Summer Institute in Statistics and Modeling of Infectious Diseases
Module 8: MCMC Methods for Infectious Disease Studies

Instructors: Kari Auranen, Elizabeth Halloran and Vladimir Minin

July 14 – July 16, 2021

1 Probability refresher (self-study material)

We assume that we can assign probabilities to *events* — outcomes of a random experiment. For example, tossing a coin results in one of two possible events: H ="heads" and T="tails." We also need a concept of a random variable. Informally, a random variable X is a function or variable, whose value is generated by a random experiment. For example, we can define a binary random variable associated with a toss of a coin:

$$X = \begin{cases} 1 & \text{if heads,} \\ 0 & \text{if tails.} \end{cases}$$

Example: Discrete uniform random variable
Let $X \in \{1, 2, \ldots, n\}$, with $\Pr(X = i) = 1/n$ for all $i = 1, \ldots, n$.

Example: Bernoulli r.v.
$X \in \{0, 1\}$ with $\Pr(X = 1) = p$, $\Pr(X = 0) = 1 - p$ for $0 \leq p \leq 1$.

Example: Binomial r.v.
Let $X_i \sim \text{Bernoulli}(p)$. Then the number of successes $S_n = \sum_{i=1}^{n} X_i$ is called a *binomial r.v.* with

$$\Pr(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Example: Geometric r.v.
$X_1, X_2, \ldots$ ordered Bernoulli(p). Let $N = \min\{n : X_n = 1\}$ be the number of trials until the first success occurs, including the the successful trial.

$$\Pr(N = n) = (1-p)^{n-1} p \text{ for } n = 1, 2, \ldots$$

Note. There is an alternative definition of the geometric distribution does not count the successful trial so that $\Pr(N = n) = (1-p)^n p$.

We defined all discrete random variables above using probabilities of X taking a particular value. A function that assigns probabilities to random variable values is called a *probability mass function*. However, a more general way to define random variables is by specifying a cumulative distribution function.

Definition. $F(x) = \Pr(X \leq x)$ is called the *cumulative distribution function* (cdf) of X.

Properties of cdf:

1. $0 \leq F(x) \leq 1$.
2. $F(x) \leq F(y)$ for $x \leq y$.
3. $\lim_{x \to y^+} = F(y)$ ($F(x)$ is right-continuous).
4. $\lim_{x \to -\infty} = \Pr(X = -\infty)$ (usually $= 0$)
5. $\lim_{x \to \infty} = 1 - \Pr(X = \infty)$ (usually $= 1$)
6. $\Pr(X = x) = F(x) - F(x^-)$

Example: Discrete uniform random variable
For random variable U uniformly distributed over $\{1, 2, \ldots, n\}$, its cdf is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 1, \\ \frac{1}{n} & \text{if } 1 \leq x < 2, \\ \frac{2}{n} & \text{if } 2 \leq x < 3, \\ \quad \vdots & \\ \frac{n-1}{n} & \text{if } n-1 \leq x < n, \\ 1 & \text{if } x \geq n. \end{cases}$$

The probability mass function and cdf of U, with $n = 10$, are shown in Figure **??**, which also contains the probability mass function and cdf of a geometric random variable.

For continuous random variables, the analog of the probability mass function is a probability density function, defined as follows.

Definition. If $F(x) = \int_{-\infty}^{x} f(x)dx$ for some $f(x) \geq 0$, then $f(x)$ is called probability density function of X. If X has a probability density function, we say that X is absolutely continuous.

Note. $\int_a^b f(x)\mathrm{d}x = F(b) - F(a) = \Pr(a \leq X \leq b)$ for $a \leq b$. Moreover, $\frac{d}{dx}F(x) = f(x)$.

Example: Uniform random variable on $[0, 1]$
Random variable U with density

$$f(x) = \begin{cases} 1 & \text{if } x \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

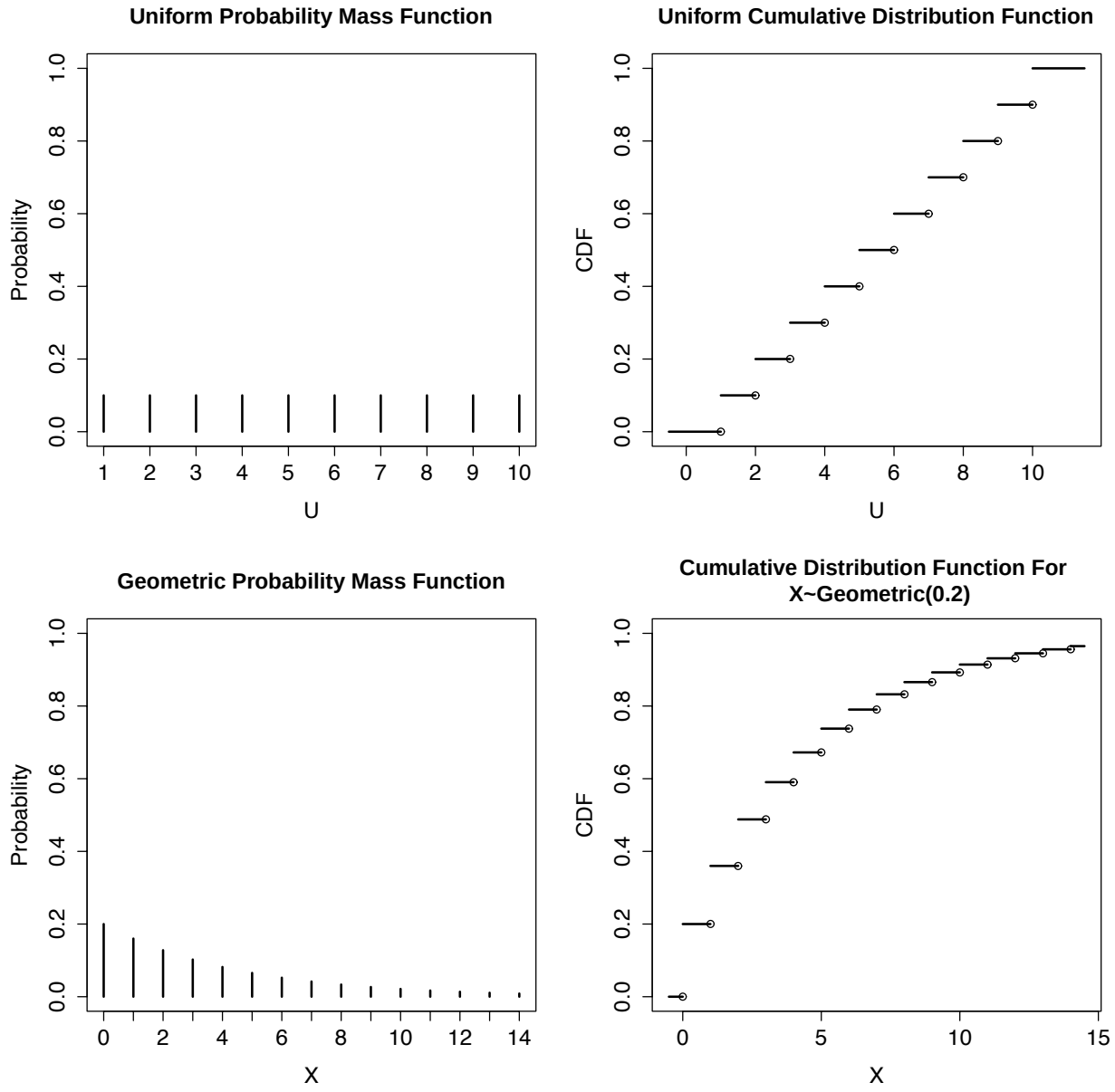


Figure 1: Probability mass functions (left column) and cumulative distribution functions (right column) of the discrete uniform random variable over $\{1, 2, \ldots, 10\}$ (top row) and geometric random variable with success probability $p = 0.2$ (bottom row).

The cdf of U is

$$F(x) = \begin{cases} 0 & \text{if } x < 0. \\ x & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1. \end{cases}$$

The top row of Figure **??** shows the probability mass function and cdf of U.

Definition. *Expectation* is defined as $E[g(X)] = \int_{-\infty}^{\infty} g(x)dF(x)$, where the integral is taken with respect to the measure induced by the cdf, aka probability measure. More concretely,

1. For discrete random variable X, $E[g(X)] = \sum_{k=1}^{\infty} g(x_k)\Pr(X = x_k)$.
2. For absolutely continuous random variable X, $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$.

Example: Exponential r.v.
Exponential random variable has density $f(x) = \lambda e^{-\lambda x}1_{\{x \geq 0\}}$, where $\lambda > 0$ is the rate parameter. Let $X \sim \text{Exp}(\lambda)$. The probability mass function and cdf of an exponential random variable are shown in the bottom row of Figure **??**. Then

$$E(X) = \int_0^{\infty} x\lambda e^{-\lambda x}dx = \begin{bmatrix} u = x & e^{-\lambda x}dx = dv \\ du = dx & -\frac{e^{-\lambda x}}{\lambda} = v \end{bmatrix} = \lambda \left[-x\frac{e^{-\lambda x}}{\lambda}\Big|_0^{\infty} + \int_0^{\infty} \frac{e^{-\lambda x}}{\lambda}dx \right] = \lambda \left[0 + \frac{1}{\lambda^2} \right] = \frac{1}{\lambda}.$$

Expectations are linear operators, meaning that for any collection of random variables $X_1, \ldots, X_n$,

$$\text{E}\left(\sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i \text{E}(X_i).$$

Linearity does not hold for the variance in general. However, if random variables $X_1, \ldots, X_n$ are independent, then

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i).$$

Definition. For events A and B in Ω we define *conditional probability*

$$\Pr(B \mid A) = \frac{\Pr(A \bigcap B)}{\Pr(A)}.$$

If we have a r.v. X defined on Ω, then we can define *conditional expectation*

$$\text{E}(X \mid A) = \frac{E(X1_{\{A\}})}{\Pr(A)}.$$

Conditioning on random variables is a little tricky, so we'll limit our discussion of this concept to

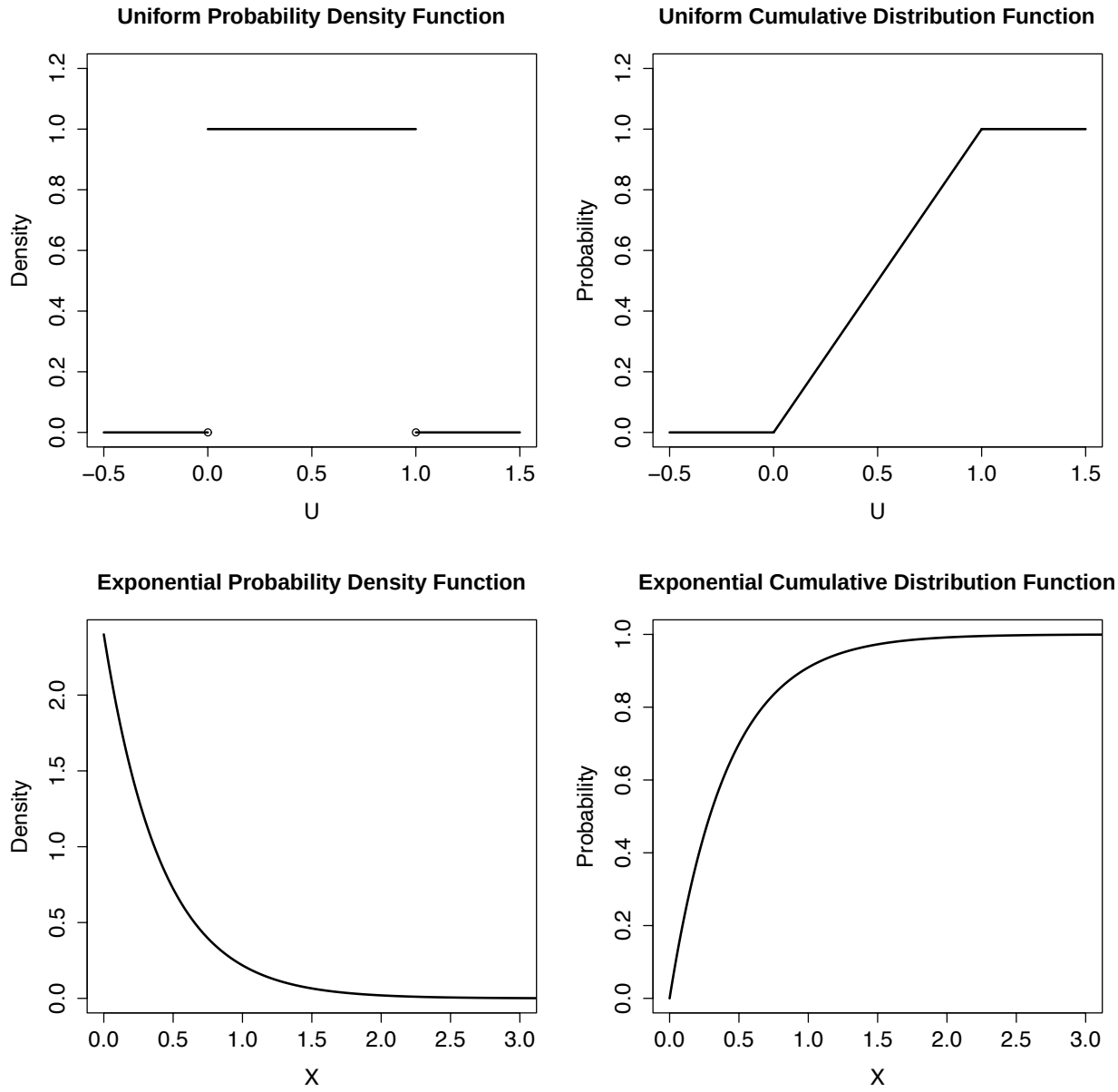


Figure 2: Probability density functions (left column) and cumulative distribution functions (right column) of the continuous uniform random variable on $[0, 1]$ (top row) and exponential random variable with rate parameter $\lambda = 2.4$ (bottom row).

1. discrete case:
$$\Pr(X = x \mid Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)},$$
and

2. absolutely continuous case:
$$F_{X \mid Y}(x|y) = \frac{\int_{-\infty}^{x} f_{XY}(z, y)dz}{f_Y(y)} \qquad \text{and} \qquad f_{X \mid Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)},$$
where $f_{XY}(x, y)$ is the joint density of X and Y and $f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)dx$ is the marginal density of Y.

Definition. Events A and B are *independent* if $\Pr(A \bigcap B) = \Pr(A)\Pr(B)$. Random variables X and Y are called independent if events $\{X \le a\}$ and $\{Y \le b\}$ are independent for all $a, b \in \mathbb{R}$, i.e. $\Pr(X \le a, Y \le b) = \Pr(X \le a)\Pr(Y \le b)$.

Note. If r.v.s X and Y are independent, then $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ and $\mathrm{E}(X \mid Y) = \mathrm{E}(X)$. The last equality says that Y carries no information about X.

Example: Hypergeometric distribution
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, $S_n = \sum_{i=1}^{n} X_i$, and $S_m = \sum_{i=1}^{m} X_i$ for $m < n$. We want to find the distribution of S_m conditional on S_n. We start with probability mass function

$$\Pr(S_m = j \mid S_n = k) = \frac{\Pr(S_m = j, S_n = k)}{\Pr(S_n = k)} = \frac{\Pr(\sum_{i=1}^{m} X_i = j, \sum_{i=1}^{n} X_i = k)}{\Pr(S_n = k)}$$
$$= \frac{\Pr\left(\sum_{i=1}^{m} X_i = j, \sum_{i=m+1}^{n} X_i = k - j\right)}{\Pr(S_n = k)} = [\text{independence}] = \frac{\Pr(\sum_{i=1}^{m} X_i = j)\Pr\left(\sum_{i=m+1}^{n} X_i = k - j\right)}{\Pr(S_n = k)}$$
$$= \frac{\binom{m}{j} p^j (1-p)^{m-j} \binom{n-m}{k-j} p^{k-j} (1-p)^{n-m-k+j}}{\binom{n}{k} p^k (1-p)^{n-k}} = \frac{\binom{m}{j}\binom{n-m}{k-j}}{\binom{n}{k}}.$$

This is the probability mass function of the hypergeometric distribution, which usually is defined as the number of red balls among the m balls drawn from an urn with k red and $n - k$ blue balls.

$$\mathrm{E}(S_m \mid S_n = k) = \sum_{i=1}^{m} \mathrm{E}(X_i \mid S_n = k) = [\text{symmetry}] = m\mathrm{E}(X_1 \mid S_n = k) = \frac{m}{n} \sum_{i=1}^{n} \mathrm{E}(X_i \mid S_n = k)$$
$$= \frac{m}{n}\mathrm{E}(S_n \mid S_n = k) = \frac{mk}{n}.$$

Notice that $X_1, \ldots, X_n$ don't have to be Bernoulli for $\mathrm{E}(S_m \mid S_n) = mS_n/n$ to hold.

Law of total probability If $B_1, \ldots, B_n$ are mutually exclusive events and $\bigcup_{i=1}^{n} B_i = \Omega$, then

$$\Pr(A) = \sum_{i=1}^{n} \Pr\left(A \bigcap B_i\right) = \sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i).$$

Law of total expectation Recall that $\mathrm{E}(X)$ is a scalar, but $\mathrm{E}(X \mid Y)$ is a random variable. Let X and Y be discrete r.v.s.

$$\mathrm{E}(X \mid Y = y) = \sum_{k=1}^{\infty} x_k \mathrm{Pr}(X = x_k \mid Y = y).$$

Proof.

$$\mathrm{E}[\mathrm{E}(X \mid Y)] = \sum_{k=1}^{\infty} \mathrm{E}(X \mid Y = y_k) \mathrm{Pr}(Y = y_k) = \sum_{k=1}^{\infty} \frac{\mathrm{E}\left(X 1_{\{Y=y_k\}}\right)}{\mathrm{Pr}(Y = y_k)} \mathrm{Pr}(Y = y_k)$$
$$= \sum_{k=1}^{\infty} \mathrm{E}\left(X 1_{\{Y=y_k\}}\right) = \mathrm{E}\left(X 1_{\{\bigcup_{k=1}^{\infty}\{Y=y_k\}\}}\right) = \mathrm{E}(X).$$

□

In general, $\mathrm{E}[\mathrm{E}(X \mid Y)] = \mathrm{E}(X)$. In fact, this equality is often used as a definition of the conditional expectation, when conditioning on a random variable (**?**).

Law of total variance Decomposing variance using conditioning is only slightly more complicated:

$$\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 = [\text{law of total expectation}] = \mathrm{E}\left[\mathrm{E}\left(X^2 \mid Y\right)\right] - \mathrm{E}[\mathrm{E}(X \mid Y)]^2$$
$$= [\text{def of variance}] = \mathrm{E}\left[\mathrm{Var}(X \mid Y) + \mathrm{E}(X \mid Y)^2\right] - \mathrm{E}[\mathrm{E}(X \mid Y)]^2 = \mathrm{E}[\mathrm{Var}(X \mid Y)]$$
$$+ \left\{\mathrm{E}\left[\mathrm{E}(X \mid Y)^2\right] - \mathrm{E}[\mathrm{E}(X \mid Y)]^2\right\} = [\text{def of variance}] = \mathrm{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}[\mathrm{E}(X \mid Y)].$$

Later in the course, we will be using the following two limit theorems that describe asymptotic behavior of empirical averages of random variables.

Theorem. *Strong Law of Large Numbers (SLLN). Let $X_1, X_2, \ldots$ be independent and identically distributed (iid) random variables with $\mu = E(X_1) < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu.$$

SLLN says that the empirical average of iid random variables converges to the theoretical average/expectation.

Theorem. *Central Limit Theorem (CLT). Let $X_1, X_2, \ldots$ be independent and identically distributed (iid) random variables with $\mu = E(X_1) < \infty$ and $0 < \sigma^2 = Var(X_1) < \infty$ and let $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Then*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1) \text{ approximately for large } n.$$

Informally, CLT says that for large n, the empirical average behaves as $\mathcal{N}(\mu, \sigma^2/n)$. Scaling of the variance by $1/n$ implies that averaging reduces variability, which makes intuitive sense.

2 Monte Carlo methods

The rest of the notes are largely based on (**?**). Although our driving applications of Monte Carlo integration will mostly revolve around Bayesian inference, we would like to point out that all Monte Carlo methods can (should?) be viewed as a numerical integration problem. Such problems usually start with either discrete ($\mathbf{x}$) or continuous ($\boldsymbol{\theta}$) vector of random variables. Despite the fact that distributions of these vectors are known only up to a proportionality constant, we are interested in taking expectations with respect to these distributions. Compare the following integration problems faced by physicists and Bayesian statisticians.

Statistical mechanics	*Bayesian statistics*
$\Pr(\mathbf{x}) = \frac{1}{Z} e^{-\mathcal{E}(\mathbf{x})}$	$\Pr(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{1}{C} \Pr(\mathbf{y} \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta})$
Objective: $\mathrm{E}[f(\mathbf{x})] = \sum_{\mathbf{x}} f(\mathbf{x}) \Pr(\mathbf{x})$	Objective: $\mathrm{E}[f(\boldsymbol{\theta}) \mid \mathbf{y}] = \int f(\boldsymbol{\theta}) \Pr(\boldsymbol{\theta} \mid \mathbf{y}) \, \mathrm{d}\boldsymbol{\theta}$

Note. Many applications involve both, intractable summation and integration:

$$\mathrm{E}[f(\mathbf{x}, \boldsymbol{\theta})] = \sum_{\mathbf{x}} \int f(\mathbf{x}, \boldsymbol{\theta}) \Pr(\mathbf{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}.$$

The above integration problems are difficult to solve even numerically, especially in high dimensions, e.g. when the length of $\mathbf{x}$ and/or $\boldsymbol{\theta}$ is on the order of $10^3 - 10^6$. All Monte Carlo techniques attempt to solve such high dimensional integration problems by stochastic simulation.

2.1 Classical Monte Carlo

In general, Monte Carlo integration aims at approximating expectations of the form

$$\mathrm{E}[h(X)] = \int h(x) f(x) \mathrm{d}x. \tag{1}$$

If $X_1, ..., X_n \overset{\text{iid}}{\sim} f(x)$ and $\mathrm{E}[h(X_1)] < \infty$, then we know from the strong law of large number (SLLN) that

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i) \overset{\text{a.s.}}{\to} \mathrm{E}_f[h(X_1)].$$

Therefore, we can approximate the desired expectation with

$$\bar{h}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \approx \mathrm{E}_f[h(X_1)]$$

for some large, yet finite n. Conveniently, the variance of this Monte Carlo estimator can be approximated as

$$\mathrm{Var}(\bar{h}_n) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} h(X_i) \right) = \frac{1}{n^2} \times n \times \mathrm{Var}[h(X_1)] = \approx \frac{1}{n} \times \underbrace{\frac{1}{n-1} \sum_{i=1}^{n} \left[h(X_i) - \bar{h}_n \right]^2}_{v_n},$$

where v_n is the sample variance of $h(X_1), \ldots, h(X_n)$ that can be obtained from the realized samples $X_1, \ldots, X_n$. Moreover, the central limit theorem says that

$$\frac{\bar{h}_n - \mathrm{E}_f[h(X_1)]}{\sqrt{v_n/n}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1),$$

allowing us to estimate the Monte Carlo error, e.g. $\bar{h}_n \pm 1.96\sqrt{v_n/n}$. The above derivation demonstrates that Monte Carlo error decreases with the number of samples n with the same rate as $1/\sqrt{n}$. In practice this means that you want to decrease Monte Carlo error by a factor of 10, you need to increase the number of samples by a factor of 100 — unfortunate, but c'est la vie.

Example: Second moment of beta distribution
Suppose, we know that $X \sim \mathrm{beta}(2,2)$, but want to compute the second of moment of X, $\mathrm{E}(X^2)$, without browsing Wikipedia. We simulate 1000 realizations from this distribution and save them in a vector (e.g, in R we type `x = rbeta(1000,2,2)`). Then we square each element of this vector (`y = x^2`) and obtain sample mean and sample variance of the resulting vector (`sample_mean = mean(y)` and `sample_var = var(y)`). Suppose the number we got are `mean = 0.29 and var = 0.05`, from which we form 95% confidence interval $0.29 \pm 1.96 * \sqrt{0.05/1000} = (0.276, 0.304)$. If we are unsatisfied with the accuracy of our Monte Carlo approximation, we can increase the number of iterations and recompute the Monte Carlo error.

Importance Sampling

In many situations classical Monte Carlo is impossible, because we can not sample from the target distribution $f(x)$. Therefore, we would like to be able to compute the integral (**??**) by sampling from some other, perhaps simpler, distribution $g(x)$. Importance sampling allows us to accomplish this task. The main idea is to rewrite the expectation of interest as

$$\mathrm{E}_f[h(X)] = \int h(x)\frac{f(x)}{g(x)}g(x)\mathrm{d}x = \mathrm{E}_g\left[h(X)\frac{f(X)}{g(X)}\right].$$

This representation suggests that we can generate $X_1, ..., X_n \overset{\mathrm{iid}}{\sim} g(x)$ and use the SLLN again to arrive at the approximation

$$\mathrm{E}_f[h(X)] \approx \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)}{g(X_i)}h(X_i).$$

Notice that the above approximation still requires knowledge of the normalizing constant of $f(x)$, which is unrealistic in most applications of importance sampling. Luckily there is an alternative importance sampling estimator that is as easy to compute as the original one:

$$\mathrm{E}_f[h(X)] \approx \frac{\sum_{i=1}^{n} h(X_i)\frac{f(X_i)}{g(X_i)}}{\sum_{i=1}^{n}\frac{f(X_i)}{g(X_i)}}.$$

In this estimator, the normalizing constants of both $f(x)$ and $g(x)$ cancel out and the denominator converges to $\int \frac{f(x)}{g(x)} g(x)\mathrm{d}x = \int f(x)\mathrm{d}x = 1$ by the SLLN again.

As illustrated by the next example, the importance sampling can be useful even if we can easily simulate from $f(x)$, because importance sampling can be used to reduce the Monte Carlo variance. In conclusion, we point out that the most difficult aspect of classical Monte Carlo is generating iid samples. Even importance sampling has severe limitations in high dimensions. In such difficult cases, Markov chain Monte Carlo (MCMC) can come to rescue. Before we master this numerical integration technique we need to refresh our knowledge of Markov chains.

Example: Estimating the tail of the standard normal distribution
See practical in 'dtmc-lab.pdf'.

2.2 Elementary Markov chain theory

In this section we will cover some basic results for Markov chains. For a more detailed treatment, see for example (**?**).

2.2.1 Definitions and examples

Definition. A stochastic process is a family of ordered random variables X_t, where t ranges over a suitable index set T, e.g. $T_1 = [0, \infty)$, $T_2 = \{1, 2, ...\}$.

Definition. A discrete time stochastic process $\{X_n\}_{n=0}^{\infty}$ is called a Markov chain if for all $n \geq 0$ and for all $i_0, i_1, ..., i_{n-1}, i, j$,

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

We call X_n a homogeneous Markov chain if $\Pr(X_{n+1} = j \,|\, X_n = i)$ is independent of n, and inhomogeneous otherwise. We also define 1-step transition probabilities

$$p_{ij} = \Pr(X_1 = j \,|\, X_0 = i) \quad \sum_j p_{ij} = 1, p_{ij} \geq 0 \text{ for all } i, j$$

and n-step transition probabilities

$$p_{ij}^{(n)} = \Pr(X_n = j \,|\, X_0 = i), \quad \sum_j p_{ij}^{(n)} = 1, p_{ij}^{(n)} \geq 0 \text{ for all } i, j$$

and collect them into transition probability matrix $\mathbf{P} = \{p_{ij}\}$ and n-step transition probability matrix $\mathbf{P}^{(n)} = \{p_{ij}^{(n)}\}$.

Note. A Markov chained is fully specified by its transition probability matrix $\mathbf{P}$ and an initial distribution $\boldsymbol{\nu}$.

Note. It is easy to show that n-step transition probabilities can be obtained by repeatedly multiplying transition probability matrix by itself. More precisely, $\mathbf{P}^{(n)} = \mathbf{P}^n$. This observation makes it easy to compute the marginal distribution of X_n, $\boldsymbol{\nu}^{(n)} = (\nu_1^{(n)}, \ldots, \nu_s^{(n)})$, where $\nu_i^{(n)} = \Pr(X_n = i)$. Then

$$\boldsymbol{\nu}^{(n)} = \boldsymbol{\nu}\mathbf{P}^n.$$

Definition. A Markov chain with transition probability matrix $\mathbf{P}$ is called irreducible if for any pair of states (i, j) there exists $n > 0$ such that $p_{ij}^{(n)} > 0$ and reducible otherwise. In other words, an irreducible Markov chain can get from any state to any state in a finite number of steps with positive provability.

Example: SIS 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 $\{X_n\}_{n=1}^{\infty}$. One possibility is to assume that X_ns are independent and $\mathrm{P}(X_n = I) = 1 - \mathrm{P}(X_n = S) = p$. 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

$$\mathbf{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

The directed graph with labeled edges, shown next to the matrix, graphically encodes the same information contained in the transition probability matrix. Such graphs are called transition graphs of Markov chains. If p and q are strictly positive, then the Markov chain is irreducible.

2.2.2 Stationary distribution and long term behavior

Definition. Any probability distribution $\boldsymbol{\pi}$ on state space E that satisfies $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T\mathbf{P}$ (also called the global balance equation) is called a stationary (or equilibrium) distribution of the corresponding homogeneous Markov chain.

Note. $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T P$ if and only if $\pi(i) = \sum_{j \in E} \pi_j p_{ji}$ for all $i \in E$.

Example: SIS model continued

Let's assume that $0 < p < 1$ and $0 < q < 1$ in the SIS model. Then global equations become

$$(\pi_1, \pi_2)\begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} = (\pi_1, \pi_2) \Rightarrow \begin{cases} \pi_1(1-p) + \pi_2 q & = \pi_1 \\ \pi_1 p + (1-q)\pi_2 & = \pi_2 \end{cases} \Rightarrow \pi_1 = \frac{q}{p}\pi_2.$$

Adding the constraint $\pi_1 + \pi_2 = 1$, we obtain the unique solution

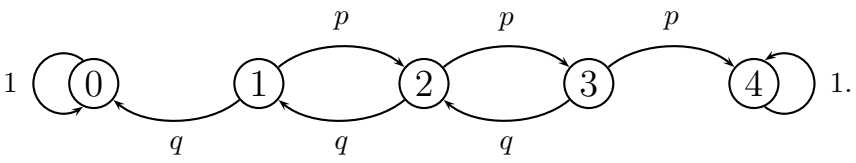
$$\pi_1 = \frac{q}{p+q} \qquad \text{and} \qquad \pi_2 = \frac{p}{p+q}.$$

Not all Markov chains have a stationary distribution and if a stationary distribution exists, it may be not unique as illustrated by the following example.

Example: Gambler's ruin

In this example, we assume that a gambler can increase or decrease his/her fortune by one with corresponding probabilities p and $q = 1-p$. The game ends as soon the gambler runs out of money or reaches a predefined fortune, 4 in our example. The transition matrix and the corresponding transition graph are shown below.

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 \\ 0 & q & 0 & p & 0 \\ 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$



The chain is reducible, because it is impossible to get out of states 0 and 4. Such states are called absorbing states. It is easy to show that vector $\boldsymbol{\pi}^T = (\alpha, 0, 0, 0, 1-\alpha)$ satisfies $\boldsymbol{\pi}^T\mathbf{P} = \boldsymbol{\pi}$ for any $\alpha \in [0, 1]$.

Global balance equations can be hard to check in practice when the Markov chain state space is large. However, there is an easier set of equations that one can check to ensure that a stationary distribution exists.

Definition. A probability vector $\boldsymbol{\pi}$ is said to satisfy detailed balance equations with respect to stochastic matrix $\mathbf{P}$ if

$$\pi_i p_{ij} = \pi_j p_{ji} \text{ for all } i, j.$$

Proposition. *(detail balance $\Rightarrow$ global balance) Let $\mathbf{P}$ be a transition probability matrix of X_n on E and let $\boldsymbol{\pi}$ be a probability distribution on E. If $\boldsymbol{\pi}$ satisfies detailed balance equations, then $\boldsymbol{\pi}$ also satisfies global balance equations.*

Proof: $\pi_i p_{ij} = \pi_j p_{ji} \Rightarrow \sum_{j\in E} \pi_i p_{ij} = \pi_i \cdot 1 = \sum_{j\in E} \pi_j p_{ji}$. $\square$

Note. Markov chains with a stationary distribution that satisfies detailed balance equations are often called reversible Markov chains. However, there is some disagreement among textbook authors about this term. For example, some authors require reversible chains to have initial distribution being equal to the stationary distribution. Irreducibility is also often added to the list of requirements for reversible Markov chains.

Example: Ehrenfest model of diffusion
See practical in 'dtmc-lab.pdf'.

Definition. An irreducible Markov chain is called recurrent if starting from any state the chain returns this state eventually with probability one. The recurrent chain is called positive recurrent if all expected return times are finite.

Proposition. *If a Markov chain is irreducible and positive recurrent, then there exists a stationary distribution and this distribution is unique.*

Note. Irreducible Markov chains on finite state spaces are always positive recurrent.

Proposition. *An irreducible Markov chain is positive recurrent if and only if the chain possesses a stationary distribution.*

Theorem. *(Ergodic Theorem) Let $\{X_n\}$ be an irreducible positive recurrent Markov chain with stationary distribution $\boldsymbol{\pi}$ and let $f : E \to \mathbb{R}$ be an arbitrary function that maps Markov chain states to real numbers satisfying $\sum_{i\in E} |f(i)|\pi_i < \infty$. Then for any initial distribution*

$$\lim_{N\to\infty} \underbrace{\frac{1}{N}\sum_{k=1}^{N} f(X_k)}_{time\ average} = \sum_{i\in E} f(i)\pi_i = \underbrace{E_{\boldsymbol{\pi}}[f(X)]}_{space\ average}.$$

Example: Ehrenfest model of diffusion (continued)
Separate practical.

Note. Just as the strong law of large numbers is the key behind Monte Carlo simulations, the ergodic theorem for Markov Chains is the reason why Markov chain Monte Carlo (MCMC) works. This remark naturally leads us to the next section.

2.3 Markov chain Monte Carlo

Before we dive into MCMC, let's ask ourselves why we are not happy with classical Monte Carlo and if there is any need to invent something more complicated. The main motivation for developing MCMC is the fact that classical Monte Carlo is very hard to implement in high dimensional spaces. MCMC also often experiences difficulties in high dimensions. However, for almost any high dimensional integration, it is fairly straightforward to formulate an MCMC algorithm, while the same is not true for classical Monte Carlo.

Recall that our objective in MCMC is the same as in classical Monte Carlo: to estimate expectations of the form

$$\mathrm{E}_{\boldsymbol{\pi}}[h(\mathbf{x})] = \sum_{\boldsymbol{x}\in E} \pi_{\boldsymbol{x}} h(\mathbf{x}).$$

Notice that here we assume that our state space is discrete so the above expectation is a finite sum. However we assume that the size of E is so large that carrying out this summation even on fastest computers is impractical. We also assume that we do not know how to produce iid samples from $\boldsymbol{\pi}$. The general MCMC strategy then is to construct an ergodic Markov chain $\{X_n\}$ with stationary distribution $\boldsymbol{\pi}$. Then from the ergodic theorem and N realizations from the Markov chain, we get

$$\mathrm{E}_{\boldsymbol{\pi}}[h(\mathbf{x})] \approx \frac{1}{N}\sum_{i=1}^{N} h(X_i).$$

The question is how to construct such a Markov chain, $\{X_n\}$.

2.3.1 Metropolis-Hastings algorithm

As always in MCMC, we start with a target distribution $\boldsymbol{\pi}$. Given some initial value $X_0 = x_0$, we construct a Markov chain according to the following set of rules (**?**).

Algorithm 2.1 Metropolis-Hastings Algorithm: approximate $\mathrm{E}_{\boldsymbol{\pi}}[h(\mathbf{x})]$

1: Start with some initial value $X_0 = x_0$.
2: **for** $n = 0$ to N **do**
3: Simulate a candidate value $Y \sim q(j \mid X_n = i)$. Suppose $Y = j$.
4: Compute the Metropolis-Hastings acceptance probability

$$a_{ij} = \min\left\{\frac{\pi_j q(i \mid j)}{\pi_i q(j \mid i)}, 1\right\}$$

5: Generate $U \sim \mathrm{Unif}[0, 1]$.
6: Accept the candidate $Y = j$ if $U \leq a_{ij}$, otherwise set $X_{n+1} = X_n$. More specifically, set

$$X_{n+1} = \begin{cases} Y & \text{if } U \leq a_{ij} \\ X_n & \text{if } U > a_{ij} \end{cases}$$

7: **end for**
8: **return** $\frac{1}{N}\sum_{i=1}^{N} h(X_i)$.

Proposition. *The Metropolis-Hastings algorithm generates a Markov chain with stationary distribution* $\boldsymbol{\pi}$.

Proof: Let $\mathbf{P} = \{p_{ij}\}$ be the transition matrix for X_n. Then for $i \neq j$,

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

Again, for $i \neq j$,

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

and

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

So we have shown $\pi_i p_{ij} = \pi_j p_{ji}$. We require $\pi_i > 0$ for all i and $q(i|j) > 0 \Leftrightarrow q(j|i) > 0$. Since we have detailed balance, we conclude that $\boldsymbol{\pi}$ is a stationary distribution. □

Note. If we choose $\{q(i,j)\}$ so that $\{X_n\}$ is irreducible, then $\{X_n\}$ is positive recurrent by the stationary distribution criterion. Therefore, we can use the Ergodic theorem.

Note. We do not need a normalizing constant of $\boldsymbol{\pi}$ in order to execute the Metropolis-Hastings algorithm. This is important, because in most applications of MCMC (e.g., Bayesian statistics) the target distribution is not normalized and the normalization constant is unknown.

Example: Toric Ising model on a circle

We model ferromagnetism with a set of n electron spins, $\mathbf{x}$. We assume that spins are arranged on a circles and have two directions, denoted by 1 and -1. The Gibbs distribution of configuration $\mathbf{x}$ is

$$\pi(\mathbf{x}) = \frac{1}{Z} e^{\beta \sum_{i=1}^{n} x_i x_{i+1}},$$

where the normalizing constant

$$Z = \sum_{\mathbf{x} \in \{1,-1\}^n} e^{\beta \sum_{i=1}^{n} x_i x_{i+1}}$$

is called a partition function. In this particular example, Z can be computed using a transfer matrix method, but we will pretend that Z is not available to us.

To set up a Metropolis-Hastings algorithm, we need a proposal mechanism to move from one configuration to another. At each step, let's choose a site uniformly at random and change the direction of the spin. This translates to the proposal probabilities

$$q(\mathbf{y} \,|\, \mathbf{x}) = q(\mathbf{x} \,|\, \mathbf{y}) = \begin{cases} \frac{1}{n} & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ differ at exactly one location,} \\ 0 & \text{otherwise.} \end{cases}$$

If $\mathbf{x}^{(t)}$ is the current state of the Markov chain and $\mathbf{x}'$ is a proposed state with the jth site changed to the opposite direction, then

$$a_{\mathbf{x}^{(t)}, \mathbf{x}'} = \frac{\pi(\mathbf{x}') \frac{1}{n}}{\pi(\mathbf{x}^{(t)}) \frac{1}{n}} = \frac{e^{\beta \sum_{i \notin \{j, j-1\}} x_i^{(t)} x_{i+1}^{(t)}} e^{\beta(-x_{j-1}^{(t)} x_j^{(t)} - x_j^{(t)} x_{j+1}^{(t)})}}{e^{\beta \sum_{i \notin \{j, j-1\}} x_i^{(t)} x_{i+1}^{(t)}} e^{\beta(x_{j-1}^{(t)} x_j^{(t)} + x_j^{(t)} x_{j+1}^{(t)})}} = e^{-2\beta x_j^{(t)} (x_{j-1}^{(t)} + x_{j+1}^{(t)})}.$$

Clearly, this proposal mechanism makes it possible to get from any state to any other state of spin configurations, so the Metropolis-Hastings chain is irreducible.

Variants of Metropolis-Hastings:

1. $q(i \,|\, j) = q(j \,|\, i)$ - symmetric proposal. This is the original Metropolis algorithm (**?**). Here, the acceptance probability simplifies to $a_{ij} = \min\left\{\frac{\pi_j}{\pi_i}, 1\right\}$. So we move to a more probable state with probability 1, and move to less probable states sometimes (more rarely if the candidate is much less probable).

2. Independence sampler: $q(j \,|\, i) = q(j)$. Note this is *not* the same as iid sampling. Independence sampler is still a Markov chain, since the sampler can stay in the same place with some probability at each step of the algorithm.

Metropolis-Hastings algorithm can be executed without any difficulties on continuous state spaces. This requires defining Markov chains on continuous state spaces.

Definition. A sequence of r.v.s $X_0, X_1, \ldots$ is called a Markov chain on a state space E if $\forall t$ and $\forall A \subset E$

$$\Pr\left(X_{n+1} \in A \,|\, X_n, X_{n-1}, \ldots, X_0\right) = \Pr\left(X_{n+1} \in A \,|\, X_n\right) = [\text{in homogeneous case}] = \Pr\left(X_1 \in A \,|\, X_0\right).$$

A family of functions $\Pr\left(X_1 \in A \,|\, x\right) = K(x, A)$ is called transition kernel.

If there exists $f(x, y)$ such that

$$\Pr\left(X_1 \in A \,|\, x\right) = \int_A f(x, y)dy,$$

then $f(x, y)$ is called transition kernel density. This is a direct analog of a transition probability matrix in discrete state spaces.

A lot of notions transfer from discrete to continuous state spaces: irreducibility, periodicity, etc. Chapman-Kolmogorov, for example takes the following form:

$$K^{m+n}(x, A) = \int_E K^n(y, A)K^m(x, dy),$$

where $K^n(x, A) = \Pr\left(X_n \in A \,|\, x\right)$.

Definition. A probability distribution π on E is called a stationary distribution of a Markov process with transition kernel $K(x, A)$ if for any Borel set B in E

$$\pi(B) = \int_E K(x, B)\pi(dx).$$

If transition kernel density is available, then global balance equation can be re-written

$$\pi(y) = \int_E \pi(x)f(x, y)dx.$$

Using the introduced terminology, we define a Metropolis-Hastings algorithm for continuous state spaces Let $f(\mathbf{x})$ be a target density, where $\mathbf{x}$ is a vector in $\mathbb{R}^n$ now. Then we simply can replace proposal probabilities $q(j \mid i)$ with proposal densities $q(\mathbf{y} \mid \mathbf{x})$ so that Metropolis-Hastings acceptance ratio becomes

$$a(\mathbf{x}, \mathbf{y}) = \min \left\{ \frac{f(\mathbf{y})q(\mathbf{x} \mid \mathbf{y})}{f(\mathbf{x})q(\mathbf{y} \mid \mathbf{x})}, 1 \right\} \tag{2}$$

The rest of the algorithm remains intact. As before, we need to ensure that the resulting Markov chain is irreducible. One way to do this is to require that $q(\mathbf{y} \mid \mathbf{x}) > 0$ for all $\mathbf{x}, \mathbf{y} \in E$. Alternately, a less restrictive assumption is that there exists some fixed δ and ϵ so that $q(\mathbf{y} \mid \mathbf{x}) > \epsilon$ if $|\mathbf{x}-\mathbf{y}| < \delta$.

A common example of a proposal scheme is a random walk. The proposal is given by

$$Y = X_n + \epsilon_n \tag{3}$$

where ϵ_n is some random perturbation independent of X_n with $E(\epsilon_n) = 0$. By convention, random walk proposals are always taken to be symmetric and have the following form

$$q(y \mid x) = q(|y - x|). \tag{4}$$

Example: Approximating standard normal distribution
Separate practical

2.3.2 Combining Markov kernels

Suppose we have constructed m transition kernels with stationary distribution $\boldsymbol{\pi}$. In discrete state spaces, this means that we have m transition matrices, $\mathbf{P}_1, \ldots, \mathbf{P}_m$, where $\boldsymbol{\pi}^T \mathbf{P}_i = \boldsymbol{\pi}$ for all $i = 1, \ldots, m$. There are two simple ways to combine these transition kernels. First, we can construct a Markov chain, where at each step we sequentially generate new states from all kernels in a predetermined order. The transition probability matrix of this new Markov chain is

$$\mathbf{S} = \mathbf{P}_1 \times \cdots \times \mathbf{P}_m.$$

It is easy to show that $\boldsymbol{\pi}^T \mathbf{S} = \boldsymbol{\pi}$. So as long the new Markov chain is irreducible, we can use the Ergodic theorem applied to the new Markov chain. In the second method of combining Markov kernels, we first create a probability vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$. Next, we first randomly select kernel i with probability α_i and then use this kernel to advance the Markov chain. The corresponding transition kernel is

$$\mathbf{R} = \sum_{i=1}^{m} \alpha_i \mathbf{P}_i.$$

Again, $\boldsymbol{\pi}^T \mathbf{R} = \boldsymbol{\pi}$, so this MCMC sampling strategy is valid as long as we can guarantee irreducibility.

2.3.3 Gibbs sampling

Suppose now that our state space is a Cartesian product of smaller subspaces, $\mathbf{E} = E_1 \times \cdots \times E_m$. The target distribution or density is $f(\mathbf{x})$ and we still want to calculate $\mathrm{E}_f[h(\mathbf{x})]$. We assume that we can sample from full conditional distributions $x_i \,|\, \mathbf{x}_{-i}$, where the notation $\mathbf{x}_{-i}$ means all elements of $\mathbf{x}$ except the ith component. It turns out that if keep iteratively sampling from these full conditionals, we will form a Markov chain with the required target distribution or density $f(\mathbf{x})$. More formally, let's look at the sequential scan Gibbs sampling algorithm below.

Algorithm 2.2 *Sequential Scan* Gibbs Sampling Algorithm: approximate $\mathrm{E}_{\mathbf{f}}[h(\mathbf{x})]$

1: Start with some initial value $\mathbf{x}^{(0)}$.
2: **for** $t = 0$ to N **do**
3: Sample $x_1^{(t+1)} \sim f_1\left(x_1 \,|\, \mathbf{x}_{-1}^{(t)}\right)$
4: Sample $x_2^{(t+1)} \sim f_2\left(x_2 \,|\, x_1^{(t+1)}, x_3^{(t)}, ..., x_m^{(t)}\right)$
$\vdots$
5: Sample $x_m^{(t+1)} \sim f_m\left(x_m \,|\, \mathbf{x}_{-m}^{(t+1)}\right)$
6: **end for**
7: **return** $\frac{1}{N}\sum_{t=1}^{N} h(\mathbf{x}^{(t)})$.

The question remains why the Gibbs sampling algorithm actually works. Consider one possible move in the Gibbs sampling procedure from $\mathbf{x}^{\text{cur}} \to \mathbf{x}^{\text{new}}$, where $\mathbf{x}^{\text{new}}$ is obtained by replacing the ith component in $\mathbf{x}^{\text{cur}}$ with a draw from the full conditional $f_i\left(x_i \,|\, \mathbf{x}_{-i}^{\text{cur}}\right)$. Now, let's view this "move" in light of the Metropolis-Hastings algorithm. Our proposal density will be the full conditional itself. Then the Metropolis-Hastings acceptance ratio becomes

$$a(\mathbf{x}^{\text{cur}}, \mathbf{x}^{\text{new}}) = \min\left\{\frac{f\left(x_i^{\text{new}}, \mathbf{x}_{-i}^{\text{cur}}\right) f_i\left(x_i^{\text{cur}} \,|\, \mathbf{x}_{-i}^{\text{cur}}\right)}{f\left(x_i^{\text{cur}}, \mathbf{x}_{-i}^{\text{cur}}\right) f_i\left(x_i^{\text{new}} \,|\, \mathbf{x}_{-i}^{\text{cur}}\right)}, 1\right\} = \min\left\{\frac{f\left(\mathbf{x}_{-i}^{\text{cur}}\right)}{f\left(\mathbf{x}_{-i}^{\text{cur}}\right)}, 1\right\} = 1. \tag{5}$$

So when we use full conditionals as our proposals in the Metropolis-Hastings step, we always accept. This means that drawing from a full conditional distribution produces a Markov chain with stationary distribution $f(\mathbf{x})$. Clearly, we can not keep updating just the ith component, because we will not be able to explore the whole state space this way. Therefore, we update each component in turn. This is not the only way to execute Gibbs sampling. We can also randomly select an component to update. This is called a random scan Gibbs sampling.

Note. Although it is not obvious, but in many cases sampling from full conditional distribution does not require knowing the normalizing constant of the target distribution.

Example: Ising model (continued)
Recall that in the Ising model

$$\pi(\mathbf{x}) = \frac{1}{Z} e^{\beta \sum_{i=1}^{k} x_i x_{i+1}},$$

Algorithm 2.3 *Random Scan* Gibbs Sampling Algorithm: approximate $\mathrm{E}_{\mathbf{f}}[h(\mathbf{x})]$

1: Start with some initial value $\mathbf{x}_0$.
2: **for** $t = 0$ to N **do**
3: Sample index i by drawing a random variable with probability mass function $\{\alpha_1, \ldots, \alpha_m\}$.
4: Sample $x_i^{(t+1)} \sim f_i\left(x_i \mid \mathbf{x}_{-i}^{(t)}\right)$
5: **end for**
6: **return** $\frac{1}{N}\sum_{t=1}^{N} h(\mathbf{x}^t)$.

where $\mathbf{x} = (x_1, \ldots, x_k)$. The full conditional is

$$\pi(x_j \mid \mathbf{x}_{-j}) = \frac{\pi(\mathbf{x})}{\pi(\mathbf{x}_{-j})} = \frac{\pi(\mathbf{x})}{\sum_{y\in\{-1,1\}} \pi(y, \mathbf{x}_{-j})} = \frac{\frac{1}{Z}e^{\beta\sum_{i=1}^{k} x_i x_{i+1}}}{\frac{1}{Z}e^{\beta\sum_{i\notin\{j,j-1\}} x_i x_{i+1}}\left[e^{\beta(x_{j-1}+x_{j+1})} + e^{-\beta(x_{j-1}+x_{j+1})}\right]}$$
$$= \frac{e^{\beta(x_{j-1}x_j + x_j x_{j+1})}}{e^{\beta(x_{j-1}+x_{j+1})} + e^{-\beta(x_{j-1}+x_{j+1})}}.$$

2.3.4 Combining Gibbs and Metropolis-Hastings samplers

Our discussion of combining Markov kernels suggests that it it is possible to combine Gibbs and Metropolis-Hastings steps in MCMC sampler.

Example: Beta-binomial hierarchical model
Separate practical

2.3.5 Variance of MCMC estimators

Let $X_1, X_2, \ldots$ be an ergodic Markov chain and

$$\hat{h} = \frac{1}{N}\sum_{i=1}^{N} h(X_i)$$

be the corresponding estimate of $\mathrm{E}_f[h(X)]$, where f is the stationary distribution of the chain. Estimating the variance of this estimator is complicated by the dependence among $X_1, X_2, \ldots, X_N$. One simple way to get around it is to subsample the Markov chain output so that the resulting sample is approximately iid. Then, the variance can be approximated as before with

$$\hat{v} = \frac{1}{N^2}\sum_{i=1}^{N}[h(X_i) - \hat{h}]^2.$$

Subsampling can be wasteful and impractical for slow mixing chains. One way to quantify the loss of efficiency due to dependence among samples is to compute the effective sample size,

$$\hat{N}_{eff} = \frac{N}{\kappa_h},$$

where

$$\kappa_h = 1 + 2\sum_{i=1}^{\infty} \text{corr}[h(X_0), h(X_i)]$$

is the autocorrelation time that can be estimated using spectral analysis for time series. After $\hat{N}_{eff}$ is obtained, the variance of $\hat{h}$ is computed as

$$\tilde{v} = \frac{1}{N}\frac{1}{N_{eff}}\sum_{i=1}^{N}[h(X_i) - \hat{h}]^2.$$

2.3.6 Convergence diagnostics

Although there is no definitive way to tell whether one ran a Markov chain long enough, several useful diagnostic tools can illuminate problems with the sampler, bugs in the code, and suggest ways to improve the design of the MCMC sampler. We organize these tools into the following categories:

1. Visualizing MCMC output. Trace plots provide a useful method for detecting problems with MCMC convergence and mixing. Ideally, trace plots of unnormalized log posterior and model parameters should look like stationary time series. Slowly mixing Markov chains produce trace plots with high autocorrelation, which can be further visualized by autocorrelation plots at different lags. Slow mixing does not imply lack of convergence.

2. Comparing batches. We take two vectors from MCMC output: $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{T/2})$ and $(\boldsymbol{\theta}^{(T/2+1)}, \ldots, \boldsymbol{\theta}^{T})$. If MCMC achieved stationarity at the time of collecting these batches, then both vectors follow the same stationary distribution. To test this hypothesis, we can apply Kolmogorov-Smirnov test, for example.

3. Renewal theory methods. Monitor return times of the Markov chain to a particular state and check whether these return times are iid. Care is needed on continuous state-spaces. See (**?**) for details.

4. Comparing multiple chains, started from random initial conditions. There are many ways of performing such a comparison. One popular method is called Potential Scale Reduction Factor (PSRF) due to **?**.

Many useful diagnostic tools are implemented in R package CODA (**?**). **?** and **?** review many of the methods in depth.

2.3.7 Special topics

1. Perfect sampling. Strictly speaking perfect sampling is a Monte Carlo, not Markov chain Monte Carlo method. However, the algorithm relies on running Markov chains. Coupling

these Markov chains in a certain way (coupling from the past), allows one to generate a sample from the stationary distribution exactly (**?**).

2. **?** formally introduced a Metropolis-Hastings algorithm for sampling parameter spaces with variable dimensions. This class of MCMC is called reversible jump MCMC (rjMCMC). **?** and **?** have developed reversible jump procedure before Peter Green popularized these algorithms with his now classical 1995 paper.

3. Simulated tempering. Simulated tempering, proposed by **?**, constructs a multivariate Markov chain $(X^{(1)}, \ldots, X^{(n)}$ to sample from the vector-valued function $(f(\mathbf{x}), f^{1/\tau_1}(\mathbf{x}), \ldots, f^{1/\tau_n}(\mathbf{x}))^T$, The auxiliary "heated" chains allow for better exploration of multimodal targets. The idea is similar in spirit to simulated annealing.

4. Sequential importance sampling and particle filters. These methods are useful for sequential building of instrumental densities in high dimensions. The main idea is to use the following representation:

$$f(x_1, \ldots, x_n) = f(x_1 \mid x_2, \ldots, x_n) f(x_2 \mid x_3, \ldots, x_n) \cdots f(x_n).$$

   Using specific structure of the problem at hand, conditioning often simplifies due to conditional independences (**??**).

5. Last, but not least, Hamiltonian Monte Carlo.