Simulation - Lectures

Yee Whye Teh

Part A Simulation

TT 2013

Administrivia

- **Lectures**: Wednesdays and Fridays 12-1pm Weeks 1-4.
- **Departmental problem classes**: Wednesdays 4-5pm Weeks 3-6.
- Hand in problem sheet solutions by
  Mondays noon in 1 South Parks Road.
- Webpage: http://www.stats.ox.ac.uk/%7Eteh/simulation.html
- This course builds upon the notes of Mattias Winkel, Geoff Nicholls, and Arnaud Doucet.

Outline

Outline

Introduction

Inversion Method

Transformation Methods

Rejection Sampling

Importance Sampling

Markov Chain Monte Carlo

Metropolis-Hastings

Monte Carlo Simulation Methods

- Computational tools for the simulation of random variables.
- These simulation methods, aka Monte Carlo methods, are used in many fields including statistical physics, computational chemistry, statistical inference, genetics, finance etc.
- The Metropolis algorithm was named the top algorithm of the 20th century by a committee of mathematicians, computer scientists & physicists.
- With the dramatic increase of computational power, Monte Carlo methods are increasingly used.

Objectives of the Course

- Introduce the main tools for the simulation of random variables:
  - inversion method,
  - transformation method,
  - rejection sampling,
  - importance sampling,
  - Markov chain Monte Carlo including Metropolis-Hastings.
- Understand the theoretical foundations and convergence properties of these methods.
- Learn to derive and implement specific algorithms for given random variables.

Computing Expectations

- Assume you are interested in computing

$$\theta = \mathbb{E}\left(\phi(X)\right) = \int_{\Omega} \phi(x) F(dx)$$

where X is a random variable (r.v.) taking values in Ω with distribution F and $\phi : \Omega \to \mathbb{R}$.

- It is impossible to compute θ exactly in most realistic applications.
- Example: $\Omega = \mathbb{R}^d$, $X \sim \mathcal{N}(\mu, \Sigma)$ and $\phi(x) = \mathbb{I}\left(\sum_{k=1}^{d} x_k^2 \geq \alpha\right)$.
- Example: $\Omega = \mathbb{R}^d$, $X \sim \mathcal{N}(\mu, \Sigma)$ and $\phi(x) = \mathbb{I}\left(x_1 < 0, ..., x_d < 0\right)$.

Example: Queuing Systems

- Customers arrive at a shop and queue to be served. Their requests require varying amount of time.
- The manager cares about customer satisfaction and not excessively exceeding the 9am-5pm working day of his employees.
- Mathematically we could set up stochastic models for the arrival process of customers and for the service time based on past experience.
- **Question**: If the shop assistants continue to deal with all customers in the shop at 5pm, what is the probability that they will have served all the customers by 5.30pm?
- If we call X the number of customers in the shop at 5.30pm then the probability of interest is

$$\mathbb{P}(X = 0) = \mathbb{E}(\mathbb{I}(X = 0)).$$

- For realistic models, we typically do not know analytically the distribution of X.

Example: Particle in a Random Medium

- A particle $(X_t)_{t=1,2,\ldots}$ evolves according to a stochastic model on $\Omega = \mathbb{R}^d$.
- At each time step t, it is absorbed with probability $1 - G(X_t)$ where $G : \Omega \to [0, 1]$.
- **Question**: What is the probability that the particle has not yet been absorbed at time T?
- The probability of interest is

$$\mathbb{P}\left(\text{not absorbed at time } T\right) = \mathbb{E}\left[G(X_1)G(X_2)\cdots G(X_T)\right].$$

- For realistic models, we cannot compute this probability.

Example: Ising Model

- The Ising model serves to model the behavior of a magnet and is the best known/most researched model in statistical physics.
- The magnetism of a material is modelled by the collective contribution of dipole moments of many atomic spins.
- Consider a simple 2D-Ising model on a finite lattice $\mathcal{G} = \{1, 2, ..., m\} \times \{1, 2, ..., m\}$ where each site $\sigma = (i, j)$ hosts a particle with a $+1$ or -1 spin modeled as a r.v. X_σ.
- The distribution of $X = \{X_\sigma\}_{\sigma \in \mathcal{G}}$ on $\{-1, 1\}^{m^2}$ is given by

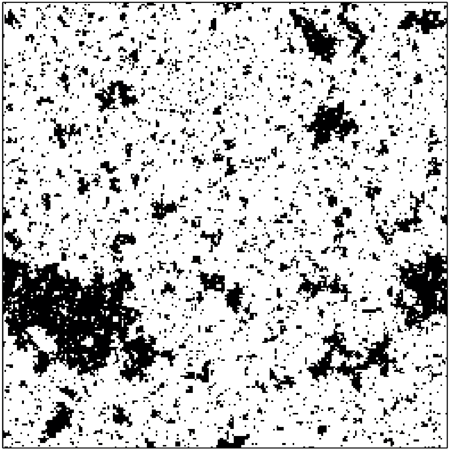
  $$\pi(x) = \frac{\exp(-\beta U(x))}{Z_\beta}$$

  where $\beta > 0$ is the inverse temperature and the potential energy is

  $$U(x) = -J \sum_{\sigma \sim \sigma'} x_\sigma x_{\sigma'}$$

- Physicists are interested in computing $\mathbb{E}[U(X)]$ and Z_β.

Example: Ising Model



Sample from an Ising model for $m = 250$.

Bayesian Inference

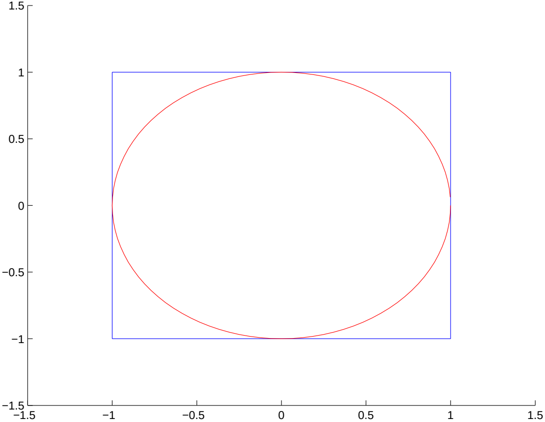
- Suppose (X, Y) are both continuous with a joint density $f_{X,Y}(x, y)$.
- We have
$$f_{X,Y}(x, y) = f_X(x)\ f_{Y|X}(y|x)$$
where, in many statistics problems, $f_X(x)$ can be thought of as a prior and $f_{Y|X}(y|x)$ as a likelihood function for a given $Y = y$.
- Using Bayes' rule, we have
$$f_{X|Y}(x|y) = \frac{f_X(x)\ f_{Y|X}(y|x)}{f_Y(y)}.$$
- For most problems of interest, $f_{X|Y}(x|y)$ does not admit an analytic expression and we cannot compute
$$\mathbb{E}\left(\phi(X)|Y = y\right) = \int \phi(x) f_{X|Y}(x|y) dx.$$

Monte Carlo Integration

- Monte Carlo methods can be thought of as a stochastic way to approximate integrals.
- Let $X_1, ..., X_n$ be a sample of independent copies of X and build the estimator
$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i),$$
for the expectation
$$\mathbb{E}\left(\phi(X)\right).$$
- **Monte Carlo algorithm**
  - Simulate independent $X_1, ..., X_n$ from F.
  - Return $\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$.

Computing Pi with Monte Carlo Methods

- Consider the 2×2 square, say $\mathcal{S} \subseteq \mathbb{R}^2$ with inscribed disk $\mathcal{D}$ of radius 1.



A 2×2 square $\mathcal{S}$ with inscribed disk $\mathcal{D}$ of radius 1.

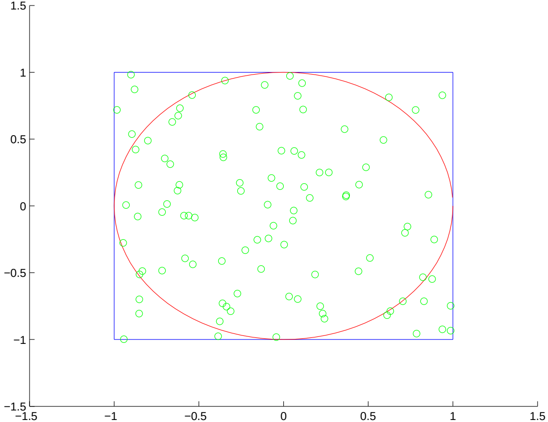
Computing Pi with Monte Carlo Methods

- We have
$$\frac{\int\int_{\mathcal{D}} dx_1 dx_2}{\int\int_{\mathcal{S}} dx_1 dx_2} = \frac{\pi}{4}.$$
- How could you estimate this quantity through simulation?
$$\begin{aligned}\frac{\int\int_{\mathcal{D}} dx_1 dx_2}{\int\int_{\mathcal{S}} dx_1 dx_2} &= \int\int_{\mathcal{S}} \mathbb{I}\left((x_1, x_2) \in \mathcal{D}\right) \frac{1}{4} dx_1 dx_2 \\ &= \mathbb{E}\left[\phi(X_1, X_2)\right] = \theta\end{aligned}$$
where the expectation is w.r.t. the uniform distribution on $\mathcal{S}$ and
$$\phi(X_1, X_2) = \mathbb{I}\left((X_1, X_2) \in \mathcal{D}\right).$$
- To sample uniformly on $\mathcal{S} = (-1, 1) \times (-1, 1)$ then simply use
$$X_1 = 2U_1 - 1,\ X_2 = 2U_2 - 1$$
where $U_1, U_2 \sim \mathcal{U}(0, 1)$.

Computing Pi with Monte Carlo Methods

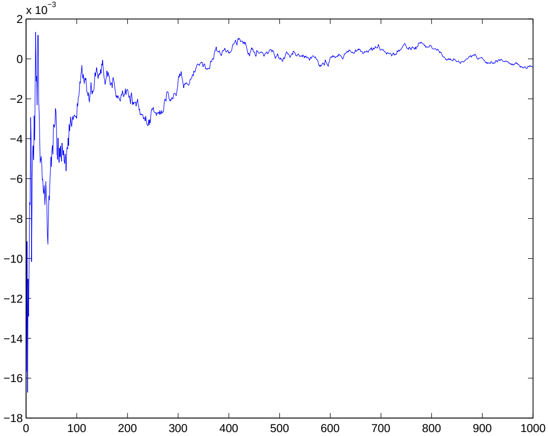
```
n <- 1000
x <- array(0, c(2,1000))
t <- array(0, c(1,1000))

for (i in 1:1000) {
  # generate point in square
  x[1,i] <- 2*runif(1)-1
  x[2,i] <- 2*runif(1)-1

  # compute phi(x); test whether in disk
  if (x[1,i]*x[1,i] + x[2,i]*x[2,i] <= 1) {
    t[i] <- 1
  } else {
    t[i] <- 0
  }
}
print(sum(t)/n*4)
```

Computing Pi with Monte Carlo Methods



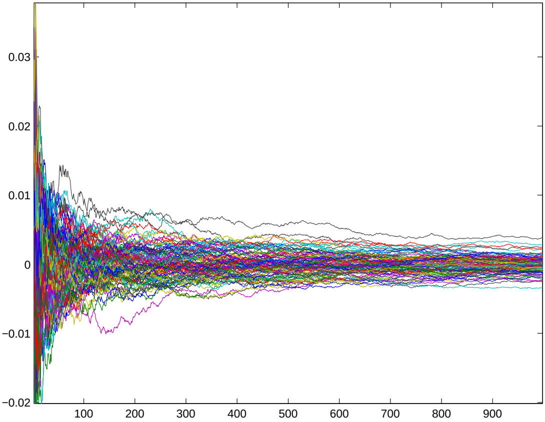
A 2×2 square $\mathcal{S}$ with inscribed disk $\mathcal{D}$ of radius 1 and Monte Carlo samples.

Computing Pi with Monte Carlo Methods



$\widehat{\theta}_n - \theta$ as a function of the number of samples n.

Computing Pi with Monte Carlo Methods



$\widehat{\theta}_n - \theta$ as a function of the number of samples n, 100 independent realizations.

Monte Carlo Integration: Law of Large Numbers

- **Proposition**: Assume $\theta = \mathbb{E}(\phi(X))$ exists then $\widehat{\theta}_n$ is an unbiased and consistent estimator of θ.
- *Proof*: We have

$$\mathbb{E}\left(\widehat{\theta}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\phi(X_i)) = \theta.$$

Weak (or strong) consistency is a consequence of the weak (or strong) law of large numbers applied to $Y_i = \phi(X_i)$ which is applicable as $\theta = \mathbb{E}(\phi(X))$ is assumed to exist.

Applications

- *Toy example*: simulate a large number n of independent r.v. $X_i \sim \mathcal{N}(\mu, \Sigma)$ and

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(\sum_{k=1}^{d} X_{k,i}^2 \geq \alpha \right).$$

- *Queuing*: simulate a large number n of days using your stochastic models for the arrival process of customers and for the service time and compute

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(X_i = 0\right)$$

where X_i is the number of customers in the shop at 5.30pm for ith sample.

- *Particle in Random Medium*: simulate a large number n of particle paths $(X_{1,i}, X_{2,i}, ..., X_{T,i})$ where $i = 1, ..., n$ and compute

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} G(X_{1,i}) G(X_{2,i}) \cdots G(X_{T,i})$$

Monte Carlo Integration: Central Limit Theorem

- **Proposition**: Assume $\theta = \mathbb{E}(\phi(X))$ and $\sigma^2 = \mathbb{V}(\phi(X))$ exist then

$$\mathbb{E}\left((\widehat{\theta}_n - \theta)^2\right) = \mathbb{V}\left(\widehat{\theta}_n\right) = \frac{\sigma^2}{n}$$

  and

$$\frac{\sqrt{n}}{\sigma}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{D} \mathcal{N}(0,1).$$

- Proof. We have $\mathbb{E}\left((\widehat{\theta}_n - \theta)^2\right) = \mathbb{V}\left(\widehat{\theta}_n\right)$ as $\mathbb{E}\left(\widehat{\theta}_n\right) = \theta$ and

$$\mathbb{V}\left(\widehat{\theta}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(\phi(X_i)) = \frac{\sigma^2}{n}.$$

  The CLT applied to $Y_i = \phi(X_i)$ tells us that

$$\frac{Y_1 + \cdots + Y_n - n\theta}{\sigma\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1)$$

  so the result follows as $\widehat{\theta}_n = \frac{1}{n}(Y_1 + \cdots + Y_n)$.

Monte Carlo Integration: Variance Estimation

- **Proposition**: Assume $\sigma^2 = \mathbb{V}\left(\phi(X)\right)$ exists then

$$S^2_{\phi(X)} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\phi(X_i) - \widehat{\theta}_n\right)^2$$

is an unbiased sample variance estimator of σ^2.

- Proof. Let $Y_i = \phi(X_i)$ then we have

$$\begin{aligned}
\mathbb{E}\left(S^2_{\phi(X)}\right) &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left(\left(Y_i - \overline{Y}\right)^2\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n} Y_i^2 \quad - n\overline{Y}^2\right) \\
&= \frac{n\left(\mathbb{V}\left(Y\right) + \theta^2\right) - n\left(\mathbb{V}\left(\overline{Y}\right) + \theta^2\right)}{n-1} \\
&= \mathbb{V}\left(Y\right) = \mathbb{V}\left(\phi(X)\right).
\end{aligned}$$

where $Y = \phi(X)$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

How Good is The Estimator?

- Chebyshev's inequality yields the bound

$$\mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| > c\frac{\sigma}{\sqrt{n}}\right) \leq \frac{\mathbb{V}\left(\widehat{\theta}_n\right)}{c^2\sigma^2/n} = \frac{1}{c^2}.$$

- Another estimate follows from the CLT for large n

$$\frac{\sqrt{n}}{\sigma}\left(\widehat{\theta}_n - \theta\right) \approx \mathcal{N}(0,1) \Rightarrow \mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| > c\frac{\sigma}{\sqrt{n}}\right) \approx 2\left(1 - \Phi(c)\right).$$

- Hence by choosing $c = c_\alpha$ s.t. $2\left(1 - \Phi(c_\alpha)\right) = \alpha$, an approximate $(1-\alpha)100\%$-CI for θ is

$$\left(\widehat{\theta}_n \pm c_\alpha \frac{\sigma}{\sqrt{n}}\right) \approx \left(\widehat{\theta}_n \pm c_\alpha \frac{S_{\phi(X)}}{\sqrt{n}}\right).$$

Monte Carlo Integration

- Whatever being Ω; e.g. $\Omega = \mathbb{R}$ or $\Omega = \mathbb{R}^{1000}$, the error is still in $\sigma/\sqrt{n}$.
- This is in contrast with deterministic methods. The error in a product trapezoidal rule in d dimensions is $\mathcal{O}(n^{-2/d})$ for twice continuously differentiable integrands.
- It is sometimes said erroneously that it beats the curse of dimensionality but this is generally not true as σ^2 typically depends of $\dim(\Omega)$.

Mathematical "Formulation"

- From a mathematical point of view, the aim of the game is to be able to generate complicated random variables and stochastic models.
- Henceforth, we will assume that we have access to a sequence of independent random variables $(U_i, i \geq 1)$ that are uniformly distributed on $(0,1)$; i.e. $U_i \sim \mathcal{U}[0,1]$.
- In R, the command `u←runif(100)` return 100 realizations of uniform r.v. in $(0,1)$.
- Strictly speaking, we only have access to pseudo-random (deterministic) numbers.
- The behaviour of modern random number generators (constructed on number theory $N_{i+1} = (aN_i + c) \bmod m$ for suitable a, c, m and $U_{i+1} = N_{i+1}/(m+1)$) resembles mathematical random numbers in many respects. Standard tests for uniformity, independence, etc. do not show significant deviations.

Outline

Generating Random Variables Using Inversion

- A function $F : \mathbb{R} \to [0, 1]$ is a cumulative distribution function (cdf) if
  - F is increasing; i.e. if $x \leq y$ then $F(x) \leq F(y)$
  - F is right continuous; i.e. $F(x + \epsilon) \to F(x)$ as $\epsilon \to 0$ ($\epsilon > 0$)
  - $F(x) \to 0$ as $x \to -\infty$ and $F(x) \to 1$ as $x \to +\infty$.
- A random variable $X : \Omega \to \mathbb{R}$ has cdf F if $\mathbb{P}(X \leq x) = F(x)$ for all $x \in \mathbb{R}$.
- If F is differentiable on $\mathbb{R}$, with derivative f, then X is continuously distributed with probability density function (pdf) f.

Generating Random Variables Using Inversion

- **Proposition**. Let F be a continuous and strictly increasing cdf on $\mathbb{R}$, we can define its inverse $F^{-1} : [0,1] \to \mathbb{R}$. Let $U \sim \mathcal{U}[0,1]$ then $X = F^{-1}(U)$ has cdf F.
- Proof. We have

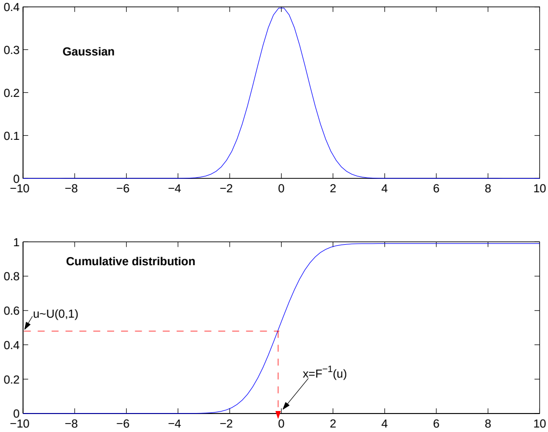
$$\begin{aligned} \mathbb{P}(X \le x) &= \mathbb{P}\left(F^{-1}(U) \le x\right) \\ &= \mathbb{P}(U \le F(x)) \\ &= F(x). \end{aligned}$$

- **Proposition**. Let F be a cdf on $\mathbb{R}$ and define its generalized inverse $F^{-1} : [0,1] \to \mathbb{R}$,

$$F^{-1}(u) = \inf \{x \in \mathbb{R}; F(x) \ge u\}.$$

Let $U \sim \mathcal{U}[0,1]$ then $X = F^{-1}(U)$ has cdf F.

Illustration of the Inversion Method



Top: pdf of a normal, bottom: associated cdf.

Examples

- *Weibull distribution*. Let $\alpha, \lambda > 0$ then the Weibull cdf is given by

$$F(x) = 1 - \exp\left(-\lambda x^{\alpha}\right), \ x \geq 0.$$

We calculate

$$\begin{aligned} u &= F(x) \Leftrightarrow \log\left(1-u\right) = -\lambda x^{\alpha} \\ &\Leftrightarrow x = \left(-\frac{\log\left(1-u\right)}{\lambda}\right)^{1/\alpha}. \end{aligned}$$

- As $(1-U) \sim \mathcal{U}[0,1]$ when $U \sim \mathcal{U}[0,1]$ we can use

$$X = \left(-\frac{\log U}{\lambda}\right)^{1/\alpha}.$$

Examples

- *Cauchy distribution*. It has pdf and cdf

$$f(x) = \frac{1}{\pi\left(1 + x^2\right)}, \; F(x) = \frac{1}{2} + \frac{arc\tan x}{\pi}$$

  We have

$$\begin{aligned} u &= F(x) \Leftrightarrow u = \frac{1}{2} + \frac{arc\tan x}{\pi} \\ &\Leftrightarrow x = \tan\left(\pi\left(u - \frac{1}{2}\right)\right) \end{aligned}$$

- *Logistic distribution.* It has pdf and cdf

$$\begin{aligned} f(x) &= \frac{\exp(-x)}{\left(1 + \exp(-x)\right)^2}, \; F(x) = \frac{1}{1 + \exp(-x)} \\ &\Leftrightarrow x = \log\left(\frac{u}{1 - u}\right). \end{aligned}$$

- Practice: Derive an algorithm to simulate from an Exponential random variable with rate $\lambda > 0$.

Generating Discrete Random Variables Using Inversion

- If X is a discrete $\mathbb{N}-$r.v. with $\mathbb{P}(X = n) = p(n)$, we get $F(x) = \sum_{j=0}^{\lfloor x \rfloor} p(j)$ and $F^{-1}(u)$ is $x \in \mathbb{N}$ such that

$$\sum_{j=0}^{x-1} p(j) < u < \sum_{j=0}^{x} p(j)$$

  with the LHS$= 0$ if $x = 0$.
- Note: the mapping at the values $F(n)$ are irrelevant.
- Note: the same method is applicable to any discrete valued r.v. X, $\mathbb{P}(X = x_n) = p(n)$.

Example: Geometric Distribution

- If $0 < p < 1$ and $q = 1 - p$ and we want to simulate $X \sim Geom(p)$ then
$$p(x) = pq^{x-1}, F(x) = 1 - q^x \quad x = 1, 2, 3...$$
- The smallest $x \in \mathbb{N}$ giving $F(x) \geq u$ is the smallest $x \geq 1$ satisfying
$$x \geq \log(1 - u)/\log(q)$$
and this is given by
$$x = F^{-1}(u) = \left\lceil \frac{\log(1 - u)}{\log(q)} \right\rceil$$
where $\lceil x \rceil$ rounds up and we could replace $1 - u$ with u.

Outline

Transformation Methods

- Suppose we have a random variable $Y \sim Q$, $Y \in \Omega_Q$, which we can simulate (eg, by inversion) and some other variable $X \sim P$, $X \in \Omega_P$, which we wish to simulate.
- Suppose we can find a function $\varphi : \Omega_Q \to \Omega_P$ with the property that $X = \varphi(Y)$.
- Then we can simulate from X by first simulating $Y \sim Q$, and then set $X = \varphi(Y)$.
- Inversion is a special case of this idea.
- We may generalize this idea to take functions of collections of variables with different distributions.

Transformation Methods

- Example: Let Y_i, $i = 1, 2, ..., \alpha$, be iid variables with $Y_i \sim Exp(1)$ and $X = \beta^{-1} \sum_{i=1}^{\alpha} Y_i$ then $X \sim Gamma(\alpha, \beta)$.
  Proof: The MGF of the random variable X is

$$\mathbb{E}\left(e^{tX}\right) = \prod_{i=1}^{\alpha} \mathbb{E}\left(e^{\beta^{-1} t Y_i}\right) = (1 - t/\beta)^{-\alpha}$$

  which is the MGF of a $\Gamma(\alpha, \beta)$ variate.
  Incidentally, the $Gamma(\alpha, \beta)$ density is $f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x > 0$.
- Practice: A generalized gamma variable Z with parameters $a > 0, b > 0, \sigma > 0$ has density

$$f_Z(z) = \frac{\sigma b^a}{\Gamma(a/\sigma)} z^{a-1} e^{-(bz)^\sigma}.$$

  Derive an algorithm to simulate from Z.

Transformation Methods: Box-Muller Algorithm

- For continuous random variables, a tool is the transformation/change of variables formula for pdf.
- **Proposition**. If $R^2 \sim Exp(\frac{1}{2})$ and $\Theta \sim \mathcal{U}[0, 2\pi]$ are independent then $X = R\cos\Theta$, $Y = R\sin\Theta$ are independent with $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$.
  Proof: We have $f_{R^2,\Theta}(r^2,\theta) = \frac{1}{2}\exp\left(-r^2/2\right)\frac{1}{2\pi}$ and

$$f_{X,Y}(x,y) = f_{R^2,\Theta}(r^2,\theta)\left|\det\frac{\partial(r^2,\theta)}{\partial(x,y)}\right|$$

where

$$\left|\det\frac{\partial(r^2,\theta)}{\partial(x,y)}\right|^{-1} = \left|\det\begin{pmatrix} \frac{\partial x}{\partial r^2} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r^2} & \frac{\partial y}{\partial \theta} \end{pmatrix}\right| = \left|\det\begin{pmatrix} \frac{\cos\theta}{2r} & -r\sin\theta \\ \frac{\sin\theta}{2r} & r\cos\theta \end{pmatrix}\right| = \frac{1}{2}.$$

Transformation Methods: Box-Muller Algorithm

- Let $U_1 \sim \mathcal{U}[0,1]$ and $U_2 \sim \mathcal{U}[0,1]$ then

$$\begin{aligned} R^2 &= -2\log(U_1) \sim Exp\left(\frac{1}{2}\right) \\ \Theta &= 2\pi U_2 \sim \mathcal{U}[0,2\pi] \end{aligned}$$

and

$$\begin{aligned} X &= R\cos\Theta \sim \mathcal{N}(0,1) \\ Y &= R\sin\Theta \sim \mathcal{N}(0,1), \end{aligned}$$

- This still requires evaluating $\log, \cos$ and $\sin$.

Simulating Multivariate Normal

- Let consider $X \in \mathbb{R}^d$, $X \sim N(\mu, \Sigma)$ where μ is the mean and Σ is the (positive definite) covariance matrix.

$$f_X(x) = (2\pi)^{-d/2} |\det \Sigma|^{-1/2} \exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right).$$

- **Proposition**. Let $Z = (Z_1, ..., Z_d)$ be a collection of d independent standard normal random variables. Let L be a real $d \times d$ matrix satisfying

$$LL^T = \Sigma,$$

and

$$X = LZ + \mu.$$

Then

$$X \sim \mathcal{N}(\mu, \Sigma).$$

Simulating Multivariate Normal

- Proof. We have $f_Z(z) = (2\pi)^{d/2} \exp\left(-\frac{1}{2} z^T z\right)$.The joint density of the new variables is

$$f_X(x) = f_Z(z) \left| \det \frac{\partial z}{\partial x} \right|$$

where $\frac{\partial z}{\partial x} = L^{-1}$ and $\det(L) = \det(L^T)$ so $\det(L^2) = \det(\Sigma)$, and $\det(L^{-1}) = 1/\det(L)$ so $\det(L^{-1}) = \det(\Sigma)^{-1/2}$. Also

$$\begin{aligned} z^T z &= (x-\mu)^T \left(L^{-1}\right)^T L^{-1} (x-\mu) \\ &= (x-\mu)^T \Sigma^{-1} (x-\mu). \end{aligned}$$

- If $\Sigma = VDV^T$ is the eigendecomposition of Σ, we can pick $L = VD^{1/2}$.
- Cholesky factorization $\Sigma = LL^T$ where L is a lower triangular matrix.
- See numerical analysis.

Outline

Rejection Sampling

- Consider X a discrete random variable on Ω with a probability mass function $p(x)$, a "target distribution"
- We want to sample from $p(x)$ using a proposal pmf $q(x)$ which we can sample.
- **Proposition**. Suppose we can find a constant M such that $p(x)/q(x) \leq M$ for all $x \in \Omega$. The following 'Rejection' algorithm returns $X \sim p$.
- **Rejection Sampling**.
  Step 1 - Simulate $Y \sim q$ and $U \sim \mathcal{U}[0,1]$, with simulated value y and u respectively.
  Step 2 - If $u \leq p(y)/q(y)/M$ then stop and return $X = y$,
  Step 3 - otherwise go back to Step 1.

Rejection Sampling: Proof 1

- We have

$$\begin{aligned}\Pr(X = x) &= \sum_{n=1}^{\infty} \Pr(\text{reject } n-1 \text{ times, draw } Y = x \text{ and accept it}) \\ &= \sum_{n=1}^{\infty} \Pr(\text{reject } Y)^{n-1} \Pr(\text{draw } Y = x \text{ and accept it})\end{aligned}$$

- We have

$$\begin{aligned}&\Pr(\text{draw } Y = x \text{ and accept it}) \\ &= \Pr(\text{draw } Y = x) \Pr(\text{accept } Y \mid Y = x) \\ &= q(x) \Pr\left(U \leq \frac{p(Y)}{q(Y)}/M \,\middle|\, Y = x \right) \\ &= \frac{p(x)}{M}\end{aligned}$$

- The probability of having a rejection is

$$\begin{aligned}\Pr(\text{reject } Y) &= \sum_{x\in\Omega} \Pr(\text{draw } Y = x \text{ and reject it}) \\ &= \sum_{x\in\Omega} q(x) \Pr\left(U \geq \frac{p(Y)}{q(Y)}/M \,\middle|\, Y = x \right) \\ &= \sum_{x\in\Omega} q(x) \left(1 - \frac{p(x)}{q(x)M} \right) = 1 - \frac{1}{M}\end{aligned}$$

- Hence we have

$$\begin{aligned}\Pr(X = x) &= \sum_{n=1}^{\infty} \Pr(\text{reject } Y)^{n-1} \Pr(\text{draw } Y = x \text{ and accept it}) \\ &= \sum_{n=1}^{\infty} \left(1 - \frac{1}{M}\right)^{n-1} \frac{p(x)}{M} \\ &= p(x).\end{aligned}$$

- Note the number of accept/reject trials has a geometric distribution of success probability $1/M$, so the mean number of trials is M.

Rejection Sampling: Proof 2

- Here is an alternative proof given for a continuous scalar variable X, the rejection algorithm still works but p, q are now pdfs.
- We accept the proposal Y whenever $(U, Y) \sim p_{U,Y}$ where $p_{U,Y}(u, y) = q(y)\mathbb{I}_{(0,1)}(u)$ satisfies $U \leq p(Y)/Mq(Y)$.
- We have

$$
\begin{aligned}
\Pr(X \leq x) &= \Pr(Y \leq x | U \leq p(Y)/Mq(Y)) \\
&= \frac{\Pr(Y \leq x, U \leq p(Y)/Mq(Y))}{\Pr(U \leq p(Y)/Mq(Y))} \\
&= \frac{\int_{-\infty}^{x} \int_{0}^{p(y)/Mq(y)} p_{U,Y}(u, y) du dy}{\int_{-\infty}^{\infty} \int_{0}^{p(y)/Mq(y)} p_{U,Y}(u, y) du dy} \\
&= \frac{\int_{-\infty}^{x} \int_{0}^{p(y)/Mq(y)} q(y) du dy}{\int_{-\infty}^{\infty} \int_{0}^{p(y)/Mq(y)} q(y) du dy} = \int_{-\infty}^{x} p(y) dy.
\end{aligned}
$$

Example: Beta Density

- Assume you have for $\alpha, \beta \geq 1$
$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1},\ 0 < x < 1$$
which is upper bounded on $[0,1]$.
- We propose to use as a proposal $q(x) = \mathbb{I}_{(0,1)}(x)$ the uniform density on $[0,1]$.
- We need to find a bound M s.t. $p(x)/Mq(x) = p(x)/M \leq 1$. The smallest M is $M = \max_{0<x<1} p(x)$ and we obtain by solving for $p'(x) = 0$
$$M = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\left(\frac{\alpha-1}{\alpha+\beta-2}\right)^{\alpha-1} \left(\frac{\beta-1}{\alpha+\beta-2}\right)^{\beta-1}}_{M'}$$
which gives
$$\frac{p(y)}{Mq(y)} = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{M'}.$$

Dealing with Unknown Normalising Constants

- In most practical scenarios, we only know $p(x)$ and $q(x)$ up to some normalising constants; i.e.

$$p(x) = \tilde{p}(x)/Z_p \text{ and } q(x) = \tilde{q}(x)/Z_q$$

  where $\tilde{p}(x), \tilde{q}(x)$ are known but $Z_p = \int_\Omega \tilde{p}(x)dx$, $Z_q = \int_\Omega \tilde{q}(x)dx$ are unknown/expensive to compute.
- Rejection can still be used: Indeed $p(x)/q(x) \leq M$ for all $x \in \Omega$ iff $\tilde{p}(x)/\tilde{q}(x) \leq M'$, with $M' = Z_p M/Z_q$.
- Practically, this means we can ignore the normalising constants from the start: if we can find M' to bound $\tilde{p}(x)/\tilde{q}(x)$ then it is correct to accept with probability $\tilde{p}(x)/M'\tilde{q}(x)$ in the rejection algorithm. In this case the mean number N of accept/reject trials will equal $Z_q M'/Z_p$ (that is, M again).

Simulating Gamma Random Variables

- We want to simulate a random variable $X \sim$Gamma(α, β) which works for any $\alpha \geq 1$ (not just integers);

$$p(x) = \frac{x^{\alpha-1}\exp(-\beta x)}{Z_p} \text{ for } x > 0, \quad Z_p = \Gamma(\alpha)/\beta^{\alpha}$$

  so $\tilde{p}(x) = x^{\alpha-1}\exp(-\beta x)$ will do as our unnormalised target.
- When $\alpha = a$ is a positive integer we can simulate $X \sim \text{Gamma}(a, \beta)$ by adding a independent $\text{Exp}(\beta)$ variables, $Y_i \sim \text{Exp}(\beta)$, $X = \sum_{i=1}^{a} Y_i$.
- Hence we can sample densities 'close' in shape to $\text{Gamma}(\alpha, \beta)$ since we can sample $\text{Gamma}(\lfloor\alpha\rfloor, \beta)$. Perhaps this, or something like it, would make an envelope/proposal density?

- Let $a = \lfloor \alpha \rfloor$ and let's try to use Gamma(a, b) as the envelope, so $Y \sim$ Gamma(a, b) for integer $a \geq 1$ and some $b > 0$. The density of Y is

$$q(x) = \frac{x^{a-1} \exp(-bx)}{Z_q} \text{ for } x > 0, \quad Z_q = \Gamma(a)/b^a$$

  so $\tilde{q}(x) = x^{a-1} \exp(-bx)$ will do as our unnormalised envelope function.
- We have to check whether the ratio $\tilde{p}(x)/\tilde{q}(x)$ is bounded over $\mathbb{R}$ where

$$\tilde{p}(x)/\tilde{q}(x) = x^{\alpha - a} \exp(-(\beta - b)x).$$

- Consider (a) $x \to 0$ and (b) $x \to \infty$. For (a) we need $a \leq \alpha$ so $a = \lfloor \alpha \rfloor$ is indeed fine. For (b) we need $b < \beta$ (not $b = \beta$ since we need the exponential to kill off the growth of $x^{\alpha - a}$).

- Given that we have chosen $a = \lfloor \alpha \rfloor$ and $b < \beta$ for the ratio to be bounded, we now compute the bound.
- $\frac{d}{dx}(\tilde{p}(x)/\tilde{q}(x)) = 0$ at $x = (\alpha - a)/(\beta - b)$ (and this must be a maximum at $x \geq 0$ under our conditions on a and b), so $\tilde{p}/\tilde{q} \leq M$ for all $x \geq 0$ if

$$M = \left(\frac{\alpha - a}{\beta - b}\right)^{\alpha - a} \exp(-(\alpha - a)).$$

- Accept Y at step 2 of Rejection Sampler if $U \leq \tilde{p}(Y)/M\tilde{q}(Y)$ where $\tilde{p}(Y)/M\tilde{q}(Y) = Y^{\alpha - a} \exp(-(\beta - b)Y)/M$.

Simulating Gamma Random Variables: Best choice of b

- Any $0 < b < \beta$ will do, but is there a best choice of b?
- Idea: choose b to minimize the expected number of simulations of Y per sample X output.
- Since the number N of trials is Geometric, with success probability $Z_p/(MZ_q)$, the expected number of trials is $\mathbb{E}(N) = Z_q M/Z_p$. Now $Z_p = \Gamma(\alpha)\beta^{-\alpha}$ where Γ is the Gamma function related to the factorial.
- Practice: Show that the optimal b solves $\frac{d}{db}(b^{-a}(\beta - b)^{-\alpha+a}) = 0$ so deduce that $b = \beta(a/\alpha)$ is the optimal choice.

Simulating Normal Random Variables

- Let $p(x) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2)$ and $q(x) = 1/\pi(1 + x^2)$. We have

$$\frac{\tilde{p}(x)}{\tilde{q}(x)} = (1 + x^2) \exp\left(-\frac{1}{2}x^2\right) \leq 2/\sqrt{e} = M$$

  which is attained at ± 1.

- Hence the probability of acceptance is

$$\mathbb{P}\left(U \leq \frac{\tilde{p}(x)}{M\tilde{q}(x)}\right) = \frac{Z_p}{MZ_q} = \frac{\sqrt{2\pi}}{\frac{2}{\sqrt{e}}\pi} = \sqrt{\frac{e}{2\pi}} \approx 0.66$$

  and the mean number of trials to success is approximately $1/0.66 \approx 1.52$.

Rejection Sampling in High Dimension

- Consider

$$\tilde{p}(x_1, ..., x_d) = \exp\left(-\frac{1}{2}\sum_{k=1}^{d} x_k^2\right)$$

and

$$\tilde{q}(x_1, ..., x_d) = \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{d} x_k^2\right)$$

- For $\sigma > 1$, we have

$$\frac{\tilde{p}(x_1, ..., x_d)}{\tilde{q}(x_1, ..., x_d)} = \exp\left(-\frac{1}{2}\left(1-\sigma^{-2}\right)\sum_{k=1}^{d} x_k^2\right) \leq 1 = M.$$

- The acceptance probability of a proposal for $\sigma > 1$ is

$$\mathbb{P}\left(U \leq \frac{\tilde{p}(X_1, ..., X_d)}{M\tilde{q}(X_1, ..., X_d)}\right) = \frac{Z_p}{MZ_q} = \sigma^{-d}.$$

- The acceptance probability goes exponentially fast to zero with d.

Outline

Importance Sampling

- Importance sampling (IS) can be thought, among other things, as a strategy for recycling samples.
- It is also useful when we need to make an accurate estimate of the probability that a random variable exceeds some very high threshold.
- In this context it is referred to as a *variance reduction* technique.
- There is a slight variation on the basic set up: we can generate samples from q but we want to estimate an expectation $\mathbb{E}_p(f(X))$ of a function f under p.
  (Previously, it was "we want samples distributed according to p".)
- In IS, we avoid sampling the target distribution p. Instead, we take samples distributed according to q and *reweight* them.

Importance Sampling Identity

- **Proposition**. Let q and p be pdf on Ω. Assume $p(x) > 0 \Rightarrow q(x) > 0$, then for any function $\phi : \Omega \to \mathbb{R}$ we have
$$\mathbb{E}_p(\phi(X)) = \mathbb{E}_q(\phi(X)w(X))$$
where $w : \Omega \to \mathbb{R}^+$ is the importance weight function
$$w(x) = \frac{p(x)}{q(x)}.$$
- Proof: We have
$$\begin{aligned}\mathbb{E}_p(\phi(X)) &= \int_\Omega \phi(x)p(x)dx \\ &= \int_\Omega \phi(x)\frac{p(x)}{q(x)}q(x)dx \\ &= \int_\Omega \phi(x)w(x)q(x)dx \\ &= \mathbb{E}_q(\phi(X)w(X)).\end{aligned}$$
- Similar proof holds in the discrete case.

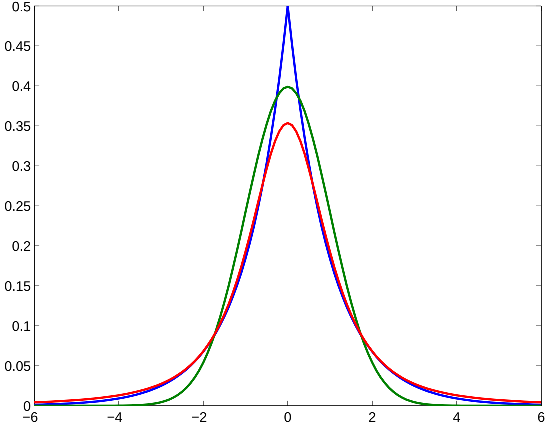
Importance Sampling Estimator

- **Proposition**. Let q and p be pdf on Ω. Assume $p(x)\phi(x) \neq 0 \Rightarrow q(x) > 0$ and let $\phi : \Omega \rightarrow \mathbb{R}$ such that $\theta = \mathbb{E}_p(\phi(X))$ exists. Let $Y_1, ..., Y_n$ be a sample of independent random variables distributed according to q then the estimator

$$\widehat{\theta}_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i) w(Y_i)$$

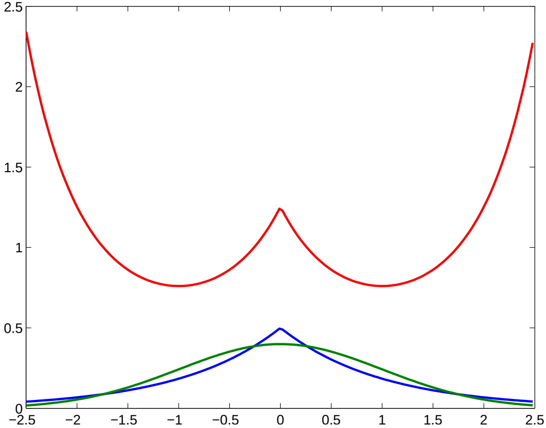
  is unbiased and consistent.
- Proof. Unbiasedness follows directly from $\mathbb{E}_p(\phi(X)) = \mathbb{E}_q(\phi(Y_i)w(Y_i))$ and $Y_i \sim q$. Weak (or strong) consistency is a consequence of the weak (or strong) law of large numbers applied to $Z_i = \phi(Y_i)w(Y_i)$ which is applicable as θ is assumed to exist.

Target and Proposal Distributions



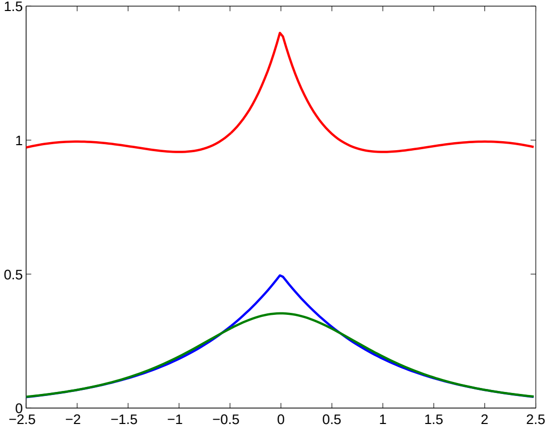
Target double exponential distributions and two IS distributions (normal and student-t).

Weight Function



Weight function evaluated at the Normal IS random variables realizations.

Weight Function



Weight function evaluated at the Student-t IS random variables realizations.

Example: Gamma Distribution

- Say we have simulated $Y_i \sim$Gamma(a, b) and we want to estimate $\mathbb{E}_p(\phi(X))$ where $X \sim$Gamma(α, β).
- Recall that the Gamma(α, β) density is

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$

so

$$w(x) = \frac{p(x)}{q(x)} = \frac{\Gamma(a)\beta^\alpha}{\Gamma(\alpha) b^a} x^{\alpha-a} e^{-(\beta-b)x}$$

- Hence

$$\widehat{\theta}_n^{\text{IS}} = \frac{\Gamma(a)\beta^\alpha}{\Gamma(\alpha) b^a} \; \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i) \;\; {Y_i}^{\alpha-a} e^{-(\beta-b)Y_i}$$

is an unbiased and consistent estimate of $\mathbb{E}_p(\phi(X))$.

Variance of the Importance Sampling Estimator

- **Proposition**. Assume $\theta = \mathbb{E}_p(\phi(X))$ and $\mathbb{E}_p(w(X)\phi^2(X))$ are finite. Then $\widehat{\theta}_n^{\text{IS}}$ satisfies

$$\begin{aligned}
\mathbb{E}\left(\left(\widehat{\theta}_n^{\text{IS}} - \theta\right)^2\right) &= \mathbb{V}\left(\widehat{\theta}_n^{\text{IS}}\right) = \tfrac{1}{n}\mathbb{V}_q\left(w(Y_1)\phi(Y_1)\right) \\
&= \tfrac{1}{n}\left(\mathbb{E}_q\left(\tfrac{p^2(Y_1)}{q^2(Y_1)}\phi^2(Y_1)\right) - \mathbb{E}_q\left(\tfrac{p(Y_1)}{q(Y_1)}\phi(Y_1)\right)^2\right) \\
&= \tfrac{1}{n}\left(\mathbb{E}_p\left(w(X)\phi^2(X)\right) - \theta^2\right).
\end{aligned}$$

- Each time we do IS we should check that this variance is finite, otherwise our estimates are somewhat untrustworthy! We check $\mathbb{E}_p(w\phi^2)$ is finite.

Example: Gamma Distribution

- Let us check that the variance of $\widehat{\theta}_n^{\mathrm{IS}}$ in previous Example is finite if $\theta = \mathbb{E}_p(\phi(X))$ and $\mathbb{V}_p(\phi(X))$ are finite.
- It is enough to check that $\mathbb{E}_p\left(w(Y_1)\phi^2(Y_1)\right)$ is finite.
- The normalisation constants are finite so we can ignore those, and begin with

$$w(x)\phi^2(x) \propto x^{\alpha-a}e^{-(\beta-b)X}\phi^2(x).$$

- The expectation of interest is

$$\begin{aligned}
&\mathbb{E}_p\left(w(X)\phi^2(X)\right) \\
\propto&\mathbb{E}_p\left(X^{\alpha-a}e^{-(\beta-b)X}\phi^2(X)\right) \\
=&\int_0^\infty p(x)\, x^{\alpha-a}\exp(-(\beta-b)x))\phi^2(x)\, dx \\
\leq&M\int_0^\infty p(x)\phi(x)^2\, dx = M\mathbb{E}_p(\phi^2).
\end{aligned}$$

where $M = \max_{x>0} x^{\alpha-a}\exp(-(\beta-b)x)$ is finite if $a < \alpha$ and $b < \beta$ (see rejection sampling section).

- Since $\theta = \mathbb{E}_p(\phi(X))$ and $\mathbb{V}_p(\phi(X))$ are finite, we have $\mathbb{E}_p(\phi^2(X)) < \infty$ if these conditions on a, b are satisfied. If not, we cannot conclude as it depends on ϕ.
- These same (sufficient) conditions apply to our rejection sampler for Gamma(α, β).
- For IS it is enough just for M to exist—we do not have to work out its value.

Choice of the Importance Sampling Distribution

- While p is given, q needs to cover $p\phi$ (i.e. $p(x)\phi(x) \neq 0 \Rightarrow q(x) > 0$) and be simple to sample.
- The requirement $\mathbb{V}\left(\widehat{\theta}_n^{\mathrm{IS}}\right) < \infty$ further constrains our choice: we need $\mathbb{E}_p\left(w(X)\phi^2(X)\right) < \infty$.
- If $\mathbb{V}_p(\phi(X))$ is known finite then, it may be easy to get a sufficient condition for $\mathbb{E}_p\left(w(X)\phi^2(X)\right) < \infty$; e.g. $w(x) \leq M$. Further analysis will depend on ϕ.
- What is the choice q_{opt} of q that actually minimizes the variance of the IS estimator? Consider $\phi : \Omega \to \mathbb{R}^+$ then

$$q_{\mathrm{opt}}(x) = \frac{p(x)\phi(x)}{\mathbb{E}_p(\phi(X))} \;\Rightarrow\; \mathbb{V}(\widehat{\theta}_n^{\mathrm{IS}}) = 0.$$

- This optimal zero-variance estimator cannot be implemented as

$$w(x) = p(x)/q_{\mathrm{opt}}(x) = \mathbb{E}_p(\phi(X))/\phi(x)$$

where $\mathbb{E}_p(\phi(X))$ is the thing we are trying to estimate! This can however be used as a guideline to select q.

Importance Sampling for Rare Event Estimation

- One important class of applications of IS is to problems in which we estimate the probability for a rare event.
- In such scenarios, we may be able to sample from p directly but this does not help us. If, for example, $X \sim p$ with $\mathbb{P}(X > x_0) = \mathbb{E}_p\left(\mathbb{I}[X > x_0]\right) = \theta$ say, with $\theta \ll 1$, we may not get any samples $X_i > x_0$ and our estimate $\widehat{\theta}_n = \sum_i \mathbb{I}(X_i > x_0)/n$ is simply zero.
- Generally, we have

$$\mathbb{E}\left(\widehat{\theta}_n\right) = \theta,\ \mathbb{V}\left(\widehat{\theta}_n\right) = \frac{\theta(1-\theta)}{n}$$

but the relative variance

$$\frac{\mathbb{V}\left(\widehat{\theta}_n\right)}{\theta^2} = \frac{(1-\theta)}{\theta n} \overset{\theta \to 0}{\to} \infty.$$

- By using IS, we can actually reduce the variance of our estimator.

Importance Sampling for Rare Event Estimation

- Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a scalar normal random variable and we want to estimate $\theta = \mathbb{P}(X > x_0)$ for some $x_0 \gg \mu + 3\sigma$. We can *exponentially tilt* the pdf of X towards larger values so that we get some samples in the target region, and then correct for our tilting via IS.
- If $p(x)$ is pdf of X then $q(x) = p(x)e^{tx}/M_p(t)$ is called an *exponentially tilted* version of p where $M_p(t) = \mathbb{E}_p(e^{tX})$ is the moment generating function of X.
- For $p(x)$ the normal density,

$$q(x) \propto e^{-(x-\mu)^2/2\sigma^2} e^{tx} = e^{-(x-\mu-t\sigma^2)^2/2\sigma^2} e^{\mu t + t^2\sigma^2/2}$$

so we have

$$q(x) = \mathcal{N}(x; \mu + t\sigma^2, \sigma^2), \quad M_p(t) = e^{\mu t + t^2\sigma^2/2}.$$

Importance Sampling for Rare Event Estimation

- The IS weight function is $p(x)/q(x) = e^{-tx} M_p(t)$ so

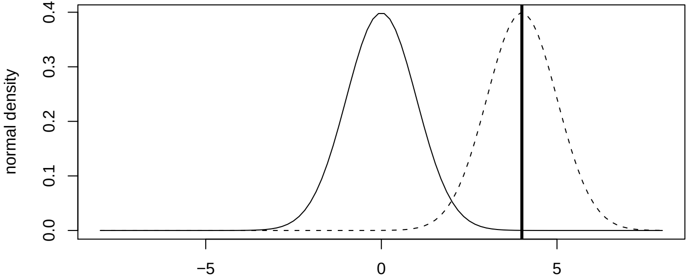
$$w(x) = e^{-t(x-\mu-t\sigma^2/2)}.$$

- We take samples $Y_i \sim \mathcal{N}(\mu + t\sigma^2, \sigma^2)$, compute $w_i = w(Y_i)$ and form our IS estimator for $\theta = \mathbb{P}(X > x_0)$

$$\widehat{\theta}_n^{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} w_i \mathbb{I}_{Y_i > x_0}$$

  since $\phi(Y_i) = \mathbb{I}_{Y_i > x_0}$.
- We have not said how to choose t. The point here is that we want samples in the region of interest. We choose the mean of the tiled distribution so that it equals x_0, this ensure we have samples in the region of interest; that is $\mu + t\sigma^2 = x_0$, or $t = (x_0 - \mu)/\sigma^2$.

Original and Exponentially Tilt Densities



(solid) $\mathcal{N}(0,1)$ density p. (i.e. $\mu = 0$, $\sigma = 1$) (dashed) $\mathcal{N}(x_0,1)$ tilted density q.

Optimal Tilt Densities

- We selected t such that $\mu + t\sigma^2 = x_0$ somewhat heuristically.
- In practice, we might be interested in selecting the t value which minimizes the variance of $\widehat{\theta}_n^{\text{IS}}$ where

$$\begin{aligned}\mathbb{V}(\widehat{\theta}_n^{\text{IS}}) &= \frac{1}{n}\left(\mathbb{E}_p\left(w(X)\mathbb{I}_{X>x_0}\right) - \mathbb{E}_p\left(\mathbb{I}_{X>x_0}\right)^2\right) \\ &= \frac{1}{n}\left(\mathbb{E}_p\left(w(X)\mathbb{I}_{X>x_0}\right) - \theta^2\right).\end{aligned}$$

- Hence we need to minimize $\mathbb{E}_p\left(w(X)\mathbb{I}_{X>x_0}\right)$ w.r.t t where

$$\begin{aligned}\mathbb{E}_p\left(w(X)\mathbb{I}_{X>x_0}\right) &= \int_{x_0}^{\infty} p(x)e^{-t(x-\mu-t\sigma^2/2)}dx \\ &= M_p(t)\int_{x_0}^{\infty} p(x)e^{-tx}dx\end{aligned}$$

Normalised Importance Sampling

- In most practical scenarios,

$$p(x) = \tilde{p}(x)/Z_p \text{ and } q(x) = \tilde{q}(x)/Z_q$$

where $\tilde{p}(x), \tilde{q}(x)$ are known but $Z_p = \int_\Omega \tilde{p}(x)dx$, $Z_q = \int_\Omega \tilde{q}(x)dx$ are unknown or difficult to compute.

- The previous IS estimator is not applicable as it requires evaluating $w(x) = p(x)/q(x)$.
- An alternative IS estimator can be proposed based on the following alternative IS identity.
- **Proposition**. Let q and p be pdf on Ω. Assume $p(x) > 0 \Rightarrow q(x) > 0$, then for any function $\phi : \Omega \to \mathbb{R}$ we have

$$\mathbb{E}_p(\phi(X)) = \frac{\mathbb{E}_q(\phi(X)\tilde{w}(X))}{\mathbb{E}_q(\tilde{w}(X))}$$

where $\tilde{w} : \Omega \to \mathbb{R}^+$ is the importance weight function

$$\tilde{w}(x) = \tilde{p}(x)/\tilde{q}(x).$$

Normalised Importance Sampling

- Proof: We have

$$\begin{aligned}
\mathbb{E}_p(\phi(X)) &= \int_\Omega \phi(x)p(x)dx \\
&= \frac{\int_\Omega \phi(x)\frac{p(x)}{q(x)}q(x)dx}{\int_\Omega \frac{p(x)}{q(x)}q(x)dx} \\
&= \frac{\int_\Omega \phi(x)\tilde{w}(x)q(x)dx}{\int_\Omega \tilde{w}(x)q(x)dx} \\
&= \frac{\mathbb{E}_q(\phi(X)\tilde{w}(X))}{\mathbb{E}_q(\tilde{w}(X))}.
\end{aligned}$$

- Remark: Even if we are interested in a simple function ϕ, we do need $p(x) > 0 \Rightarrow q(x) > 0$ to hold instead of $p(x)\phi(x) \neq 0 \Rightarrow q(x) > 0$ for the previous IS identity.

Importance Sampling Pseudocode

1. **Inputs:**
   - Function to draw samples from p
   - Function $\tilde{w}(x) = \tilde{p}(x)/\tilde{q}(x)$
   - Function ϕ
   - Number of samples n
2. **For** $i = 1, \ldots, n$:
   2.1 Draw $Y_i \sim \tilde{p}$.
   2.2 Compute $w_i = \tilde{w}(Y_i)$.
3. **Return**

$$\frac{\sum_{i=1}^{n} w_i \phi(Y_i)}{\sum_{i=1}^{n} w_i}.$$

Normalised Importance Sampling Estimator

- **Proposition**. Let q and p be pdf on Ω. Assume $p(x) > 0 \Rightarrow q(x) > 0$ and let $\phi : \Omega \to \mathbb{R}$ such that $\theta = \mathbb{E}_p(\phi(X))$ exists. Let $Y_1, ..., Y_n$ be a sample of independent random variables distributed according to q then the estimator

$$\widehat{\theta}_n^{\mathrm{NIS}} = \frac{\frac{1}{n}\sum_{i=1}^n \phi(Y_i)\tilde{w}(Y_i)}{\frac{1}{n}\sum_{i=1}^n \tilde{w}(Y_i)} = \frac{\sum_{i=1}^n \phi(Y_i)\tilde{w}(Y_i)}{\sum_{i=1}^n \tilde{w}(Y_i)}$$

  is consistent.
- Remark: It is easy to show that $\widehat{A}_n = \frac{1}{n}\sum_{i=1}^n \phi(Y_i)\tilde{w}(Y_i)$ (resp. $\widehat{B}_n = \frac{1}{n}\sum_{i=1}^n \tilde{w}(Y_i)$) is an unbiased and consistent estimator of $A = \mathbb{E}_q\left(\phi(Y)\tilde{w}(Y)\right)$ (resp. $B = \mathbb{E}_q\left(\tilde{w}(Y)\right)$). However $\widehat{\theta}_n^{\mathrm{NIS}}$, which is a ratio of estimates, is biased for finite n.

Normalised Importance Sampling Estimator

- Proof strong consistency (not examinable). The strong law of large numbers yields

$$\lim \mathbb{P}\left(\widehat{A}_n \to A\right) = \lim \mathbb{P}\left(\widehat{B}_n \to B\right) = 1$$

This implies

$$\lim \mathbb{P}\left(\widehat{A}_n \to A, \widehat{B}_n \to B\right) = 1$$

and

$$\lim \mathbb{P}\left(\frac{\widehat{A}_n}{\widehat{B}_n} \to \frac{A}{B}\right) = 1.$$

Normalised Importance Sampling Estimator

- Proof weak consistency (not examinable). The weak law of large numbers states that for any $\varepsilon > 0$ and $\delta > 0$, there exists $n_0 \geq 0$ such that for all $n \geq n_0$: $\mathbb{P}\left(\left|\widehat{B}_n - B\right| > \frac{B}{2}\right) < \frac{\delta}{3}$, $\mathbb{P}\left(\left|\widehat{A}_n - A\right| > \frac{\varepsilon B}{2}\right) < \frac{\delta}{3}$, $\mathbb{P}\left(A\left|\widehat{B}_n - B\right| > \frac{\varepsilon B^2}{4}\right) < \frac{\delta}{3}$. Then, we also have for all $n \geq n_0$

$$
\begin{aligned}
&\mathbb{P}\left(\left|\frac{\widehat{A}_n}{\widehat{B}_n} - \frac{A}{B}\right| > \varepsilon\right) \leq \mathbb{P}\left(\left|\widehat{B}_n - B\right| > \frac{B}{2}\right) \\
&+\mathbb{P}\left(\left|\widehat{B}_n - B\right| \leq \frac{B}{2}, \left|\widehat{A}_n B - A\widehat{B}_n\right| > \varepsilon \widehat{B}_n B\right) \\
&< \frac{\delta}{3} + \mathbb{P}\left(\left|\widehat{A}_n B - AB\right| > \frac{\varepsilon B^2}{4}\right) + \mathbb{P}\left(\left|AB - A\widehat{B}_n\right| > \frac{\varepsilon B^2}{4}\right) < \delta
\end{aligned}
$$

where the middle step use $\widehat{B}_n > B/2$, and

$$
\begin{aligned}
&\mathbb{P}\left(\left|\widehat{A}_n B - A\widehat{B}_n\right| > \frac{\varepsilon B^2}{2}\right) \leq \mathbb{P}\left(\left|\widehat{A}_n B - AB\right| > \frac{\varepsilon B^2}{4}\right) \\
&+ \mathbb{P}\left(\left|AB - A\widehat{B}_n\right| > \frac{\varepsilon B^2}{4}\right).
\end{aligned}
$$

Example Revisited: Gamma Distribution

- We are interested in estimating $\mathbb{E}_p\left(\phi(X)\right)$ where $X \sim$Gamma(α, β) using samples from a Gamma(a, b) distribution; i.e.

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad q(x) = \frac{b^a}{\Gamma(a)} e^{-bx}$$

- Suppose we do not remember the expression of the normalising constant for the Gamma, so that we use

$$\begin{aligned} \tilde{p}(x) &= x^{\alpha-1} e^{-\beta x},\ \tilde{q}(x) = x^{a-1} e^{-bx} \\ &\Rightarrow \tilde{w}(x) = x^{\alpha-a} e^{-(\beta-b)x} \end{aligned}$$

- Practially, we simulate $Y_i \sim$Gamma(a, b), for $i = 1, 2, ..., n$ then compute

$$\begin{aligned} \tilde{w}(Y_i) &= Y_i^{\alpha-a} e^{-(\beta-b)Y_i}, \\ \widehat{\theta}_n^{\text{NIS}} &= \frac{\sum_{i=1}^n \phi(Y_i)\tilde{w}(Y_i)}{\sum_{i=1}^n \tilde{w}(Y_i)}. \end{aligned}$$

Importance Sampling in High Dimension

- Consider

$$\tilde{p}(x_1, ..., x_d) = \exp\left(-\tfrac{1}{2}\sum_{k=1}^{d} x_k^2\right),$$

$$\tilde{q}(x_1, ..., x_d) = \exp\left(-\tfrac{1}{2\sigma^2}\sum_{k=1}^{d} x_k^2\right).$$

- We have

$$\tilde{w}(x) = \frac{\tilde{p}(x_1, ..., x_d)}{q(x_1, ..., x_d)} = \exp\left(-\frac{1}{2}(1-\sigma^{-2})\sum_{k=1}^{d} x_k^2\right).$$

- For $Y_i \sim q$, $\widehat{B}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}(Y_i)$ is a consistent estimate of $B = \mathbb{E}_q(\tilde{w}(Y)) = Z_p/Z_q$ with for $\sigma^2 > \frac{1}{2}$

$$\mathbb{V}\left(\widehat{B}_n\right) = \frac{\mathbb{V}_q\left(\tilde{w}(Y)\right)}{n} = \frac{1}{n}\left(\frac{Z_p}{Z_q}\right)^2\left(\left(\frac{\sigma^4}{2\sigma^2-1}\right)^{d/2} - 1\right)$$

with $\sigma^4\left(2\sigma^2-1\right)^{-1} > 1$ for $\sigma^2 > \frac{1}{2}$.

Outline

Markov chain Monte Carlo Methods

- Our aim is to estimate $\mathbb{E}_p(\phi(X))$ for $p(x)$ some pmf (or pdf) defined for $x \in \Omega$.
- Up to this point we have based our estimates on iid draws from either p itself, or some proposal distribution with pmf q.
- In MCMC we simulate a correlated sequence $X_0, X_1, X_2, ...$ which satisfies $X_t \sim p$ (or at least X_t converges to p in distribution) and rely on the usual estimate $\widehat{\phi}_n = n^{-1} \sum_{t=0}^{n-1} \phi(X_t)$.
- We will suppose the space of states of X is finite (and therefore discrete) but it should be kept in mind that MCMC methods are applicable to countably infinite and continuous state spaces.

Markov chains

- Let $\{X_t\}_{t=0}^{\infty}$ be a homogeneous Markov chain of random variables on Ω with starting distribution $X_0 \sim p^{(0)}$ and transition probability

$$P_{i,j} = \mathbb{P}(X_{t+1} = j | X_t = i).$$

- Denote by $P_{i,j}^{(n)}$ the *n*-step transition probabilities

$$P_{i,j}^{(n)} = \mathbb{P}(X_{t+n} = j | X_t = i)$$

and by $p^{(n)}(i) = \mathbb{P}(X_n = i)$.

- Recall that P is *irreducible* if and only if, for each pair of states $i, j \in \Omega$ there is n such that $P_{i,j}^{(n)} > 0$. The Markov chain is *aperiodic* if $P_{i,j}^{(n)}$ is non zero for all sufficiently large n.

Markov chains

- Here is an example of a periodic chain:
  $\Omega = \{1, 2, 3, 4\}, p^{(0)} = (1, 0, 0, 0)$, and transition matrix

$$P = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix},$$

  since $P_{1,1}^{(n)} = 0$ for n odd.
- **Exercise**: show that if P is irreducible and $P_{i,i} > 0$ for some $i \in \Omega$ then P is aperiodic.

Markov chains and Reversible Markov chains

- Recall that the pmf $\pi(i), i \in \Omega$, $\sum_{i \in \Omega} \pi(i) = 1$ is a stationary distribution of P if $\pi P = \pi$. If $p^{(0)} = \pi$ then

$$p^{(1)}(j) = \sum_{i \in \Omega} p^{(0)}(i) P_{i,j},$$

  so $p^{(1)}(j) = \pi(j)$ also. Iterating, $p^{(t)} = \pi$ for each $t = 1, 2, \ldots$ in the chain, so the distribution of $X_t \sim p^{(t)}$ doesn't change with t, it is stationary.
- In a reversible Markov chain we cannot distinguish the direction of simulation from inspection of a realization of the chain (so, you simulate a piece of the chain, toss a coin and reverse the order of states if the coin comes up heads; now you present me the sequence of states; I cannot tell whether or not you have reversed the sequence, though I know the transition matrix of the chain).
- Most MCMC algorithms are based on reversible Markov chains.

Reversible Markov chains

- Denote by $P'_{i,j} = \mathbb{P}(X_{t-1} = j | X_t = i)$ the transition matrix for the time-reversed chain.
- It seems clear that a Markov chain will be reversible if and only if $P = P'$, so that any particular transition occurs with equal probability in forward and reverse directions.
- **Theorem**. (i) If there is a probability mass function $\pi(i), i \in \Omega$ satisfying $\pi(i) \geq 0$, $\sum_{i \in \Omega} \pi(i) = 1$ and

  $$\text{“Detailed balance”:} \quad \pi(i) P_{i,j} = \pi(j) P_{j,i} \qquad \text{for all pairs } i, j \in \Omega, \quad (1)$$

  then $\pi = \pi P$ so π is stationary for P.
  (ii) If in addition $p^{(0)} = \pi$ then $P' = P$ and the chain is reversible with respect to π.

Reversible Markov chains

- Proof of (i): sum both sides of Eqn. 1 over $i \in \Omega$. Now $\sum_i P_{j,i} = 1$ so $\sum_i \pi(i) P_{i,j} = \pi(j)$.
- Proof of (ii), we have π a stationary distribution of P so $\mathbb{P}(X_t = i) = \pi(i)$ for all $t = 1, 2, \ldots$ along the chain. Then

$$\begin{aligned} P'_{i,j} &= \mathbb{P}(X_{t-1} = j | X_t = i) \\ &= \mathbb{P}(X_t = i | X_{t-1} = j) \frac{\mathbb{P}(X_{t-1} = j)}{\mathbb{P}(X_t = i)} \text{ (Bayes rule)} \\ &= P_{j,i} \pi(j) / \pi(i) \text{ (stationarity)} \\ &= P_{i,j} \text{ (detailed balance).} \end{aligned}$$

Reversible Markov chains

- Why bother with reversibility? If we can find a transition matrix P satisfying $p(i)P_{i,j} = p(j)P_{j,i}$ for all i, j then $pP = p$ so 'our' p is a stationary distribution. Given P it is far easier to verify detailed balance, than to check $p = pP$ directly.
- We will be interested in using simulation of $\{X_t\}_{t=0}^{n-1}$ in order to estimate $\mathbb{E}_p(\phi(X))$. The idea will be to arrange things so that the stationary distribution of the chain is $\pi = p$: if $X_0 \sim p$ (*ie* start the chain in its stationary distribution) then $X_t \sim p$ for all t and we get some useful samples.
- The 'obvious' estimator is

$$\widehat{\phi}_n = n^{-1} \sum_{t=0}^{n-1} \phi(X_t),$$

but we may be concerned that we are averaging correlated quantities.

Ergodic Theorem

- **Theorem**. If $\{X_t\}_{t=0}^{\infty}$ is an irreducible and aperiodic Markov chain on a finite space of states Ω, with stationary distribution p then, as $n \to \infty$, for any bounded function $\phi : \Omega \to R$,

$$\mathbb{P}(X_n = i) \to p(i) \text{ and } \widehat{\phi}_n = \frac{1}{n}\sum_{t=0}^{n-1} \phi(X_t) \ \to \mathbb{E}_p(\phi(X)).$$

- $\widehat{\phi}_n$ is weakly and strongly consistent. In Part A Proba the Ergodic theorem asks for positive recurrent $X_0, X_1, X_2, ...$, and the stated conditions are simpler here because we are assuming a finite state space for the Markov chain.
- We would really like to have a CLT for $\widehat{\phi}_n$ formed from the Markov chain output, so we have confidence intervals $\pm\sqrt{\text{var}(\widehat{\phi}_n)}$ as well as the central point estimate $\widehat{\phi}_n$ itself. These results hold for all the examples discussed later but are a little beyond us at this point.

How Many Samples

- The problem of how large n must be for the guaranteed convergence to give a usefully accurate estimate does not have a simple honest answer.
- However we can repeat the entire simulation and check that independent estimates $\widehat{\phi}_n$ have an acceptably small variance.
- We can also check also that 'most' of the samples are not biased in any obvious way by the choice of X_0.
- We can also repeat the entire simulation for various choices of X_0 and check that independent estimates $\widehat{\phi}_n$ have an acceptably small variance.

Outline

Metropolis-Hastings Algorithm

- The Metropolis-Hastings (MH) algorithm allows to simulate a Markov Chain with any given equilibrium distribution.
- If we are given a pdf or pmf p then we may be able to simulate an iid sequence $X_1, X_2, ..., X_n$ of r.v. satisfying $n^{-1} \sum_i \phi(X_i) \to \mathbb{E}_p(\phi(X))$ as $n \to \infty$, using the Rejection algorithm.
- In a similar way, if we are given a pdf or pmf p then we may be able to simulate an correlated sequence $X_1, X_2, ..., X_n$ of r.v. (*ie*, a Markov chain) satisfying $n^{-1} \sum_i \phi(X_i) \to \mathbb{E}_p(\phi(X))$ as $n \to \infty$, using the MCMC algorithm.
- In each case convergence in probability is 'easily' established, whilst the more useful CLT 'usually' applies, but is harder to verify, at least in the MCMC case.

Metropolis-Hastings Algorithm

- We will start with simulation of random variable X on a finite state space.
- Let $p(x) = \tilde{p}(x)/Z_p$ be the pmf on finite state space $\Omega = \{1, 2, ..., m\}$. We will call p the (pmf of the) target distribution. Fix a 'proposal' transition matrix $q(y|x)$. We will use the notation $Y \sim q(\cdot|x)$ to mean $\Pr(Y = y|X = x) = q(y|x)$.
- If $X_t = x$, then X_{t+1} is determined in the following way.
  1. Let $Y \sim q(\cdot|x)$ and $U \sim U(0,1)$. Simulate $Y = y$ and $U = u$.
  2. If
     $$u \leq \alpha(y|x) \text{ where } \alpha(y|x) = \min\left\{1, \frac{\tilde{p}(y)q(x|y)}{\tilde{p}(x)q(y|x)}\right\}$$
     set $X_{t+1} = y$, otherwise set $X_{t+1} = x$.

Metropolis-Hastings Algorithm

- **Theorem**. The transition matrix P of the Markov chain generated by the M-H algorithm satisfies $p = pP$.
- Proof: Since p is a pmf, we just check detailed balance. The case $x = y$ is trivial. If $X_t = x$, then the proba to come out with $X_{t+1} = y$ for $y \neq x$ is the proba to propose y at step 1 times the proba to accept it at step 2. Hence we have $P_{x,y} = \mathbb{P}(X_{t+1} = y | X_t = x) = q(y|x)\alpha(y|x)$ and

$$\begin{aligned}
p(x)P_{x,y} &= p(x)q(y|x)\alpha(y|x) \\
&= p(x)q(y|x)\min\left\{1, \frac{p(y)q(x|y)}{p(x)q(y|x)}\right\} \\
&= \min\left\{p(x)q(y|x), p(y)q(x|y)\right\} \\
&= p(y)q(x|y)\min\left\{\frac{p(x)q(y|x)}{p(y)q(x|y)}, 1)\right\} \\
&= p(y)q(x|y)\alpha(x|y) = p(y)P_{y,x}.
\end{aligned}$$

Metropolis-Hastings Algorithm

- To run the MH algo., we need to specify $X_0 = x_0$ and a proposal $q(y|x)$. We then repeat steps 1 and 2 to generate a sequence $X_0, X_1, ..., X_n$, and these are our correlated samples distributed according to p (at least for large n when $p^{(n)}$ has converged to p).
- We only need to know the target p up to a normalizing constant as α depends only $p(y)/p(x) = \tilde{p}(y)/\tilde{p}(x)$.
- If the Markov chain simulated by the M-H algorithm is irreducible and aperiodic then the ergodic theorem applies.
- Verifying aperiodicity is usually straightforward, since the MCMC algo. may reject the candidate state y, so $P_{x,x} > 0$ for at least some states $x \in \Omega$. In order to check irreducibility we need to check that q can take us anywhere in Ω (so q itself is an irreducible transition matrix), and then that the acceptance step doesn't trap the chain (as might happen if $\alpha(y|x)$ is zero too often).

Example: Simulating a Discrete Distribution

- We will use MCMC to simulate $X \sim p$ on $\Omega = \{1, 2, ..., m\}$ with $\tilde{p}(i) = i$ so $Z_p = \sum_{i=1}^{m} i = \frac{m(m+1)}{2}$.
- One simple proposal distribution is $Y \sim q$ on Ω such that $q(i) = 1/m$.
- This proposal scheme is clearly irreducible (we can get from A to B in a single hop).
- If $X_t = x$, then X_{t+1} is determined in the following way.
  1. Let $Y \sim U\{1, 2, ..., m\}$ and $U \sim U(0, 1)$. Simulate $Y = y$ and $U = u$.
  2. If
  $$u \leq \min\left\{1, \frac{\tilde{p}(y)q(x|y)}{\tilde{p}(x)q(y|x)}\right\} = \min\left\{1, \frac{y}{x}\right\}$$
  set $X_{t+1} = y$, otherwise set $X_{t+1} = x$.

Example: Simulating a Poisson Distribution

- We want to simulate $p(x) = e^{-\lambda}\lambda^x/x! \propto \lambda^x/x!$
- For the proposal we use

$$q(y|x) = \begin{cases} \frac{1}{2} & \text{for } y = x \pm 1 \\ 0 & \text{otherwise,} \end{cases}$$

  i.e. toss a coin and add or substract 1 to x to obtain y.
- If $X_t = x$, then X_{t+1} is determined in the following way.
  1. Let $V \sim \mathcal{U}(0,1)$ and set $y = x + 1$ if $V \leq \frac{1}{2}$ and $y = x - 1$ otherwise. Simulate $U \sim U(0,1)$.
  2. Let $\alpha(y|x) = \min\left\{1, \frac{\tilde{p}(y)q(x|y)}{\tilde{p}(x)q(y|x)}\right\}$ then

$$\alpha(y|x) = \begin{cases} \min\left(1, \frac{\lambda}{x+1}\right) & \text{if } y = x + 1 \\ \min\left(1, \frac{x}{\lambda}\right) & \text{if } y = x - 1 \geq 0 \\ 0 & \text{if } y = -1. \end{cases}$$

  and if $u \leq \alpha(y|x)$, set $X_{t+1} = y$, otherwise set $X_{t+1} = x$.

Estimating Normalizing Constants

- Assume we are interested in estimating Z_p.
- If we have an irreducible and aperiodic Markov chain then the ergodic theorem tells us that $\widehat{\phi}_n = \frac{1}{n}\sum_{t=0}^{n-1}\phi(X_t) \ \rightarrow \mathbb{E}_p(\phi(X))$ so for $\phi(x) = \mathbb{I}_{x_0}(x)$, $\mathbb{E}_p(\phi(X)) = p(x_0)$

$$\widehat{p}_n(x_0) = \frac{1}{n}\sum_{t=0}^{n-1}\mathbb{I}_{x_0}(X_t) \rightarrow p(x_0).$$

- For any x_0 s.t. $p(x_0) > 0$, we have

$$p(x_0) = \frac{\tilde{p}(x_0)}{Z_p} \Leftrightarrow Z_p = \frac{\tilde{p}(x_0)}{p(x_0)}.$$

- Hence a consistent estimate of Z_p is

$$\widehat{Z}_{p,n} = \frac{\tilde{p}(x_0)}{\widehat{p}_n(x_0)}.$$