28/04/2022. LECTURE 20: NOTES MADE ON THE VISUALISER.

3.3.4 A BRIEF INSIGHT INTO WHY THE M-H ALGORITHM WORKS.

We illustrate why the M-H algorithm works. For simplicity, we shall work in the discrete setting.

Let X_t be a discrete time M-C with transition probabilities

$$P_{\theta\phi} = \mathbb{P}(X_t = \phi \mid X_{t-1} = \theta)$$

for all $\theta, \phi \in S$, the state space of the chain. Let

$$P_{\theta\phi} = q_1(\phi \mid \theta)$$

to make the link between the discrete and continuous cases explicit.

[NB. q_1 is the transition probability. It is NOT the proposal distribution which is q]

Suppose that the chain has a stationary distribution π (i.e. the chain is irreducible and aperiodic) Then:

$$\pi(\phi) = \sum_{\theta \in S} \pi(\theta) q_1(\phi \mid \theta)$$

[$\pi = \{\pi(\theta), \theta \in S, \pi(\theta) \geq 0, \sum_{\theta \in S} \pi(\theta) = 1$]

Note that, as the chain has to go somewhere,

$$\sum_{\phi \in S} q_1(\phi \mid \theta) = 1.$$

If the equation

$$\pi(\theta) q_1(\phi \mid \theta) = \pi(\phi) q_1(\theta \mid \phi)$$

is satisfied for all $\theta, \phi \in S$ then the $q_1(\phi \mid \theta)$ satisfy time reversibility or detailed balance.

If so, then the $q_1(\phi|\theta)$ give the transition probabilities of a Markov chain with stationary distribution π.

$$\sum_{\theta \in S} \pi(\theta) q_1(\phi|\theta) = \sum_{\theta \in S} \pi(\phi) q_1(\theta|\phi) \text{ by detailed balance}$$

$$= \pi(\phi) \underbrace{\sum_{\theta \in S} q_1(\theta|\phi)}_{=1 \text{ AS THE CHAIN HAS TO GO SOMEWHERE}}$$

$$= \pi(\phi)$$

Hence, π is stationary. Recall the M-H algorithm with proposal $q(\phi|\theta)$.

[q easy to sample from; aperiodic and irreducible mopped up in choice of q e.g. can stay still, doesn't always move forward]

We propose a move from θ to ϕ with probability

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

π here is the desired stationary distribution.

[Typically, we want $\pi(\theta) = f(\theta|x)$, $\theta = \theta^{(t-1)}$, $\phi = \theta^{*}$]

The transition probabilities of the M-C created by the M-H algorithm are:

propose move from θ to ϕ — probability of accepting the move

$$q_1(\phi|\theta) = \begin{cases} q(\phi|\theta)\alpha(\theta, \phi) & \phi \neq \theta \\ 1 - \sum_{\phi \neq \theta} q(\phi|\theta)\alpha(\theta, \phi) & \phi = \theta \end{cases}$$

[Two ways for $q_1(\theta|\theta)$: either reject a proposal or sample θ]

We demonstrate that detailed balance holds for π and q_1. It is trivially satisfied for $\theta = \phi$.

For $\phi \neq \theta$,

$$\pi(\theta) q_1(\phi|\theta) = \pi(\theta) q(\phi|\theta) \alpha(\theta, \phi)$$
$$= \pi(\theta) q(\phi|\theta) \min\left(1, \frac{\pi(\phi) q(\theta|\phi)}{\pi(\theta) q(\phi|\theta)}\right)$$
$$= \min\left(\pi(\theta) q(\phi|\theta), \pi(\phi) q(\theta|\phi)\right)$$
$$= \pi(\phi) q(\theta|\phi) \min\left(1, \frac{\pi(\theta) q(\phi|\theta)}{\pi(\phi) q(\theta|\phi)}\right)$$
$$= \pi(\phi) q(\theta|\phi) \alpha(\phi, \theta)$$
$$= \pi(\phi) q_1(\theta|\phi)$$

Thus, detailed balance holds. Hence, π is the stationary distribution.

3.3.5 USING THE SAMPLE FOR INFERENCE.

Early values of the chain prior to convergence are called "the burn-in". The length of the burn-in depends upon the rate of convergence and it can be difficult to obtain estimates of this.

The strategy is to discard initial observations $\theta^{(1)}, \ldots, \theta^{(b)}$ ('b' is the time to approximate convergence) and work with the observations $\theta^{(b+1)}, \ldots, \theta^{(N)}$ for inference.

As the ergodic averages converge,

$$\frac{1}{N} \sum_{i=1}^{N} g(\theta^{(i)}) \to \mathbb{E}(g(\theta) \mid \theta \sim \pi(\theta))$$

stationary distribution

so, $$\frac{1}{N}\left(\underbrace{\sum_{i=1}^{b} g(\theta^{(i)})}_{\text{prior to convergence}} + \underbrace{\sum_{i=b+1}^{N} g(\theta^{(i)})}_{\text{after convergence to the stationary}} \right) \to \mathbb{E}(g(\theta) \mid \theta \sim \pi(\theta))$$

Thus, $\frac{1}{N-b} \sum_{i=b+1}^{N} g(\theta^{(i)}) \rightarrow \mathbb{E}(g(\theta) \mid \theta \sim \pi(\theta))$

and typically with better properties.

3.3.6 EFFICIENCY OF THE ALGORITHMS.

The efficiency of the M-H algorithm depends upon a "good" choice of $q(\phi \mid \theta)$

[In examples, we choose $q(\phi \mid \theta)$ to be $N(\theta, \sigma^2)$ for a chosen value of σ^2. σ^2 acts as a "tuning parameter"]