Part A Simulation and Statistical Programming HT14

Lecturer: Geoff Nicholls

University of Oxford

Lecture 8: Importance sampling; Markov chains

Notes and Problem sheets are available at

www.stats.ox.ac.uk\ ~nicholls\PartASSP

Unnormalized Importance sampling

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

Same issue as for rejection. The IS weights are $w = p/q$ so need q and p normalized.

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

$$\begin{aligned} E_q\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{p}(Y_i)}{\tilde{q}(Y_i)}\phi(Y_i)\right) &= E_q\left(\frac{1}{n}\sum_{i=1}^{n} \frac{Z_p}{Z_q}\frac{p(Y_i)}{q(Y_i)}\phi(Y_i)\right) \\ &= \frac{Z_p}{Z_q} E_p(\phi(X)). \end{aligned}$$

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.

$$\begin{aligned}
E_q\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{p}(Y_i)}{\tilde{q}(Y_i)}\right) &= E_q\left(\frac{1}{n}\sum_{i=1}^{n}\frac{Z_p}{Z_q}\frac{p(Y_i)}{q(Y_i)}\right)\\
&= \frac{Z_p}{Z_q}E_q\left(\frac{1}{n}\sum_{i=1}^{n}\frac{p(Y_i)}{q(Y_i)}\right)\\
&= Z_p/Z_q
\end{aligned}$$

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

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

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

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

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 $p(x) = \Gamma(x; \alpha = 2, \beta = 4)$.

```
> phi<-function(x) {x}
>
> theta.est<-function(n,alpha,beta) {
+   #IS estimate of E_p(phi(X)), X~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 $\tilde{\theta}_n^{\text{IS}}$. See the course texts for more on this.

Claim: If $Y_i \sim q,\ i = 1, 2, ..., n$ iid, $p(x) > 0 \Rightarrow q(x) > 0$ and

$$\tilde{\theta}_n^{\text{IS}} = \frac{\frac{1}{n}\sum_{i=1}^n \tilde{w}(Y_i)\phi(Y_i)}{\frac{1}{n}\sum_{i=1}^n \tilde{w}(Y_i)} \qquad \left(= \frac{a_n}{b_n} \text{ say}\right)$$

then $\tilde{\theta}_n^{\text{IS}}$ is consistent for $\theta = E_p(\phi(X))$.

Proof: Let $a/b = E_q(\tilde{w}\phi)/E(\tilde{w})$. We have seen that $a/b = \theta$. We need to show that

$$P(\left|\frac{a_n}{b_n} - \frac{a}{b}\right| > \epsilon) \to 0$$

as $n \to \infty$. It is easy to see (from our result for regular IS estimators) that $a_n \xrightarrow{P} a$ (and b_n etc) at large n. Then

$$\begin{aligned}
P(\left|\frac{a_n}{b_n} \right. & - & \left. \frac{a}{b}\right| > \epsilon) \\
& \leq & P(|b_n - b| > \frac{b}{2}) + P(|b_n - b| \leq \frac{b}{2}, |a_n b - a b_n| > \epsilon b_n b) \\
& \leq & P(|b_n - b| > \frac{b}{2}) + P(|a_n b - a b_n| > \epsilon \frac{b^2}{2}) \\
& < & P(|b_n - b| > \frac{b}{2}) + P(|a_n b - ab| > \frac{\epsilon b^2}{4}) + P(|ab - a b_n| > \frac{\epsilon b^2}{4}) \\
& \rightarrow & 0 \text{ as } n \rightarrow \infty \text{ by the consistency of } a_n \text{ and } b_n.
\end{aligned}$$

The middle step uses $b_n > b/2$, and

$$P(|a_n b - a b_n| > \frac{\epsilon b^2}{2}) \leq P(|a_n b - ab| > \frac{\epsilon b^2}{4}) + P(|ab - a b_n| > \frac{\epsilon b^2}{4}).$$

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 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 and widespread classes of Monte Carlo algorithms currently.

Markov chains

From Part A Probability. 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.

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 $t = 1, 2, ...$ 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": $\pi(i) P_{i,j} = \pi(j) P_{j,i} \qquad$ for all pairs $i, j \in \Omega$,

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.

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 π a stationary distribution of P so $\mathbb{P}(X_t = i) = \pi(i)$ for all $t = 1, 2, ...$ along the chain. Then

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

Convergence and the Ergodic Theorem

If the (finite state space) MC is irreducible and aperiodic then the stationary distribution is unique and $p^{(t)} \longrightarrow \pi$ as $t \longrightarrow \infty$. If we simulate the MC $X_0, X_1, ...X_n$ to large enough n from any start $X_0 = x_0$ then since $X_t \sim p^t$ and $p^t \simeq \pi$ at large t, 'most' of the samples are 'nearly' distributed according to π.

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 π.

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

$$P(X_n = i) \to \pi(i) \text{ and } \hat{\phi}_n \to \mathbb{E}_p(\phi(X)).$$

We refer to such a chain as ergodic with equilibrium π.

$\hat{\phi}_n$ is consistent. In Part A Probability the Ergodic theorem asks for positive recurrent $X_0, X_1, X_2, ...$ 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 $\hat{\phi}_n$ 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

The Metropolis-Hastings (MH) algorithm allows to simulate a Markov Chain with any given equilibrium distribution. We will start with simulation of random variable $X \sim p$ on a finite state space. We want to arrange things so that the Markov chain has equilibrium p.

We give an algorithm for simulating X_{t+1} give X_t. The algorithm determines the transition probabilities $P(X_{t+1} = j | X_t = i)$ and the transition matrix P. We have to choose the algorithm so that the transition matrix it simulates satisfies $pP = p$.

Let $p(x) = \tilde{p}(x)/Z_p$ be the pmf on finite state space $\Omega = \{1, 2, ..., m\}$. We will call p the (pmf of 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 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 $t = 1, 2, 3, ...n$ to simulate the samples we need.

Example: Simulating a Discrete Distribution

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

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

Step 1: Choose a proposal distribution $q(j|i)$. 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

$$q(i) = 1/m, \quad i = 1, 2, ..., m.$$

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

[2] If

$$\begin{aligned} u &\leq \min\left\{1, \frac{\tilde{p}(y)q(x|y)}{\tilde{p}(x)q(y|x)}\right\} \\ &= \min\left\{1, \frac{y}{x}\right\} \end{aligned}$$

set $X_{t+1} = y$, otherwise set $X_{t+1} = x$.

```
#MCMC simulate X_t according to p=[1:m]/sum(1:m).
m<-30
n<-10000; X<-rep(NA,n); X[1]<-1
for (t in 1:(n-1)) {
   x<-X[t]
   y<-ceiling(m*runif(1))
   a<-min(1,y/x)
   U<-runif(1)
   if (U<=a) {
      X[t+1]<-y
   } else {
      X[t+1]<-x
   }
}
```

Left: x-axis is Markov chain step counter $t = 1, 2, 3...200$ and y-axis is Markov chain state X_t for $\tilde{p}(i) = i, i = 1, 2, ..., m$, $m = 30$.

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