MSc HT15. Further Statistical Methods: MCMC

Lecture 3-4: Rejection (continued); Importance sampling and variance reduction; Markov chains

Notes and Practicals available at

www.stats.ox.ac.uk/~nicholls/MScMCMC15
Normalizing constants and the rejection algorithm

A probability density \( p(x), x \in \Omega \) must be normalized, so that
\[
\int_{\Omega} p(x) \, dx = 1.
\]

We can specify a probability density by just giving a function \( \tilde{p}(x) \) proportional to \( p(x) \). If
\[
p(x) \propto \tilde{p}(x)
\]
then \( p(x) = \frac{\tilde{p}(x)}{Z_p} \) with
\[
Z_p = \int_{\Omega} \tilde{p}(x) \, dx.
\]

\( p(x) \) is determined from \( \tilde{p}(x) \), even if we can’t calculate it.
Normalizing constants are hard to calculate, so we like algorithms that simulate \( X \sim p \) without our needing to calculate \( p(x) \). The rejection algorithm is one of these.

Suppose (usual setup) we can simulate \( Y \sim q \) and we want \( X \sim p \). What if we can only calculate \( \tilde{q}(x) \) and \( \tilde{p}(x) \) and not \( Z_q \) and \( Z_p \) (so not \( q \) and \( p \)).

Suppose we can find \( M \) so that \( M \geq \tilde{p}(x)/\tilde{q}(x) \) for all \( x \).

Rejection Algorithm simulating \( X \sim p \):

[1] Simulate \( y \sim q(y) \) and \( u \sim U(0, 1) \).

[2] If \( u < \tilde{p}(y)/M\tilde{q}(y) \) return \( X = y \), otherwise goto [1].

Nowhere in this algorithm did we need to compute \( p(x) \) or \( q(x) \).
Example: The random variable $X$ has probability density

$$p(x) \propto \frac{\sin^2(x)}{x^2}, \quad -\infty < x < \infty$$

Give a rejection algorithm which simulates $X$ using iid simulation of $Y \sim q$ with $q(y) \propto (1 + y^2)^{-1}$ (the Cauchy distribution).
Step 1: bound $\tilde{p}/\tilde{q}$.

$$\frac{\tilde{p}(x)}{\tilde{q}(x)} = \frac{(1 + x^2) \sin^2(x)}{x^2}$$

$$= \frac{\sin^2(x)}{x^2} + \sin^2(x)$$

$$\leq 2$$

because $|\sin(x)/x| < 1$. We can take $M = 2$.

Step 2: we have the following algorithm for $X \sim p(x)$.

[1] Simulate $y \sim t(1)$ and $u \sim U(0, 1)$.

[2] If $u < (1 + y^2) \sin^2(y)/2y^2$ return $X = y$, otherwise goto [1].

The histogram at the start of this example was produced from the implementation of this algorithm in the R-code associated with this lecture.
Comment: The t-distribution with one degree of freedom (called \( t(1) \), or the Cauchy distribution) has density

\[
q(y) \propto (1 + y^2)^{-1}.
\]

\( Y \) is easy to simulate. If

\[
Z_1, Z_2 \sim N(0, 1)
\]

are iid standard normal and we set

\[
Y = Z_1/Z_2
\]

then \( Y \sim t(1) \).

We can calculate the normalizing constants for both \( q \) and \( p \) using contour integrals but \( F^{-1}(u) \) is very hard work for \( X \).
Why does “unnormalized” rejection work? Let $M' = M Z_q / Z_p$. The test

$$u < \frac{\tilde{p}(y)}{M \tilde{q}(y)}$$

is the exactly the same as the test

$$u < \frac{p(y)}{M' q(y)}.$$ 

We cant calculate $M'$, $p$ or $q$, but if we use $M$ with $\tilde{p}$ and $\tilde{q}$ it is just as if we were using $M'$ with $p$ and $q$. However,

$$M \geq \frac{\tilde{p}(x)}{\tilde{q}(x)} \iff \frac{M Z_q}{Z_p} \geq \frac{\tilde{p}(x) Z_q}{\tilde{q}(x) Z_p} \iff M' \geq \frac{p(x)}{q(x)},$$

so the $M'$ “implied” by our choice of $M$ satisfies $M' \geq p(x)/q(x)$ for all $x$, and that means the revised algorithm is correct.

When we worked out the mean number of trials we worked with normalized densities. If we work with unnormalized densities (use $M$ to bound $\tilde{p}/\tilde{q}$) the mean number of trials is $M' = M Z_q / Z_p$. 
Importance Sampling Estimator

Slight revision on usual story: we can sample $Y \sim q, Y \in \Omega$. We want to estimate $\theta = E_p(\phi(X))$ where $X \sim p, X \in \Omega$ and $\phi$ is some given function $\phi : \Omega \rightarrow \Re$.

Idea: simulate $Y_1, Y_2, Y_3, \ldots, Y_n \sim q$ iid and form the weighted average

$$\hat{\theta}_{n}^{\text{IS}}(Y) = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i)w(Y_i)$$

with $w(Y_i) = p(Y_i)/q(Y_i)$.

Proposition: If $p(x)\phi(x) \neq 0 \Rightarrow q(x) > 0$ and $E_p(\phi(X))$ exists then $\hat{\theta}_{n}^{\text{IS}}$ is unbiased and consistent (proof by WLLN).
Proposition: If $E_p(\phi^2(X)w(X))$ and $E_p(\phi(X))$ are finite then $\hat{\theta}_n^\text{IS}$ is unbiased and consistent.

Proof. Unbiasedness:

\[
E_q(\hat{\theta}_n^\text{IS}) = \frac{1}{n} \sum_{i=1}^{n} E_q \left( \phi(Y_i) \frac{p(Y_i)}{q(Y_i)} \right)
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \int_{\Omega} \phi(y_i) \frac{p(y_i)}{q(y_i)} q(y_i) dy_i
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \int_{\Omega} \phi(x_i) p(x_i) dx_i
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} E_p(\phi(X))
\]
\[
= \theta
\]

so $\hat{\theta}_n^\text{IS}$ is unbiased.
Proof continued. Consistency: show that for each \( \epsilon > 0 \),

\[
\Pr(|\hat{\theta}^{\text{IS}}_n - \theta| \geq \epsilon) \to 0 \quad \text{as} \quad n \to \infty.
\]

By the Markov inequality for rv \( Z \geq 0 \), \( \Pr(Z \geq a) \leq E(Z)/a \).

\[
\Pr(|\hat{\theta}^{\text{IS}}_n - \theta| \geq \epsilon) = \Pr(|\hat{\theta}^{\text{IS}}_n - \theta|^2 \geq \epsilon^2)
\]

\[
\leq \frac{E_q(|\hat{\theta}^{\text{IS}}_n - \theta|^2)}{\epsilon^2}
\]

\[
= \var_q(\hat{\theta}^{\text{IS}}_n)
\]

\[
= \var_q \left( \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i) \frac{p(Y_i)}{q(Y_i)} \right)
\]

\[
= \var_q \left( \phi(Y) \frac{p(Y)}{q(Y)} \right)
\]

\[
= \frac{\var_q(\phi(Y)) p(Y)}{n \epsilon q(Y)}
\]

so the probability for a large error tends to zero as \( n \to \infty \).
Example: Gamma Distribution

Earlier on we used the transformation method to simulate

\[ Y \sim \Gamma(a, b) \]

for \( a = 1, 2, 3, \ldots \) and \( b > 0 \) by summing exponentials. Suppose we have simulated \( Y_i, i = 1, 2, \ldots, n \) iid \( \Gamma(a, b) \) rv, but want to estimate the expectation of \( \phi(X) \) in some rv \n
\[ X \sim \Gamma(\alpha, \beta) \]

for some \( \alpha, \beta > 0 \).

The Gamma(\( \alpha, \beta \)) density is

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

\[ w(y) = \frac{p(y)}{q(y)} = \frac{\Gamma(a) \beta^\alpha}{\Gamma(\alpha) b^a} y^{\alpha-a} \exp(-(\beta - b)y) \]

Hence

\[
\hat{\theta}_{n}^{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i) w(Y_i) \\
= \frac{\Gamma(a) \beta^\alpha}{\Gamma(\alpha) b^a} \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i) Y_i^{\alpha-a} \exp(-(\beta - b)Y_i)
\]

is an unbiased and consistent estimate of \( E_p(\phi(X)) \). We can actually “recycle” the \( Y \)'s and compute \( E_{\alpha,\beta}(\phi(X)) \) for lots of \( \alpha \)'s and \( \beta \)'s.

So far so good.
Variance of the Importance Sampling Estimator

If $\theta = E_p(\phi(X))$ and $E_p(w(X)\phi^2(X))$ are finite then

$$\text{var}_q(\hat{\theta}_n^{\text{IS}}) = \frac{1}{n} \left( E_p(w(X)\phi^2(X)) - \theta^2 \right).$$

Proof:

$$\text{var}_q(\hat{\theta}_n^{\text{IS}}) = \text{var}_q \left( \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i)w(Y_i) \right)$$

$$= \frac{1}{n} \text{var}_q (\phi(Y_1)w(Y_1))$$

$$= \frac{1}{n} \left( E_q \left( w(Y_1)^2 \phi(Y_1)^2 \right) - E_q (w(Y_1)\phi(Y_1))^2 \right).$$
The second expectation is $E_q(\phi(Y_1)p(Y_1)/q(Y_1)) = \theta$ as we saw earlier. The first expectation can also be converted into an expectation in $X \sim p$.

$$E_q \left( w(Y_1)^2 \phi(Y_1)^2 \right) = \int_\Omega \frac{p(y)^2}{q(y)^2} \phi(y)^2 q(y) dy$$

$$= \int_\Omega \frac{p(y)}{q(y)} \phi(y)^2 p(y) dy$$

$$= E_p \left( w(X) \phi(X)^2 \right)$$

and hence

$$\text{var}_q(\hat{\theta}^\text{IS}_n) = \frac{1}{n} \left( E_p \left( w(X) \phi(X)^2 \right) - \theta^2 \right).$$
Each time we do IS we should check that this variance is finite (and ideally small). We check $E_p(w\phi^2)$ is finite.

How can we show $E_p(w\phi^2)$ is finite? If $\phi(X)$ has finite mean and variance then $E_p(\phi^2)$ must be finite. If $w(x)$ in addition is bounded $w(x) \leq M$ for all $x \in \Omega$ then

$$E_p(w\phi^2) \leq ME_p(\phi^2) \leq \infty.$$ 

That is actually the same condition we had for rejection, 

$$p(x)/q(x) \leq M \quad \text{for all } x \in \Omega$$

though now we just show $M$ exists, we don’t have to find it.

This is just sufficient: $w(x)$ may be unbounded, but $E_p(w\phi^2)$ still finite. Importance sampling has a wider domain of application than rejection. It is also statistically more efficient (hardish proof - lecturer’s prize if you can show this).
Example: Gamma Distribution (continued)

Check that \( \text{var}(\hat{\theta}^\text{IS}_n) \) is finite if \( E_p(\phi) \) and \( \text{var}_p(\phi) \) are finite.

It is sufficient that \( E_p \left( w(Y)\phi(Y)^2 \right) \) is finite.

\[
w(x)\phi(x)^2 = \frac{\Gamma(x; \alpha, \beta)}{\Gamma(x; a, b)} \phi(x)^2 = \frac{\Gamma(a) \beta^\alpha}{\Gamma(\alpha) b^a} x^{\alpha-a} \exp(-(\beta - b)x) \phi(x)^2,
\]

so

\[
E_p \left( w(X)\phi(X)^2 \right) \propto E_p \left( X^{\alpha-a} \exp(-(\beta - b)X)\phi(X)^2 \right) \]

\[
= \int_0^\infty p(x) x^{\alpha-a} \exp(-(\beta - b)x) \phi(x)^2 \, dx
\]

\( x^{\alpha-a} \exp(-(\beta - b)x) \) bounded if \( a < \alpha \) and \( b < \beta \), sufficient for finite variance. Show eg \( a < 2\alpha \) and \( b < 2\beta \) is N&S if \( \phi(x) = 1 \).
Try $a = 2, b = 2$ and $\beta = 2.5$, $\alpha = 0.5$ (ie $\alpha$ less than $a$) and $\phi(x) = 1$. Monitor the weights $w(y_i)$ and the sequence of estimates $\hat{\theta}^\text{IS}_m$, $m = 1, 2, \ldots n$.

The estimator is hit by occasional huge weights. Exercise: What would happen if we used $\phi(x) = x$?
Rare Event Estimation and variance reduction

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 doesn’t help.

For example, suppose $X \sim p$ and we want to estimate

$$P(X > x_0) = E_p(\mathbb{I}[X > x_0])$$

with $x_0$ in the extreme upper tail of $p(x)$. We may not get any samples $X_i > x_0$ and the usual estimate

$$\hat{\theta}_n = \sum_i \mathbb{I}(X_i > x_0) / n$$

is simply zero. We can take a $q$-dbn that puts more probability at large $Y$, and then reweight to get expectations in $X$. By using IS, we can actually reduce the variance of our estimator.
Example

Say \( p(x) = N(x; \mu, \sigma^2) \) and we want to estimate \( \theta = \mathbb{P}(X > x_0) \) for some \( x_0 \gg \mu + 3\sigma \).

Take \( q \) to be some simple distribution that sits over \( x_0 \). A natural choice is \( q(y) = N(y; x_0, \sigma^2) \).

The weights \( w = p/q \) are

\[
    w(y) = \frac{N(y; \mu, \sigma^2)}{N(y; x_0, \sigma^2)} = \exp\left(-\frac{(y - \mu)^2}{2\sigma^2} + \frac{(y - x_0)^2}{2\sigma^2}\right)
\]

and the IS estimator is \( \hat{\theta}_{n}^{\text{IS}} = \frac{1}{n} \sum_{i=1}^{n} w(Y_i) \mathbb{I}_{Y_i > x_0} \).
The variance reduction can be dramatic. Here are 100 estimates of $\Pr(X > 4)$ for $X \sim N(0, 1)$ using $q(y) = N(y; 4, 1)$.
Unnormalized Importance sampling

Recall \( p(x) = \frac{\tilde{p}(x)}{Z_p} \), \( q(x) = \frac{\tilde{q}(x)}{Z_q} \) with \( Z_p, Z_q \) commonly intractable.

Same issue as for rejection. The IS weights are \( w = \frac{p}{q} \) so need \( q \) and \( p \) normalized.

Let \( \tilde{w} = \frac{\tilde{p}}{\tilde{p}} \). If we use \( \frac{1}{n} \sum_{i=1}^{n} \tilde{w}(Y_i) \phi(Y_i) \) then we find

\[
E_q \left( \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{p}(Y_i)}{\tilde{q}(Y_i)} \phi(Y_i) \right) = E_q \left( \frac{1}{n} \sum_{i=1}^{n} \frac{Z_p p(Y_i)}{Z_q q(Y_i)} \phi(Y_i) \right) = \frac{Z_p}{Z_q} E_p(\phi(X)).
\]
We need to estimate $Z_p/Z_q$ and divide. $\frac{1}{n}\sum_{i=1}^{n} \tilde{w}(Y_i)$ is the estimator we need.

\[
E_q \left( \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{p}(Y_i)}{\tilde{q}(Y_i)} \right) = E_q \left( \frac{1}{n} \sum_{i=1}^{n} \frac{Z_p p(Y_i)}{Z_q q(Y_i)} \right) = \frac{Z_p}{Z_q} E_q \left( \frac{1}{n} \sum_{i=1}^{n} \frac{p(Y_i)}{q(Y_i)} \right) = \frac{Z_p}{Z_q}
\]

since $\sum_{i=1}^{n} w(Y_i)/n$ is the IS estimator for $\phi = 1$. We will see that indeed

\[
\hat{\theta}_{n}^{\text{IS}} = \frac{\sum_{i=1}^{n} \tilde{w}(Y_i) \phi(Y_i)}{\sum_{i=1}^{n} \tilde{w}(Y_i)}
\]

is consistent for $E_p(\phi(X))$. 

Example: we saw that if $Y_i \sim \Gamma(a, b)$ and

$$w(y) = \frac{\Gamma(a)\beta^\alpha}{\Gamma(\alpha)b^a} y^{\alpha-a} \exp(-(\beta - b)y)$$

then

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

is unbiased and consistent for $E_p(\phi(X))$ with $X \sim \Gamma(\alpha, \beta)$. From above, if

$$\tilde{w}(y) = y^{\alpha-a} \exp(-(\beta - b)y)$$

then

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

is a consistent estimator for $E_p(\phi(X))$. 
Example (cont). I will take $a = b = 1$ so $Y \sim \text{Exp}(1)$ and estimate $E_p(X)$ with $p(x) = \Gamma(x; \alpha = 2, \beta = 4)$.

```r
> phi<-function(x) {x}
>
> theta.est<-function(n,alpha,beta) {
+   #IS estimate of $E_p(\phi(X))$, $X \sim \text{Gamma}(\alpha, \beta)$
+   y<-rexp(n)
+   w<-y^(alpha-1)*exp(-(beta-1)*y)
+   theta.hat<-mean(phi(y)*w)/mean(w)
+   return(theta.hat)
+ }
>
> theta.est(1000,alpha=2,beta=4)
[1] 0.5043166
```

We can use the delta method to estimate the variance of our estimate. Also, there is a CLT for $\tilde{\theta}_n$. 
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 \( X_0, X_1, X_2, \ldots \) which satisfies \( X_t \sim p \) (or at least \( X_t \) converges to \( p \) in distribution) and rely on the usual estimate \( \hat{\phi}_n = n^{-1} \sum_{t=0}^{n-1} \phi(X_t) \).

We will suppose \( \Omega \), the space of states of \( X \), is finite (and therefore discrete).

MCMC methods are applicable to countably infinite and continuous state spaces, and are one of the most versatile classes of Monte Carlo algorithms we have.
Markov chains

Let $\{X_t\}_{t=0}^\infty$ be a homogeneous Markov chain of random variables on $\Omega$ 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 $p^{(n)}(i) = \mathbb{P}(X_n = i)$.

The transition matrix $P$ 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

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

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.
The Stationary Distribution 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 and its reversal, even with knowledge of the transition matrix.

Most MCMC algorithms are based on 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

"Detailed balance": $\pi(i)P_{i,j} = \pi(j)P_{j,i}$ for all pairs $i, j \in \Omega$,

then $\pi = \pi P$ so $\pi$ is stationary for $P$.

(II) If in addition $p^{(0)} = \pi$ then $P' = P$ and the chain is reversible with respect to $\pi$. 
Proof of (I): sum both sides of detailed balance equation 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 $\pi$ a stationary distribution of $P$ so $\mathbb{P}(X_t = i) = \pi(i)$ for all $t = 1, 2, \ldots$ along the chain. Then

$$P'_{i,j} = \mathbb{P}(X_{t-1} = j | X_t = i)$$

$$= \mathbb{P}(X_t = i | X_{t-1} = j) \frac{\mathbb{P}(X_{t-1} = j)}{\mathbb{P}(X_t = i)} \quad \text{(Bayes rule)}$$

$$= P_{j,i} \pi(j) / \pi(i) \quad \text{(stationarity)}$$

$$= P_{i,j} \quad \text{(detailed balance).}$$
Convergence and the Ergodic Theorem

If the (finite state space) MC is irreducible and aperiodic then the stationary distribution is unique and \( p_t \to \pi \) as \( t \to \infty \). We say the chain “targets” \( \pi \). If we simulate the MC \( X_0, X_1, \ldots X_n \) to large enough \( n \) from any start \( X_0 = x_0 \) then since \( X_t \sim p^t \) and \( p^t \sim \pi \) at large \( t \), ’most’ of the samples are ’nearly’ distributed according to \( \pi \).

We will use \( \{X_t\}_{t=0}^{n-1} \) to estimate \( \mathbb{E}_p(\phi(X)) \). The ‘obvious’ estimator is

\[
\hat{\phi}_n = \frac{1}{n} \sum_{t=0}^{n-1} \phi(X_t),
\]

but the \( X_t \) are correlated and only converge in distribution to \( \pi \).
Theorem. If \( \{X_t\}_{t=0}^{\infty} \) is an irreducible and aperiodic Markov chain on a finite space of states \( \Omega \), and is reversible wrt \( \pi \), then as \( n \to \infty \)

\[
P(X_n = i) \to \pi(i) \quad \text{and} \quad \hat{\phi}_n \xrightarrow{P} \mathbb{E}_p(\phi(X))
\]

for any bounded function \( \phi : \Omega \to \mathbb{R} \).

We refer to such a chain as ergodic with equilibrium \( \pi \).

\( \hat{\phi}_n \) is consistent. A more general statement asks for a positive recurrent chain. The 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 \( \hat{\phi}_n \) formed from the Markov chain output, so we have confidence intervals \( \pm \sqrt{\text{var}(\hat{\phi}_n)} \) as well as the central point estimate \( \hat{\phi}_n \) itself. These results hold for all the examples discussed later but are a little beyond us at this point.
Metropolis-Hastings Algorithm

Suppose we need samples from a pmf $p(x), x \in \Omega$. We give an algorithm simulating a Markov chain targeting $p$. It is enough to give a rule simulating $X_{t+1}$ given $X_t$. The algorithm determines the transition probabilities $P(X_{t+1} = y|X_t = x)$ and the transition matrix $P$.

Let $p(x) = \tilde{p}(x)/Z_p$ be the pmf on finite state space $\Omega = \{1, 2, ..., m\}$. We will call $p$ the target distribution.

Choose a ‘proposal’ transition matrix $q(y|x)$. We will use the notation $Y \sim q(\cdot|x)$ to mean $\Pr(Y = y|X = x) = q(y|x)$. 
Metropolis Hastings MCMC: the following algorithm simulates a Markov chain. If the chain is irreducible and aperiodic then it is ergodic with equilibrium distribution $p$.

Let $X_t = x$. $X_{t+1}$ is determined in the following way.

[1] Draw $y \sim q(\cdot|x)$ and $u \sim U[0, 1]$.

[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 $X_{t+1} = x$.

We initialise this with $X_0 = x_0, p(x_0) > 0$ and iterate for $t = 1, 2, 3, \ldots n$ to simulate the samples we need.
Example: Simulating a Discrete Distribution

Let \( p(x) = x/Z_p \) with \( Z_p = \sum_{x=1}^{m} x \).

Give a MH MCMC algorithm ergodic for \( p(x), x = 1, 2, ..., m \).

Step 1: Choose a proposal distribution \( q(y|x) \). It needs to be easy to simulate and determine a irreversible chain.

A simple distribution that 'will do' is \( Y \sim U\{1, 2, ..., m\} \), so

\[
q(y) = \frac{1}{m}, \quad y = 1, 2, ..., m.
\]

This proposal scheme is clearly irreducible (we can get from \( A \) to \( B \) in a single hop).
Step 2: write down the algorithm.

If $X_t = x$, then $X_{t+1}$ is determined in the following way.

[1] Simulate $y \sim U\{1, 2, \ldots, m\}$ and $u \sim U[0, 1]$. 

[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 $X_{t+1} = x$. 


#MCMC simulate \( X_t \) according to \( p=[1:m]/\text{sum}(1:m) \).

\( m<-30 \)

\( n<-10000; \ X<-\text{rep}(\text{NA},n)\); \ X[1]<-1

for (t in 1:(n-1)) {
    x<-X[t]
    y<-\text{ceiling}(m*\text{runif}(1))
    a<-\text{min}(1,y/x)
    U<-\text{runif}(1)
    if (U<=a) {
        X[t+1]<-y
    } else {
        X[t+1]<-x
    }
}
Left: $x$-axis is step counter $t = 1, 2, 3 \ldots 200$. The $y$-axis is Markov chain state $X_t$ for $	ilde{p}(x) = x, \ x = 1, ..., m, \ m = 30$.

Right: histogram of $X_1, X_2, ..., X_n$ for $n = 1000$. 