldots, y_n) = p_X(x_1(\mathbf{y}), \ldots, x_n(\mathbf{y})) |J(y_1, \ldots, y_n; x_1, \ldots, x_n)|^{-1}.$$

You can also use

$$p_Y(y_1, \ldots, y_n) = p_X(x_1(\mathbf{y}), \ldots, x_n(\mathbf{y})) |J(x_1, \ldots, x_n; y_1, \ldots, y_n)|,$$

whichever is easier to compute.

Example 6.1 (Scaling)**.** *Let $Y = \alpha X$ for some constant $\alpha > 0$. Then,*

$$p_Y(y) = p_X(y/\alpha) \left| \frac{dx}{dy} \right| = \frac{p_X(y/\alpha)}{\alpha}.$$

So, if X is uniform on $[0,1]$, and $Y = \alpha X$, then Y is uniform on $[0,\alpha]$ with density $p_Y(y) = 1/\alpha$.

Example 6.2 (Sum of Independent Random Variables)**.** *Let X and Y be independent random variables with densities p_X and p_Y, respectively. Let $Z = X + Y$ and $W = Y$. Then, $|J(X,Y;Z,W)| = 1$, and*

$$p_{Z,W}(z,w) = p_{X,Y}(x(z,w), y(z,w)) = p_X(x(z,w))p_Y(y(z,w)),$$

where the last equality follows from the fact that X and Y are independent. But, $x = z - w$ and $y = w$, so

$$p_{Z,W}(z,w) = p_X(z-w)p_Y(w).$$

The marginal density of Z is obtained by integrating out W, i.e.

$$p_Z(z) = \int p_{Z,W}(z,w)dw = \int p_X(z-w)p_Y(w)dw = (p_X * p_Y)(z),$$

where $*$ is the convolution operator.

6.2.2 Cumulative Distribution Functions as Uniform Random Variables

Let $F_X(x)$ be the cumulative distribution of a random variable X, and $p(x)$ its density. Then, define a new random variable Y as

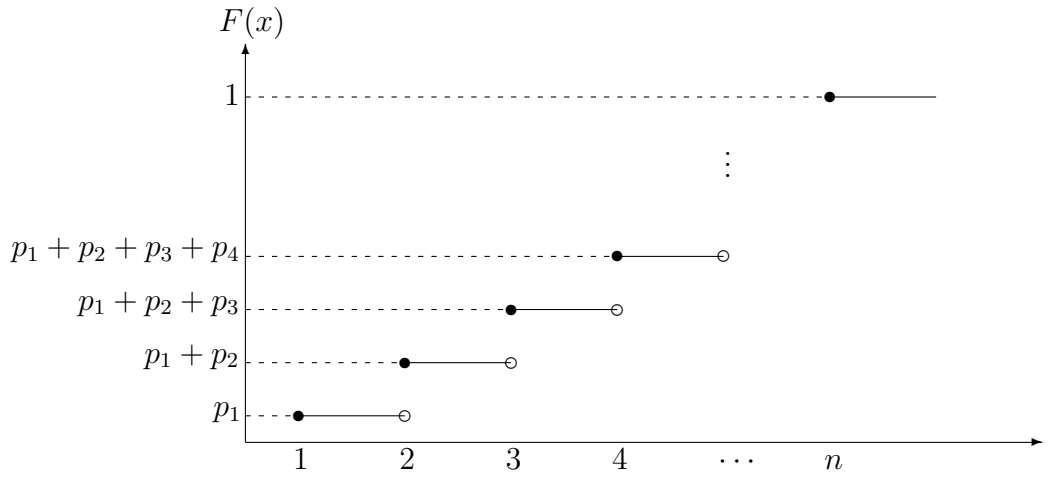
$$Y = F_X(X) = \int_{-\infty}^{X} p(t)\,dt.$$

Applying the coordinate transformation, the density of Y is

$$p(y) = p(x(y))J(x;y) = p(x(y))\left(\frac{dy}{dx}\right)^{-1} = p(x(y))p(x(y))^{-1} = 1.$$

That is, Y is uniformly distributed on $[0,1]$. Equivalently, the one-sided p-values are uniformly distributed.

6.2.3 Monte Carlo Sampling by Inverting Cumulative Distributions



Sampling a discrete random $X \in \{1, 2, \ldots, n\}$ which has a finite number of possible outcomes is easy: given $p_i = P(X = i)$, partition the unit interval $[0, 1]$ as

$$[0, 1] = [0, p_1) \cup [p_1, p_1 + p_2) \cup \cdots [p_1 + p_2 + \cdots + p_{n-1}, 1],$$

so that the width of i-th interval is the probability p_i. Now, use a computer to generate pseudo-random numbers $y_1, y_2, \ldots, y_m$ which are almost uniformly distributed in the interval $[0, 1]$, and assign $X_j = i$ if $y_j \in [\sum_{k=1}^{i-1} p_k, \sum_{k=1}^{i} p_k)$.

From Section 6.2.2, we know that for a continuous random variable X, $Y = F_X(X)$ is uniformly distributed. So, in principle, to sample X, we just need to sample Y uniformly on $[0, 1]$ and then invert the relation $y = F_X(x)$ in order to obtain $x = F_X^{-1}(y)$. Note that because F_X is a monotonically increasing function, a unique inverse exists.

REMARK 6.1. *In general, if we are able to sample a random variable Y, then we can also sample the random variable $Z = f(Y)$ by evaluating $f(y_i)$ for each sample y_i of Y.*

Example 6.3. *The density of an exponential distribution is $p(x) = \gamma e^{-\gamma x}$, for $x \geq 0$ and a constant $\gamma > 0$. Its cumulative distribution is $y = F(x) = (1 - e^{-\gamma x})$, which implies that $x = -\log(1 - y)/\gamma$. Thus, to sample a random variable that is exponetially distributed with parameter γ, first sample a uniformly distributed y, and then take the transform $x = -\log(1 - y)/\gamma$.*

In reality, inverting F_X directly is not so easy for most random variables, and we need to take alternative approaches.

Example 6.4. *Let X be distributed as standard normal, i.e.*

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

The cumulative distribution of X is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \, dx = \frac{1}{2} + \frac{1}{2} erf\left(\frac{x}{\sqrt{2}}\right),$$

where the error function erf(z) is defined as

$$erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt.$$

The cumulative function $\Phi(x)$ is difficult to invert, so we will use a trick to sample X: the trick is to sample two independent standard normal random variables X and Y simultaneously. The joint distribution of X and Y is

$$p(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}.$$

Now define $R = (x^2 + y^2)/2$, so that $x = \sqrt{2R}\cos\theta$ and $y = \sqrt{2R}\sin\theta$, for some $\theta \in [0, 2\pi)$. In this new coordinate system, we have

$$p(R, \theta) = \frac{e^{-R}}{2\pi}.$$

Hence, R is exponentially distributed with rate 1 and θ is uniformly distributed between 0 and 2π. We already know how to sample R from Example 6.3. For each pair (R_i, θ_i) of samples, $i = 1, \ldots, n$, compute $x_i = \sqrt{2R_i}\cos\theta_i$ and $y_i = \sqrt{2R_i}\sin\theta_i$. Finally, $x_1, x_2, \ldots, x_n$ provide n samples of X, and $y_1, y_2, \ldots, y_n$ provide n samples of Y.

Example 6.5. *The joint density $p(x, y)$ can be written as $p(x, y) = p_{X|Y}(x|y)p_Y(y)$. To sample from $p(x, y)$, one can first sample y from $p_Y(y)$ and then sample x from $p_{X|Y}(x|y)$.*

6.3 Rejection Sampling

Suppose you have a random variable X with a density function $p_X(x)$, which is difficult to sample. The main idea behind rejection sampling is to sample from another distribution $q(x)$ which is easy to sample and then reject the samples with some defined probability. Amazingly, this methods works even if you know p_X only up to a multiplicative constant; i.e. $p_X(x) = g(x)/c$, where $c = \int_{-\infty}^{\infty} g(x)\, dx$ is a normalization constant that may be difficult to compute. Let us now be more precise. Suppose there exists $M \geq 0$, such that

$$g(x) \leq Mq(x). \tag{6.1}$$

Then, the rejection sampling steps for sampling X are

1. Sample x from $q(x)$.
2. Compute $r = \frac{g(x)}{Mq(x)}$. (Note that $r \leq 1$).
3. Sample u uniformly in $[0, 1]$.
4. If $u \leq r$, then accept x; otherwise, go to step 1.

To see why this methods works, note that

$$p(x|\text{accept}) = \frac{p(x,\text{accept})}{P(\text{accept})} = \frac{p(\text{accept}|x)q(x)}{P(\text{accept})} = \frac{\frac{g(x)}{Mq(x)}q(x)}{\int_{-\infty}^{\infty}\frac{g(x)}{Mq(x)}q(x)\,dx} = \frac{g(x)}{c} = p_X(x).$$

Thus, the set of all accepted values are distributed as p_X. The acceptance probability is

$$p(\text{accept}) = \int_{-\infty}^{\infty} p(x,\text{accept})dx = \int_{-\infty}^{\infty} p(\text{accept}|x)q(x)dx = \int_{-\infty}^{\infty}\frac{g(x)}{Mq(x)}q(x)dx = \frac{c}{M}.$$

Hence, the larger the M, the lower the acceptance probability. We thus need to choose $q(x)$ that is easy to sample from and also has small M in (6.1).

Example 6.6 (Truncated Distribution)**.** *Let X be a random variable with density $p(x)$. A truncated distribution of X is a conditional distribution $p(x|X \in [a,b])$ defined as*

$$p(x|X \in [a,b]) = \begin{cases} p(x)/P(X \in [a,b]), & \text{for } x \in [a,b] \\ 0, & \text{otherwise.} \end{cases}$$

In (6.1), we choose $g(x) = p(x)I(x \in [a,b])$, $q(x) = p(x)$ and $M = 1$. Thus, sample from $p(x)$, and accept x iff $x \in [a,b]$. The acceptance rate of this approach is $P(X \in [a,b])$. This method is efficient if $P(X \in [a,b])$ is not much smaller than 1.

Example 6.7 (Truncated Normal)**.** *Let X be standard normal. We would like to sample from $p(x|X \geq a)$. If $a \leq 0$, then we can use the approach described in Example (6.6). If $a > 1$, then the approach can quickly become wasteful. In that case, it is better to use a translated exponential distribution as the proposal distribution. That is, choose $q(x) = \lambda e^{-\lambda(x-a)}$, for $x \geq a$, and $q(x) = 0$, for $x < a$. It can be shown that the optimal choice of λ is*

$$\lambda^* = \frac{a + \sqrt{a^2+4}}{2} \tag{6.2}$$

and the acceptance rate is

$$\sqrt{2\pi}\,(1 - \Phi(a))\,\lambda^*\,e^{\lambda^* a - {\lambda^*}^2/2},$$

where $\Phi(a)$ is the cumulative distribution of a standard normal random variable.

6.3.1 How to Sample from a Bayesian Posterior Distribution using Rejection Sampling

Recall that the posterior distribution $P(\theta|\mathbf{x})$ of parameter θ given data $\mathbf{x}$ is:

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x}|\theta)P(\theta)}{P(\mathbf{x})},$$

where $P(\mathbf{x}|\theta)$ is the likelihood and $P(\theta)$ is the prior distribution. When the prior distribution $P(\theta)$ is simple enough that we are able to sample from it using standard methods, we can

use rejection sampling to sample from the posterior distribution. Letting $L = \max_\theta P(\mathbf{x}|\theta)$, we see that

$$P(\theta|\mathbf{x}) \leq \frac{L}{P(\mathbf{x})} P(\theta),$$

which is in the form of (6.1) with $M = L/P(\mathbf{X})$. Hence, we can sample from the prior $P(\theta)$ and accept with probability

$$r = \frac{P(\theta|\mathbf{x})}{\frac{L}{P(\mathbf{X})} P(\theta)} = \frac{P(\mathbf{x}|\theta)}{L},$$

which is just the likelihood normalized by its maximum. The efficiency of this sampling scheme is $P(\mathbf{x})/L$.

6.3.2 Sampling Nonhomogeneous Poisson Processes

Another form of rejection sampling is the following scheme based on the rates of Poisson processes. Let $\Omega \in \mathbb{R}^d$ be a non-empty set with measure μ. Intuitively, given a measurable set $A \subset \Omega$, $\mu(A)$ gives the area of A. The Poisson process is then a point generating process such that

1. The number $N(A)$ of random points occurring in A is Poisson distributed with mean $\mu(A)$:
$$P(N(A) = n) = \frac{\mu(A)^n e^{-\mu(A)}}{n!}.$$
2. For any disjoint measurable sets $A_1, \ldots, A_k \subset \Omega$, the numbers $N(A_1), \ldots, N(A_k)$ are independent.

REMARK 6.2. *We can think of the Poisson process as a random counting process denoted by the random counting function N.*

If there exists a non-negative function λ on Ω, such that

$$\mu(A) = \int_A \lambda(x)\, dx,$$

then λ is called the rate function.

Theorem 6.1. *Conditioned on $N(A) = n$, the n points are i.i.d. samples from a probability distribution with density*

$$p(x) = \begin{cases} \frac{\lambda(x)}{\mu(A)}, & x \in A \\ 0, & x \notin A \end{cases}.$$

Proof. Let $B_1, \ldots, B_n \subset A$ denote disjoint balls of sufficiently small radius $\epsilon > 0$. Then,

$$\begin{aligned} P(N(B_1) = 1, \ldots, N(B_n) = 1|N(A) = n) &= \frac{\left(\prod_{i=1}^n \mu(B_i) e^{-\mu(B_i)}\right) e^{-(\mu(A) - \sum_{i=1}^n \mu(B_i))}}{\mu(A)^n e^{-\mu(A)}/n!} \\ &= n! \prod_{i=1}^n \frac{\mu(B_i)}{\mu(A)}, \end{aligned}$$

where the combinatorial factor $n!$ accounts for the indistinguishability of the n points in A. □

In the special case of $\Omega = \mathbb{R}_+$, let T_n denote the time of the n-th point. The waiting time distribution between two Poisson events is

$$P(T_{n+1} - T_n > t) = P(N([T_n, T_n + t]) = 0) = \exp\left(-\int_{T_n}^{T_n+t} \lambda(\tau)\, d\tau\right),$$

which is a nonhomogeneous exponential distribution with density

$$p_{T_{n+1}-T_n}(t) = \begin{cases} \lambda(t + T_n) \exp\left(-\int_{T_n}^{T_n+t} \lambda(\tau)\, d\tau\right), & t \geq 0 \\ 0 & t < 0 \end{cases}.$$

Sampling from this 1-dimensional (1D) Poisson process is equivalent to sampling from a 2D Poisson process with constant rate 1 in the region under the curve of λ and then marginalizing the y variable. This observation provides a efficient way to sample from a 1D nonhomogeneous Poisson process:

Theorem 6.2. *Let N be a 1D Poisson process with rate $\lambda(t)$. Let N' be a second Poisson process with rate $\lambda'(t) \geq \lambda(t)$. Then, the thinning of N' obtained by rejecting the random points of N' with probability $1 - \lambda(x)/\lambda'(x)$ is N.*

Proof. Thinking of the random points of N' as arising from the 2D Poisson with rate 1 under the curve of λ', the acceptance probability $\lambda(x)/\lambda'(x)$ implies that a random point will be accepted iff it lies under the curve of λ (via Theorem 6.1.). □

REMARK 6.3. *In particular, we can choose $\lambda'(x)$ to be a constant function equal to the maximum value $\lambda_{\max} = \max_x \lambda(x)$. Thus, a nonhomogeneous Poisson process can be simulated by thinning a homogeneous Poisson process.*

6.4 Importance Sampling and Variance Reduction of Monte Carlo Estimates

Suppose we want to numerically estimate the tail integral

$$f(z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

for $z > 2$. We can rewrite $f(z)$ as

$$f(z) = \int_{-\infty}^{\infty} I(x \geq z) p(x) dx,$$

where $I(x \geq z)$ is the indicator function for the condition $x \geq z$ and $p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the standard normal density. By sampling N i.i.d. points x_i from the standard normal

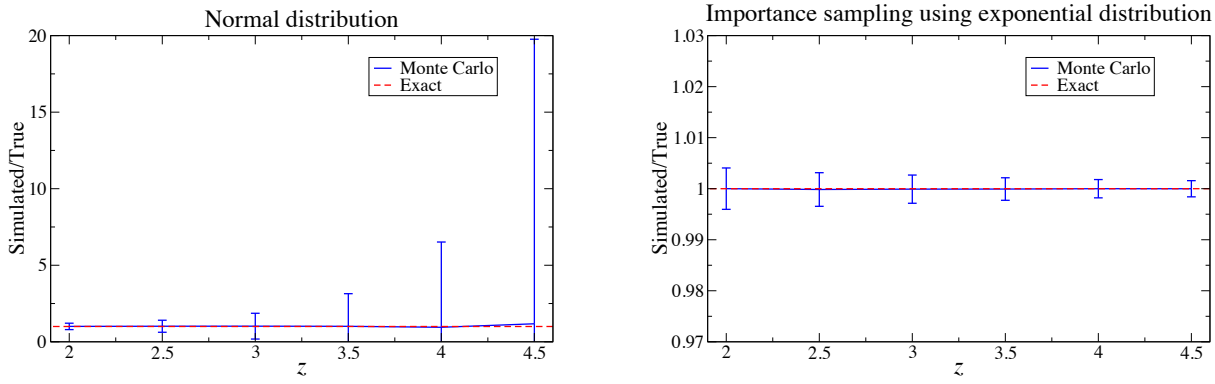


Figure 6.1: Monte Carlo estimation of the tail probability $\int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$ by sampling from the standard normal distribution (left) or shifted exponential distribution (right).

distribution, we obtain the estimate

$$\hat{f}(z) \equiv \frac{1}{N} \sum_{i=1}^{N} I(x_i \geq z) \approx f(z).$$

The expected value of $\hat{f}(z)$ is

$$E[\hat{f}(z)] = \frac{1}{N} \sum_{i=1}^{N} E_{X_i}[I(X_i \geq z)] = f(z).$$

Hence, $\hat{f}(z)$ is an unbiased estimator of $f(z)$. However, the variance of $\hat{f}(z)$ is

$$\text{var}[\hat{f}(z)] = \frac{f(z)(1 - f(z))}{N},$$

so that the relative error in the estimation is

$$\frac{\text{sd}[\hat{f}(z)]}{f(z)} = \sqrt{\frac{1}{N}} \sqrt{\frac{1 - f(z)}{f(z)}}.$$

For $z > 2$, $f(z)$ is a small number and $1 - f(z) \approx 1$, so the relative error can be large for large z, as shown in Figure 6.1. In fact, we need at least as many as $1/f(z)$ samples to have a small relative error. For $z = 5$, this required number is greater than 3 million. This kind of naive estimator is thus not very useful for practical applications.

A major problem associated with the current estimator is that obtaining a sample $x_i > z$, for large z, is a rare event; that is, a typical sample will satisfy $x_i < z$ and will not contribute to the sum. Importance sampling is a technique designed to generate random samples according to a new distribution with density $q(x)$, such that most samples will be concentrated around the large absolute values of the integrand. For this purpose, let us

rewrite $f(z)$ as

$$f(z) = \int_z^\infty I(x \geq z) \frac{p(x)}{q(x)} q(x)\, dx.$$

In order for this expression to make sense, we need to impose that the support of $q(x)$ contain the support of $I(x \geq z)p(x)$. Sampling N i.i.d points x_i from $q(x)$, our new estimate is

$$\hat{f}(z)_{Imp} \equiv \frac{1}{N} \sum_{i=1}^{N} I(x_i \geq z) \frac{p(x_i)}{q(x_i)}.$$

EXERCISE 6.1. *Show that $\hat{f}(z)_{Imp}$ is also unbiased and that its variance is*

$$\text{var}[\hat{f}(z)_{Imp}] = \frac{1}{N} \left(E_{p(x)} \left[I(x > z) \frac{p(x)}{q(x)} \right] - f(z)^2 \right).$$

Hence, the variance of $\hat{f}(z)_{Imp}$ will be smaller than that of $\hat{f}(z)$ if we choose our new sampling distribution to satisfy $q(x) > p(x)$ for $x > z$. Figure 6.1 shows the simulation result for the choice

$$q(x) = \lambda e^{-\lambda(x-z)},$$

where the optimal λ is from (6.2) with $a = z$.

6.5 Introduction to Markov Chains

6.5.1 Discrete Time, Discrete Space Markov Chains

Definition 6.1 (Markov Property)**.** *A set of jointly distributed random variables $\{X_k\}_{k \in \mathbb{N}}$ is said to satisfy a first order Markov property if $P(x_n|x_{n-1}, \ldots, x_0) = P(x_n|x_{n-1})$. In this case, the set forms a system known as a (discrete time) Markov chain. The space of values that X_k can take is called the state space. In this section, we will assume that the state space S is $\{1, \ldots, M\}$, unless stated otherwise.*

REMARK 6.4. *We will follow the convention of capitalizing the random variables and indicating their values in lower case.*

REMARK 6.5. *The Markov property can be informally stated as, "the future depends only on today and not the past."*

Definition 6.2 (Homogeneous Markov Chain)**.** *A Markov chain is called homogeneous if $P(X_n = y|X_{n-1} = x) = P(X_1 = y|X_0 = x) \equiv T_{xy}$, i.e. the transition probability T_{xy} from state x to state y is independent of time.*

REMARK 6.6. *The Markov property implies the following factorization of joint probability:*

$$P(x_n, \ldots, x_0) = \pi(x_0) \left(\prod_{k=1}^{n} T_{x_{k-1}x_k} \right),$$

where $\pi(x_0)$ is the initial distribution of X_0. By marginalizing this expression, we see that any joint distribution can be written in terms of the transition matrix T and the row vector π of initial state probabilities.

Example 6.8. *Marginalizing over $x_0, \ldots, x_{n-1}$, we get*

$$P(X_n = j) = (\pi T^n)_j,$$

which sums over all paths leading to the state $X_n = j$.

Definition 6.3 (Stationary Distribution)**.** *A distribution π_s is called stationary if $\pi_s T = \pi_s$, i.e. the distribution does not evolve under the transition T.*

A stronger form is

Definition 6.4 (Reversible Markov Chain)**.** *A distribution π on the state space S of a Markov chain with transition matrix T is said to satisfy detailed balance if $\forall x, y \in S$,*

$$\pi_x T_{xy} = \pi_y T_{yx}.$$

In this case, the Markov chain is said to be reversible.

EXERCISE 6.2. *Show that the distribution π in Definition 6.4 is automatically a stationary distribution.*

6.5.2 Two-State Model

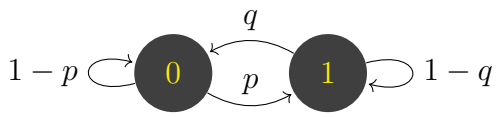


Figure 6.2: Two-state Markov chain.

Consider the Markov chain shown in Figure 6.2 with two possible states. The transition matrix is

$$T = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

The eigenvalues of this matrix are 1 and $1-p-q$. Hence, there exists a stationary distribution π_s, which by definition satisfies the equation $\pi_s T = \pi_s$. Solving this equation yields

$$\pi_s = \begin{pmatrix} \frac{q}{p+q} & \frac{p}{p+q} \end{pmatrix},$$

if p and q are not both equal to 0. If $p = q = 0$, then any distribution is a stationary distribution.

REMARK 6.7. *π_s satisfies $\pi_s T^n = \pi_s$ for any positive integer n.*

In fact, we can prove

Theorem 6.3. *Every finite Markov chain has a stationary distribution.*

Proof. Let T be the transition matrix of a Markov chain. Then, T is row stochastic and thus satisfies $T\mathbb{1} = \mathbb{1}$. Hence, T has at least one eigenvalue equal to 1, and we can find a row vector v satisfying $vT = v$. In order for v to represent probabilities, its entries must be non-negative. Suppose we break up v into non-negative and negative components: $v = v^+ - v^-$, where v^+ and v^- have non-negative entries and have disjoint positive-entry locations. Then,

$$vT = v \;\Rightarrow\; v^+(T-I) = v^-(T-I) \equiv w$$

for some vector w. The entries of w cannot be negative, because $v^+(T-I)$ has non-negative values in the zero entry locations of v^-, while $v^-(T-I)$ has non-negative values in the zero entry locations of v^+.

Note that $w\mathbb{1} = 0$ since $(T-I)\mathbb{1} = 0$; together with the fact that w has no negative entries, it thus implies that $w = 0$. Hence, v^+ and v^- are separately left eigenvectors of T with eigenvalue 1. □

EXERCISE 6.3. *Design an infinite Markov chain that does not have a stationary distribution.*

Now, it can be shown via mathematical induction that

$$T^n = \frac{1}{p+q}\begin{pmatrix} q+p(1-p-q)^n & p-p(1-p-q)^n \\ q-q(1-p-q)^n & p+q(1-p-q)^n \end{pmatrix} = \begin{pmatrix} \pi_s \\ \pi_s \end{pmatrix} + \frac{\alpha^n}{p+q}\begin{pmatrix} p & -p \\ -q & q \end{pmatrix},\ n \in \mathbb{N},$$

where $\alpha = 1-p-q$. If we assume that $0 < p+q < 2$, then $-1 < 1-p-q < 1$. In this case, we thus have $|1-p-q|^n \to 0$ as $n \to \infty$, and

$$\lim_{n\to\infty} T^n = \begin{pmatrix} \frac{q}{p+q} & \frac{p}{p+q} \\ \frac{q}{p+q} & \frac{p}{p+q} \end{pmatrix} = \begin{pmatrix} \pi_s \\ \pi_s \end{pmatrix} = \mathbb{1}\pi_s. \tag{6.3}$$

An immediate implication of this fact is that for any initial probability distribution π_0, we have $\pi_0 T^n \to \pi_s$ as $n \to \infty$. This observation lies at the heart of Markov Chain Monte Carlo (MCMC) methods. We formalize this concept by defining

Definition 6.5 (Steady State Distribution)**.** *If a stationary distribution π_s of a Markov chain with transition matrix T satisfies $\lim_{n\to\infty} T^n = \mathbb{1}\pi_s$, then π_s is called the steady state distribution.*

Theorem 6.4. *If a Markov chain has a steady state distribution, then it is the unique stationary distribution.*

Proof. Suppose π_s is a steady state distribution, and $\tilde{\pi}_s$ is a stationary distribution. Then, by definition,

$$\tilde{\pi}_s = \tilde{\pi}_s T^n, \text{ for any } n \in \mathbb{N}.$$

Taking the limit $n \to \infty$, we thus have

$$\tilde{\pi}_s = \tilde{\pi}_s \mathbb{1} \pi_s = \pi_s.$$

□

EXERCISE 6.4. *For $0 < p+q < 2$ in the two-state Markov model, suppose the initial state X_0 is distributed according to the stationary distribution. Given a positive integer M, define the ergodic average*

$$S_M := \frac{\sum_{n=0}^{M-1} X_n}{M}.$$

Compute the mean and variance of S_M. Show that in the limit $M \gg 1$,

$$\mathrm{Var}[S_M] \sim \frac{1}{M} \frac{pq(2-p-q)}{(p+q)^3}. \tag{6.4}$$

REMARK 6.8. *The limit in (6.3) breaks down when either (a) $p+q=0$ or (b) $p+q=2$. In the first case, we have $p=q=0$, so the transition matrix is an identify matrix, and the Markov chain becomes disconnected. Any initial distribution is thus a stationary distribution, and one cannot evolve a given stationary distribution into another stationary distribution. In the second case, we have $p=q=1$; i.e.*

$$T = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \Rightarrow T^{2n} = I_{2\times 2} \text{ and } T^{2n+1} = T,$$

so the limit $\lim_{n\to\infty} T^n$ does not exist. Case (a) is said to violate the irreducibility condition, and case (b) is said to violate the acyclicity condition. These two conditions are critical for the existence of a steady state distribution and for the validity of MCMC.

Definition 6.6 (Irreducible Chain)**.** *A Markov chain with state space S is called irreducible if $\forall x, y \in S$, $\exists$ a finite integer $n > 0$ such that $(T^n)_{xy} > 0$. In other words, we can go from any x to any y in the state space in a finite number of steps.*

Theorem 6.5. *An irreducible finite Markov chain has a unique stationary distribution.*

Proof. This theorm follows from the Perron-Frobenius theorem for irreducible stochastic matrices. □

REMARK 6.9. *The converse of this statement is not true. For example, when $q=0$ and $p \neq 0$ in the two-state model, $(0,1)$ is the unique stationary distribution; but, the chain is not irreducible, because $(T^n)_{21} = 0$ for all $n > 0$.*

Let us now define the notion of acyclic or aperiodic chain.

Definition 6.7 (Period)**.** *The period d_x of a state $x \in S$ is the greatest common divisor (GCD) of all positive integers n such that $(T^n)_{xx} > 0$. We say that x is aperiodic if $d_x = 1$.*

REMARK 6.10. *In the above definition, n is the number of steps in a path that starts at x and comes back to x.*

Definition 6.8 (Aperiodic Markov Chain)**.** *A Markov chain on S is called aperiodic or acyclic every element in S is aperiodic.*

Example 6.9. *The two-state model with $p = q = 1$ is not aperiodic, because each state has a period equal to 2.*

Example 6.10. *A bipartite Markov Chain is not aperiodic.*

Example 6.11. *If every state has a "self-loop," then the Markov chain is aperiodic.*

Example 6.12. *If at least one state in an irreducible Markov chain has a "self-loop," then the chain is aperiodic.*

In general, we have the following theorem:

Theorem 6.6. *Suppose there exist positive integers n and m such that $(T^n)_{xy} > 0$ and $(T^m)_{yx} > 0$. Then, $d_x = d_y$.*

Proof. We have

$$(T^{n+m})_{xx} = \sum_z (T^n)_{xz}(T^m)_{zx} \geq (T^n)_{xy}(T^m)_{yx} > 0.$$

Hence, $d_x | n + m$. Let k be any positive integer such that $(T^k)_{yy} > 0$. Then,

$$(T^{n+m+k})_{xx} \geq (T^n)_{xy}(T^k)_{yy}(T^m)_{yx} > 0,$$

and we must have $d_x | n + k + m$. But, since $d_x | n + m$, we must have $d_x | k$. By definition, d_y is the largest integer that divides all such k's, and thus $d_x | d_y$. Reversing the argument shows that $d_y | d_x$, implying that $d_x = d_y$. □

Corollary 6.1. *All elements of an irreducible Markov chain have the same period.*

REMARK 6.11. *Hence, to check whether an irreducible Markov chain is aperiodic, we only need to check that at least one state has period 1.*

Finally, we will state the following master theorem without proof:

Theorem 6.7. *A finite irreducible aperiodic Markov chain has a steady state distribution.*

REMARK 6.12. *It is important that the chain be finite in order for the irreducibility and aperiodicity conditions to suffice to guarantee the existence of a steady state distribution. This is because the finiteness condition automatically guarantees another condition known as the positive recurrence of an irreducible Markov chain (see Section 6.6).*

REMARK 6.13. *The converse is not necessarily true. That is, a Markov chain that has a steady state distribution may not be irreducible or aperiodic. For example, the two-state model with $p > 0$ and $q = 0$ is not irreducible, but has a steady state distribution.*

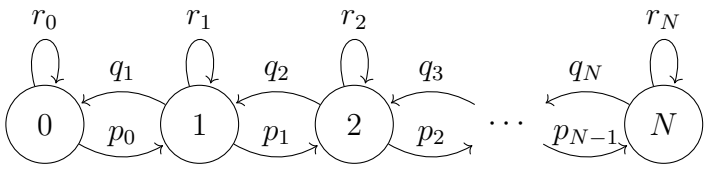


Figure 6.3: Birth and Death Markov chain. "Gambler's ruin" Markov chain, a.k.a. "you will eventually go bankrupt if you continue to gamble" model, is when $r_0 = 1$ and $N \to \infty$.

Example 6.13 (Birth and Death Process)**.** *The Markov chain shown in Figure 6.3 describes successive birth and death of individuals. By conservation of probability, the transition probabilities satisfy the constraint $r_n + p_n + q_n = 1$. Suppose $0 < p_n < 1$ and $0 < q_n < 1$ for $n = 1, \ldots, N-1$.*

Case 1: $0 < r_0 < 1, 0 < r_N < 1$. This chain is irreducible. It is also aperiodic, because 0 and N both have period 1, and the chain is irreducible. Hence, there exists a steady state distribution which is also the unique stationary distribution. Solving the detailed balance equation yields the following steady state distribution:

$$\pi_s = \left(\pi_0, \pi_0 \frac{p_0}{q_1}, \pi_0 \frac{p_0 p_1}{q_1 q_2}, \ldots, \pi_0 \frac{p_0 p_1 \cdots p_{N-1}}{q_1 q_2 \cdots q_N}\right)$$

where

$$\pi_0 = \left(1 + \frac{p_0}{q_1} + \frac{p_0 p_1}{q_1 q_2} + \cdots + \frac{p_0 p_1 \cdots p_{N-1}}{q_1 q_2 \cdots q_N}\right)^{-1}.$$

Case 2: $r_0 = 1, r_N < 1$. This case models a gambling situation where a player can either lose or gain a dollar at each bet and cannot play anymore when he/she runs out of money. In the limit $N \to \infty$, it is guaranteed that the gambler will eventually go bankrupt. The 0 state is called the absorbing state, because the chain will terminate at this node. The chain is thus not irreducible. However, it has a steady state distribution $(1, 0, 0, \ldots, 0)$.

Case 3: $r_0 = 1, r_N = 1$. There are two stationary states $(1, 0, 0, \ldots, 0)$ and $(0, 0, \ldots, 0, 1)$, but there is no steady state distribution.

6.6 Markov Chains Monte Carlo (MCMC)

The goal of MCMC is to construct a Markov chain such that its steady state distribution is the target probability distribution π_s from which we would like to sample. More precisely, we will guess an initial distribution π_0, and let it evolve under the designed Markov random walk such that

$$\pi_0 T^n \overset{n \to \infty}{\longrightarrow} \pi_s$$

As mentioned previously, Theorem 6.7, which ensures the existence of a steady state distribution for a finite Markov chain, holds because the finiteness condition guarantees the positive recurrence condition in that setting. Let us now briefly discuss what the positive recurrence condition means and how MCMC algorithm satisfies this criterion by construction.

Definition 6.9 (Positive Recurrence)**.** *A state of a Markov chain is said to be positive*

recurrent if its mean recurrence time is finite. A Markov chain is called positive recurrent if all of its states are positive recurrent.

For an irreducible Markov chain, checking the positive recurrence condition for any one state will imply that all other states are also positive recurrent. However, checking whether a state is positive recurrent or not is generally a difficult problem. But, checking this criterion is critical, because we would like to apply the following theorem:

Theorem 6.8 (Ergodic Theorem)**.** *An irreducible, positive recurrent, aperiodic Markov chain has a steady state distribution.*

Fortunately, we have the following theorem that allows us to replace the problem of checking for positive recurrence with the problem of checking for the existence of a stationary distribution:

Theorem 6.9. *An irreducible Markov chain is positive recurrent if and only if it has a stationary distribution. This stationary distribution is unique.*

Since the detailed balance condition automatically implies the existence of a stationary distribution, we will construct an irreducible aperiodic Markov chain such that the target distribution satisfies the detailed balance equations with respect to the transition matrix of the Markov chain. Then, Theorem 6.9 will ensure that this Markov chain is also positive recurrent. Finally, Theorem 6.8 will guarantee that the target distribution is the steady state distribution of the Markov chain.

6.6.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a two-step procedure for constructing an irreducible, positive recurrent, aperiodic Markov chain. Given a target distribution π_s with support S from which we would ultimately like to sample,

1. construct an irreducible proposal Markov chain on S with transition probability Q_{xy};
2. "balance" this Markov chain to satisfy detailed balance

$$(\pi_s)_x T_{xy} = (\pi_s)_y T_{yx}, \tag{6.5}$$

where T is obtained from suppressing high probability transitions in Q.

REMARK 6.14. *Because the balanced Markov chain with transition probability T is irreducible and has π_s as its stationary distribution, it will be positive recurrent by Theorem 6.9. The step of balancing Q will introduce self-loops, so the resulting Markov chain will be also aperidic. By the ergodic theorem (Theorem 6.8), π_s will be thus the steady state distribution of this balanced Markov chain.*

Without loss of generality, suppose that

$$(\pi_s)_x Q_{xy} > (\pi_s)_y Q_{yx},$$

which implies that there is an excess of transition from x to y. We need to dial down this net flow and satisfy the detailed balance condition. Hence, we choose

$$T_{xy} = Q_{xy}\left(\frac{(\pi_s)_y Q_{yx}}{(\pi_s)_x Q_{xy}}\right) \quad \text{and} \quad T_{yx} = Q_{yx}.$$

By construction, the detailed balance condition (6.5) is satisfied in this case. Generalizing this definition to arbitrary x and y, we define

$$\boxed{T_{xy} = Q_{xy}\min\left(1, \frac{(\pi_s)_y Q_{yx}}{(\pi_s)_x Q_{xy}}\right)}.$$

Metropolis-Hastings MCMC Pseudocode

Given: target distribution π_s on S.
Guess: proposal Markov chain with transition probability Q.
function MCMC(π_s, Q, burn-in n_0, nSim N):
    Initialize $x_0 \in S$
    Set $n = 0$
    While $n < N$:
        Sample y from $Q_{x_n y}$
        Set $r = \min\left(1, \frac{(\pi_s)_y Q_{yx_n}}{(\pi_s)_{x_n} Q_{x_n y}}\right)$
        Sample $u \sim \text{Unif}[0,1]$
        If $u \leq r$: $x_{n+1} = y$
        else: $x_{n+1} = x_n$
        n++
    Return $x_{n_0+1}, \ldots, x_N$

REMARK 6.15. *Notice that the target distribution appears as a ratio in the acceptance rate, so any normalization constant will cancel. So, we only need to know the target distribution up to a constant. This is very useful for sampling from Bayesian posterior distributions, where the denominator is usually very difficult to calculate.*

REMARK 6.16. *Because some proposed moves may get rejected with non-zero probability, in which case $x_{n+1} = x_n$, the new Markov chain has self-loops and is thus aperiodic. The target distribution π_s is thus the steady state distribution of this Markov chain.*

REMARK 6.17. *For sufficiently large n, the marginal distribution of x_n is roughly π_s.*

REMARK 6.18. *When two regions are separated by sharp drops in π_s, most proposed moves trying to cross the boundary may get rejected. MCMC in this case may not be able to explore the entire state space S efficiently. This phenomenon is known as poor mixing.*

REMARK 6.19. *How do you choose Q? It depends on several factors. For example, the state space S will inform which class of proposal distribution to consider. If the proposal steps are too short, then the MCMC samples may be too correlated. Some parameters such as the variance of proposal distribution may need to be adjusted to improve mixing.*

REMARK 6.20. *Why do we throw away n_0 burn-in samples? There is a chance that the MCMC steps will revisit a region near the initialization point x_0 even after n_0 steps. So, why do we need to throw away the first n_0 burn-in samples? One reason would be when we simulated multiple chains, say L of them, simultaneously and initialized the walks according to some distribution π_0 that differs from π_s. We want to reduce the number N of simulations in a single chain and speed up the entire process by parallelizing L independent chains. Then, the samples $x_n^{(i)}, i = 1, \ldots, L$, would not be distributed as π_s for small n, so we may want to wait until π_0 has stabilized to π_s. Another reason for the burn-in is to escape from bad initialization states lying in low probability regions.*

6.6.2 Gibbs Sampling

Suppose $X = (X_1, \ldots, X_d)$ is a vector of random variables jointly distributed as $P(X_1, \ldots, X_d)$. Gibbs sampling is an iterative sampling algorithm that is applicable when it is difficult to directly sample from the full joint distribution, but the conditional distributions $P(X_i|X_1, \ldots, X_{\hat{i}}, \ldots, X_d)$ are easy to handle.

Gibbs Sampler Pseudocode

Given: joint probability distribution $P(X_1, \ldots, X_d)$.
function Gibbs(P, nSim N):
    Initialize $x^0 = (x_1^{(0)}, \ldots, x_d^{(0)})$
    Set $n = 1$
    While $n \leq N$:
        Sample $x_1^{(n)}$ from $P(X_1|x_2^{(n-1)}, \ldots, x_d^{(n-1)})$
        Sample $x_2^{(n)}$ from $P(X_2|x_1^{(n)}, x_3^{(n-1)}, \ldots, x_d^{(n-1)})$
        Sample $x_3^{(n)}$ from $P(X_3|x_1^{(n)}, x_2^{(n)}, x_4^{(n-1)}, \ldots, x_d^{(n-1)})$
        $\vdots$
        Sample $x_d^{(n)}$ from $P(X_3|x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, \ldots, x_{d-1}^{(n)})$
        n++
    Return $x^0, x^1, \ldots, x^N$

Example 6.14. *Let $X_1, X_2, X_3 \in \mathbb{R}_{\geq 0}$ be jointly distributed as*

$$p(x_1, x_2, x_3) \propto \exp\left(-x_1 - x_2 - x_3 - \theta_{12}x_1x_2 - \theta_{13}x_1x_3 - \theta_{23}x_2x_3\right),$$

where $\theta_{ij} > 0$ are known constants. The full conditionals are

$$X_1|X_2 = x_2, X_3 = x_3 \sim Exp(1 + \theta_{12}x_2 + \theta_{13}x_3),$$

and similar expressions with a cyclic permutation of the indices for X_2 and X_3.

Example 6.15 (Ising Model)**.** *In Ising model, the spin configuration $s = (s_1, \ldots, s_N)$ on a periodic lattice in thermodynamic equilibrium has probability*

$$P(s) = \frac{e^{\beta(H\sum_i s_i + J\sum_{(i,j)\in\mathcal{N}} s_i s_j)}}{Z},\ s_i \in \{-1, 1\},$$