Sampling from the stationary distribution

Suppose we want to obtain samples from some distribution $\underline{\pi}$, which has a large number of states and so cannot be sampled from directly.

Idea: Set up an irreducible, aperiodic MC (i.e. define P), with stationary distribution $\underline{\pi}$. Then run this chain until it settles down to $\underline{\pi}$ – states will then be generated with the correct probabilities.

We have considered the problem of finding $\underline{\pi}$ given a specified P - now, we must find P given a specified $\underline{\pi}$ as its stationary distribution. Can we do this, and at the same time, ensure that the resulting chain is convergent?
Thankfully...yes!

Metropolis Algorithm

We can construct the required P by considering a probabilistic transition mechanism. We choose any symmetric transition matrix Q, with elements q_{ij}. Then, we define our transition mechanism as follows:
Suppose the chain is in state i...

- we select state j as a candidate for the next state of the chain, with probability q_{ij},
- and we then move to state j with probability

$$\min\left\{1, \frac{\pi_j}{\pi_i}\right\},$$

otherwise, we stay at state i. Note: if $\pi_j < \pi_i$, we will not always move.

Metropolis Algorithm

This defines P in the following way:

$$p_{ij} = \min\left\{1, \frac{\pi_j}{\pi_i}\right\} q_{ij}, \qquad i \neq j$$

$$p_{ii} = \underbrace{q_{ii}}_{\substack{\text{prob. choose} \\ \text{state } i}} + \sum_{j \neq i} \underbrace{\max\left\{0, 1 - \frac{\pi_j}{\pi_i}\right\}}_{\substack{\text{prob. not moving} \\ \text{to state } j}} q_{ij}$$

P defined in this way is certainly a transition matrix (as Q is), but will running the chain for a long time give us $\underline{\pi}$ as the limiting distribution?

Metropolis Algorithm

For this, we need to show that P is irreducible, aperiodic and that $\underline{\pi} = \underline{\pi} P$.

For the P designed above, we have

$$\begin{aligned}
\pi_i p_{ij} &= \pi_i \min\left\{1, \frac{\pi_j}{\pi_i}\right\} q_{ij} \\
&= \min\left\{\pi_i, \pi_j\right\} q_{ij} \\
&= \min\left\{\pi_i, \pi_j\right\} q_{ji} \quad \text{since } Q \text{ symmetric} \\
&= \pi_j \min\left\{1, \frac{\pi_i}{\pi_j}\right\} q_{ji} \\
&= \pi_j p_{ji}
\end{aligned}$$

...i.e. P satisfies the detailed balance equations (by design), and so the generated Markov chain will be reversible and have $\underline{\pi}$ as a stationary distribution.

Metropolis Algorithm

If we can now show that the chain is irreducible and aperiodic, then we guarantee that $\underline{\pi}$ will be the unique stationary distribution, and furthermore that the chain will be convergent.

- Irreducibility: If Q is irreducible, then so is P ($p_{ij} > 0$ iff $q_{ij} > 0$), so choose an irreducible Q!
- Aperiodicity: If we can show $p_{ii} > 0$, then state i is aperiodic, and thus all states are aperiodic (as there is only one class).

$$p_{ii} = q_{ii} + \sum_{j \neq i} \max\left\{0, 1 - \frac{\pi_j}{\pi_i}\right\} q_{ij}$$

i.e. $p_{ii} \geq q_{ii}$ – so we should choose Q with $q_{ii} > 0$. ... an easy condition to impose!

In short...choose a Q which ensures irreducibility and aperiodicity!
Note: only need to know $\underline{\pi}$ up to a normalizing constant, as it only enters the algorithm through π_j/π_i

Metropolis-Hastings for discrete state spaces

Suppose that the chain is in state i...

- select state j as a candidate state, according to some predefined transition probability q_{ij};
- then accept this candidate as the next state of the chain with probability

$$\alpha_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}.$$

The Markov chain generated by the discrete version of the Metropolis-Hastings sampler will have a transition matrix P specified by

$$p_{ij} = \min\left\{1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\} q_{ij}, \qquad i \neq j$$

$$p_{ii} = \underbrace{q_{ii}}_{\substack{\text{prob. choose} \\ \text{state } i}} + \sum_{j \neq i} \underbrace{\max\left\{0, 1 - \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right\}}_{\substack{\text{prob. not moving} \\ \text{to state } j}} q_{ij},$$

Metropolis-Hastings for discrete state spaces

This satisfies the detailed balance conditions by design:
Suppose $\pi_j q_{ji} \geq \pi_i q_{ij}$...

$$\implies p_{ij} = q_{ij}, \qquad p_{ji} = \frac{\pi_i q_{ij}}{\pi_j}$$

$$\implies \pi_i p_{ij} = \pi_i q_{ij} = \pi_j p_{ji}$$

...and swapping the indices gives the required result for $\pi_j q_{ji} \leq \pi_i q_{ij}$.

Metropolis-Hastings for discrete state spaces

We must initialise the procedure somehow! This is easy though - irreducibility means that, as long as we choose the initial value from the state space we're interested in, the algorithm will converge!

Note that, in order to recover the original Metropolis algorithm, all that is required is to choose a symmetric matrix of proposal probabilities $Q = \{q_{ij}\}$. More importantly, however...we are no longer restricted to symmetric proposal mechanisms!

In particular, with the Metropolis-Hastings sampler, we are allowed to propose candidate values for our chain independently of the chain's current state - this is Independent Metropolis-Hastings, and is sometimes useful for Bayesian inference.

Metropolis-Hastings for continuous state spaces

Suppose we wish to sample from the target density $f(x)$, defined on a continuous state space $\mathcal{X}$. As in the discrete MH procedure, we must specify a proposal mechanism for each step of the procedure. Here, we require a *proposal density* $q(y|x) \geq 0$, which satisfies

$$\int_{\mathcal{X}} q(y|x)dy = 1.$$

We must also initialise our procedure - this is once again a relatively straightforward affair. As long as our transition density $q(y|x)$ allows us to move from any part of the state space to any other part of the state space, in a finite number of iterations, then the Markov chain will be irreducible and thus converge to the required target from any arbitrary starting point chosen from within $\mathcal{X}$. Note that we can also choose to sample from a specified arbitrary initial distribution $\pi^{(0)}$.

Metropolis-Hastings Procedure (continuous state space)

1. Initialise the chain: start from an arbitrary X_0, possibly sampled from $\pi^{(0)}$. Set $n = 1$
2. Given $X_{n-1} = x$, generate a candidate value $Y = y$ from the proposal density $q(y|x)$.
3. Set $X_n = y$ with probability $\alpha(x, y)$, where

$$\alpha(x, y) = \min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\},$$

   otherwise, set $X_n = x$.
4. Replace n by $n + 1$ and return to Step 2.

Burn-in Periods

In both the discrete and continuous cases, the MH procedure is guaranteed to converge - before convergence occurs, however, the Markov chain cannot be used to represent a sample from the target distribution.

It is therefore common to allow a period of *burn-in* at the start of the algorithm, during which convergence is assumed to have not yet occurred. After a specified number of samples, the initial part of the chain is typically discarded, and the rest of the chain may be treated as a random sample from the target distribution of interest. "burn-in sample".

The chosen length of burn-in period is often subjective, and affected as much by available computer time as by formal convergence considerations, though convergence diagnostics may also be used.

Dependence

After discarding the burn-in sample, we are left with a statistically dependent sample from the target distribution. This could be slightly problematic...as Monte Carlo estimation requires an i.i.d. sample from the target distribution!

Under the existing assumptions of irreducibility and aperiodicity, the Monte Carlo estimate resulting from an M-length Markov chain realisation $X_{B+1}, \ldots, X_{B+M}$ for any test function ϕ such that $\int_{\mathcal{X}} |\phi(x)| f(x) dx < \infty$ satisfies

$$\lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \phi(X_{B+i}) = \mathrm{E}\left[\phi(X)\right] \qquad \text{with probability 1}$$

For moderate- length post-burn-in samples resulting from the MH procedure, a common step to take is to further discard all but every k-th iteration, where k is some sensible lag, chosen subjectively. e.g. through consulting the autocorrelation sequence This is known as thinning - can be quite wasteful

Scaling

The key quantity that has to be specified in the MH algorithm is the proposal density $q(y|x)$. This is often simplified by writing

$$q(y|x) = \frac{1}{h} g\left(\frac{y-x}{h}\right)$$

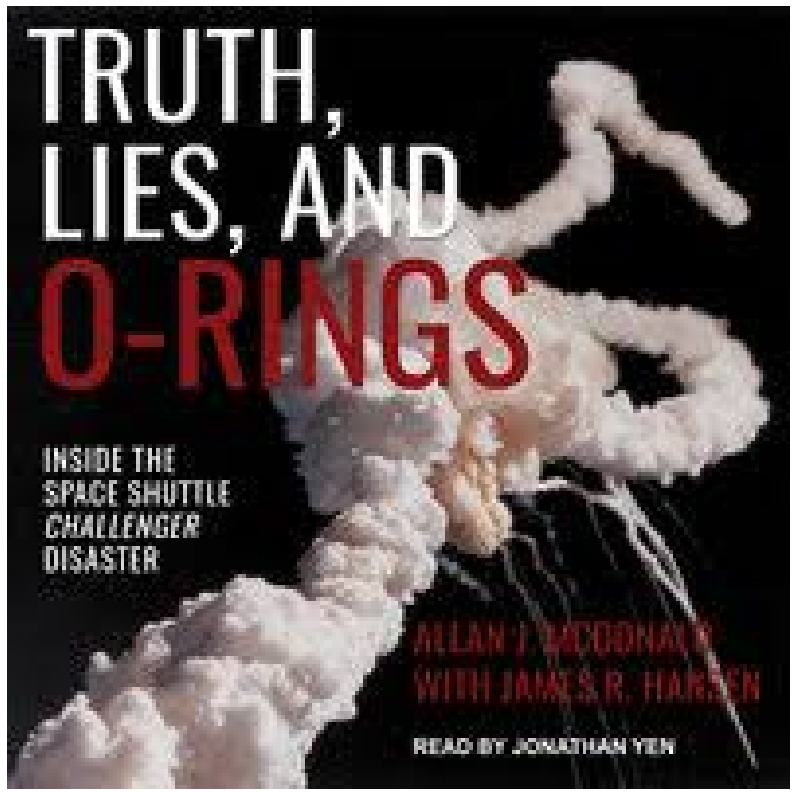
for some density g symmetric about 0 and a *scaling constant* h; note that in this case we automatically have $q(y|x) = q(x|y)$.

We may take g to be some standard form such as normal or uniform; the critical issue then becomes how to choose h so that the algorithm converges in reasonable time:

- If h is too large, then the proposal mechanism will explore the space well, but will often propose candidate values in regions where the target density is low, such that the probability of rejection is high.
- In contrast, however, small values of h can lead to high acceptance rates, but poor exploration of the space and significant dependence in the resulting Markov chain.

Metropolis-Hastings Example

Challenger Disaster: 28 June 1986
Temp: launch: 2 degrees (C) (36 F)
overnight: -8 degrees (C) (18 F)
O-rings (installed between the solid fuel segments) failed



Metropolis-Hastings Example

Data on 23 previous flights logs information on issues with the o-rings (binary) and launch temperature.
Let, Y_i represent failure of o-ring

$$Y_i \sim Bernoulli(\theta_i),\ i = 1, 2, \ldots, 23$$

with $\theta_i = \mathrm{P}(Y_i = 1) = \mathrm{E}(Y_i)$. Use the following model:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \beta x_i$$

where x_i is the temperature. i.e.

$$\theta_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

Likelihood

The likelihood is given by:

$$\mathcal{L}(y_1, y_2, \ldots, y_n \mid \alpha, \beta) = \prod_{i=1}^{n} f_{Y_i}(y_i) = \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}$$

So the log-likelihood, $\ell(y_1, y_2, \ldots, y_n \mid \alpha, \beta)$ is

$$\sum_{i=1}^{n} y_i \log\left(\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}\right) + (1-y_i)\log\left(1-\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}\right)$$

$$= \sum_{i=1}^{n} y_i(\alpha+\beta x_i) - \log(1+\exp(\alpha+\beta x_i))$$

Use random walk MH with likelihood as the target with normal proposals for α and β

Metropolis-Hastings for Bayesian inference

Suppose we adopt a Bayesian perspective. Our interest is now in establishing information about the distribution of a parameter $\theta \in \Theta \subseteq \mathbb{R}^d$, given some prior belief about the parameter's distribution, represented through the prior density $p(\theta)$, and given observation on some data, $\mathcal{D}$, dependent on the parameter with likelihood $\ell(\mathcal{D}|\theta)$.
We therefore seek to perform inference with respect to the posterior density

$$\pi(\theta|\mathcal{D}) \propto \ell(\mathcal{D}|\theta)p(\theta).$$

and it is common to use simulation-based inference methods, especially when

- the state space θ is high-dimensional, i.e. $d >> 1$
- the normalising constant is analytically intractable

(Continuous) MH Procedure for Bayesian Inference

A Metropolis-Hastings procedure may be designed for simulating a Markov chain that can be considered a sample from $\pi(\theta|\mathcal{D})$:

1. Initialise the chain: sample $\theta^{(0)} \sim p(\theta)$. Set $n = 1$
2. Given $\theta^{(n-1)} = \theta$, generate a candidate value η from a chosen proposal density $q(\eta|\theta)$.
3. Set $\theta^{(n)} = \eta$ with probability $\alpha(\theta, \eta)$, where

$$\alpha(\theta, \eta) = \min\left\{\frac{\pi(\eta|\mathcal{D})q(\theta|\eta)}{\pi(\theta|\mathcal{D})q(\eta|\theta)}, 1\right\},$$

   otherwise, set $\theta^{(n)} = \theta$.
4. Replace n by $n + 1$ and return to Step 2.

(Continuous) MH Procedure for Bayesian Inference

Now consider the ratio of densities in the acceptance probability:

$$\frac{\pi(\eta|\mathcal{D})q(\theta|\eta)}{\pi(\theta|\mathcal{D})q(\eta|\theta)} = \frac{\ell(\mathcal{D}|\eta)p(\eta)q(\theta|\eta)}{\ell(\mathcal{D}|\theta)p(\theta)q(\eta|\theta)}.$$

We can make two important notes:

- As before, our target density is only needed up to proportionality.
- This ratio of densities will simplify with a prudent choice of proposal density...

$$q(\eta|\theta) = p(\eta) \implies \alpha(\theta, \eta) = \min\left\{\frac{\ell(\mathcal{D}|\eta)}{\ell(\mathcal{D}|\theta)}, 1\right\},$$

When the state-space of the target distribution is high-dimensional, Metropolis-Hastings provides a flexible approach to performing simulation based inference...but there is a more efficient method...

Gibbs sampling for continuous state spaces

Suppose we have a d-dimensional target density, $f(x)$, $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$.

Gibbs Sampler:

1. Initialise the multivariate chain at an arbitrary initial vector, say $X^{(0)} = (X_1^{(0)}, \ldots, X_d^{(0)})$. Set $n = 1$
2. Given $X^{(n-1)}$, sample each component of $X^{(n)}$ from the full-conditionals:
   - sample $X_1^{(n)} \sim f(x_1 | X_2^{(n-1)}, X_3^{(n-1)}, \ldots, X_d^{(n-1)})$
   - sample $X_2^{(n)} \sim f(x_2 | X_1^{(n)}, X_3^{(n-1)}, \ldots, X_d^{(n-1)})$
   - $\vdots$
   - sample $X_d^{(n)} \sim f(x_d | X_1^{(n)}, X_2^{(n)}, \ldots, X_{d-1}^{(n)})$
3. Replace n by $n + 1$ and return to Step 2.