UiO : Matematisk institutt

Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2020
Exercise 13

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no

Problem 2

Consider the Metropolis-Hastings algorithm where the transition densities $P(\boldsymbol{y}|\boldsymbol{x})$ are defined through:

- Sample a candidate value $\boldsymbol{X}^*$ from a *proposal distribution* $g(\cdot|\boldsymbol{x})$.
- Compute the Metropolis-Hastings ratio

$$R(\boldsymbol{x}, \boldsymbol{X}^*) = \frac{f(\boldsymbol{X}^*)g(\boldsymbol{x}|\boldsymbol{X}^*)}{f(\boldsymbol{x})g(\boldsymbol{X}^*|\boldsymbol{x})}$$

- Put

$$\boldsymbol{Y} = \begin{cases} \boldsymbol{X}^* & \text{with probability } \min\{1, R(\boldsymbol{x}, \boldsymbol{X}^*)\}; \\ \boldsymbol{x} & \text{otherwise.} \end{cases}$$

An essential requirement for a Markov chain to converge to a stationary distribution $\pi(\boldsymbol{x})$ is that

$$\pi(\boldsymbol{y}) = \int_{\boldsymbol{x}} \pi(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{x}. \qquad (*)$$

(a) Show that a *sufficient* requirement for (*) is the *detailed balance* criterion

$$\pi(\boldsymbol{y})P(\boldsymbol{x}|\boldsymbol{y}) = \pi(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}). \qquad (1)$$

$$\int_{\boldsymbol{x}} \pi(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{x} = \int_{\boldsymbol{x}} \pi(\boldsymbol{y})P(\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{x}$$
$$= \pi(\boldsymbol{y}) \int_{\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{x} = \pi(\boldsymbol{y})$$

Problem 2

Consider the Metropolis-Hastings algorithm where the transition densities $P(\boldsymbol{y}|\boldsymbol{x})$ are defined through:

- Sample a candidate value $\boldsymbol{X}^*$ from a *proposal distribution* $g(\cdot|\boldsymbol{x})$.
- Compute the Metropolis-Hastings ratio

$$R(\boldsymbol{x}, \boldsymbol{X}^*) = \frac{f(\boldsymbol{X}^*)g(\boldsymbol{x}|\boldsymbol{X}^*)}{f(\boldsymbol{x})g(\boldsymbol{X}^*|\boldsymbol{x})}$$

- Put

$$\boldsymbol{Y} = \begin{cases} \boldsymbol{X}^* & \text{with probability } \min\{1, R(\boldsymbol{x}, \boldsymbol{X}^*)\}; \\ \boldsymbol{x} & \text{otherwise.} \end{cases}$$

An essential requirement for a Markov chain to converge to a stationary distribution $\pi(\boldsymbol{x})$ is that

$$\pi(\boldsymbol{y}) = \int_{\boldsymbol{x}} \pi(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{x}. \qquad (*)$$

$$\pi(\boldsymbol{y})P(\boldsymbol{x}|\boldsymbol{y}) = \pi(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}). \qquad (1)$$

(b) Show that the Metropolis-Hastings algorithm satisfies the detailed balance criterion.

What other criteria are needed in order for the Markov chain to converge in distribution to $\pi(\boldsymbol{x})$?

$$\begin{aligned}\pi(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x}) &= \pi(\boldsymbol{x})g(\boldsymbol{y}|x)\min\left\{1, \frac{\pi(\boldsymbol{y})g(\boldsymbol{x}|\boldsymbol{y})}{f(\boldsymbol{x})g(\boldsymbol{y}|\boldsymbol{x})}\right\} \\ &= \min\{\pi(\boldsymbol{x})g(\boldsymbol{y}|\boldsymbol{x}), \pi(\boldsymbol{y})g(\boldsymbol{x}|\boldsymbol{y})\} \\ &= \pi(\boldsymbol{y})g(\boldsymbol{x}|\boldsymbol{y})\min\left\{\frac{\pi(\boldsymbol{x})g(\boldsymbol{y}|\boldsymbol{x})}{\pi(\boldsymbol{y})g(\boldsymbol{x}|\boldsymbol{y})}, 1\right\} = \pi(\boldsymbol{y})P(\boldsymbol{x}|\boldsymbol{y})\end{aligned}$$

Problem 2

Consider the Metropolis-Hastings algorithm where the transition densities $P(\boldsymbol{y}|\boldsymbol{x})$ are defined through:

- Sample a candidate value $\boldsymbol{X}^*$ from a *proposal distribution* $g(\cdot|\boldsymbol{x})$.
- Compute the Metropolis-Hastings ratio

$$R(\boldsymbol{x}, \boldsymbol{X}^*) = \frac{f(\boldsymbol{X}^*)g(\boldsymbol{x}|\boldsymbol{X}^*)}{f(\boldsymbol{x})g(\boldsymbol{X}^*|\boldsymbol{x})}$$

- Put

$$\boldsymbol{Y} = \begin{cases} \boldsymbol{X}^* & \text{with probability } \min\{1, R(\boldsymbol{x}, \boldsymbol{X}^*)\}; \\ \boldsymbol{x} & \text{otherwise.} \end{cases}$$

An essential requirement for a Markov chain to converge to a stationary distribution $\pi(\boldsymbol{x})$ is that

$$\pi(\boldsymbol{y}) = \int_{\boldsymbol{x}} \pi(\boldsymbol{x})P(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{x}.$$

What other criteria are needed in order for the Markov chain to converge in distribution to $\pi(\boldsymbol{x})$?

Also need irreducibility, that is it is possible to move from any state $\boldsymbol{x}$ to any other state $\boldsymbol{y}$ in a finite number of steps. You also need the chain to be aperiodic, but this will be fulfilled as long as there is a positive probability for not accepting a new proposal (which will always be the case except for some degerate situations)

In in infinite state space you also need reccuence, i.e. probability 1 for returning to any set of non-neglible measure

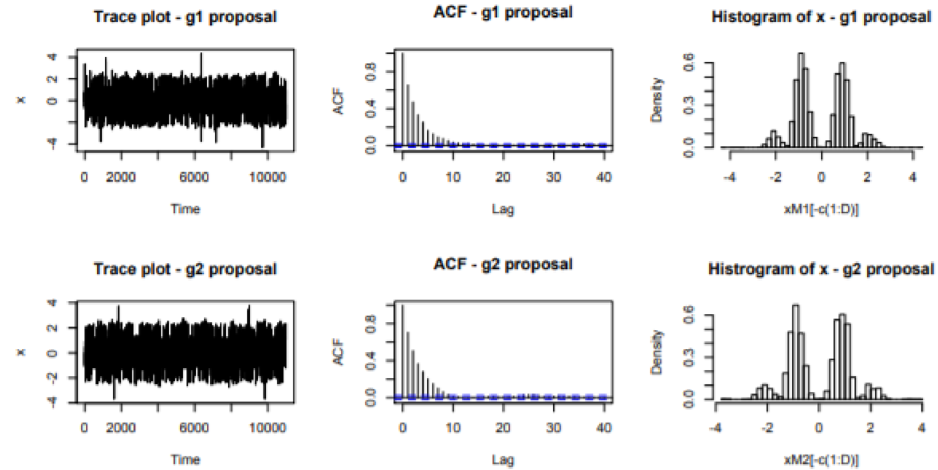
Assume now you want to use the Metropolis-Hastings algorithm to simulate from a one-dimensional distribution given by

$$\pi(x) \propto \sin^2(x) \cdot \sin^2(2x) \cdot \phi(x)$$

where $\phi(x)$ is the density for the standard Gaussian distribution. We will consider two different proposal distributions:

$$g_1(x^*|x) = N(0, \sigma_1^2);$$
$$g_2(x^*|x) = N(x, \sigma_2^2).$$

These two proposal distributions were run with $\sigma_1 = 3.5$ and $\sigma_2 = 2.5$. A total of 11000 iterations were run where the first 1000 were discarded. The plots below shows traceplots (including the first 1000 iterations), estimated autocorrelation functions and histograms for the two proposal distributions. The acceptance rates for the two proposal distributions were 0.24 and 0.28, respectively.



(c) For each of the two proposal distributions, derive formulas for the Metropolis-Hastings acceptance probabilities.

What type of Metropolis-Hastings algorithms do the g_1 and g_2 proposal distributions belong to?

Are the acceptance rates satisfactory for the two proposal distributions? If not, how would you recommend to change the proposal distributions?

For g_2 we obtain

$$R = \frac{\sin^2(x^*)\sin^2(2x^*)\phi(x^*;0,1)\phi(x;x^*,\sigma_2)}{\sin^2(x)\sin^2(2x)\phi(x;0,1)\phi(x^*;x,\sigma_2)} = \frac{\sin^2(x^*)\sin^2(2x^*)\phi(x^*;0,1)}{\sin^2(x)\sin^2(2x)\phi(x;0,1)}$$

$$R(\boldsymbol{x}, \boldsymbol{X}^*) = \frac{f(\boldsymbol{X}^*)g(\boldsymbol{x}|\boldsymbol{X}^*)}{f(\boldsymbol{x})g(\boldsymbol{X}^*|\boldsymbol{x})}$$

$$f(x) = \pi(x)$$

The g_1 proposal corresponds to an independent sampler while g_2 corresponds to random walk. For the first one we would like the acceptance rate to be as large as possible, indicating that it is too small in this case. One can try to change σ_1 to see when the acceptance probability is largest. Given that the standard gaussian distribution is involved in the target distribution, something closer to this distribution should be expected to give higher acceptance rate.

(c) For g_1:

$$R = \frac{\sin^2(x^*)\sin^2(2x^*)\phi(x^*;0,1)\phi(x;0,\sigma_1)}{\sin^2(x)\sin^2(2x)\phi(x;0,1)\phi(x^*;0,\sigma_1)}$$

For g_2, we want the acceptance rate to be somewhere between 0.25 and 0.50, which is ok in this case (perhaps we could increase the acceptance rate somewhat by decreasing the variance).

(d) Why would one discard the values obtained from the first iterations? Describe a general method for specifying the number of iterations to discard.

(d) A Markov chain typically needs some time before the simulated values are close to the target distribution. In order to reduce the bias, the first iterations should therefore be discarded.

Two possible methods for specifying the burnin:

- Looking at the trace plots and see if they have stabilized
- Calculate the Gelman-Rubin criterion, which, when running multiple chains, mainly compare within variability with between variability. This is a more formal criterion.

(e) For the two proposal distributions, the quantity $\sum_{k=0}^{\infty} \rho(k)$, with $\rho(k) = \text{cor}[(x^t)^2, (x^{t+k})^2]$ was estimated to be 3.36 and 3.73, respectively, for the g_1 and g_2 proposal distributions. How can these numbers be used to compare the two versions of the Metropolis-Hasting algorithm if estimation of $E[x^2]$ is of interest?

Note: both are squared

(e) If one wants to estimate $E[x^2]$, then the variance of the Monte Carlo estimate converges towards $\sigma^2[1 + 2\sum_{k=1}^{\infty} \rho(k)]$ where σ^2 corresponds to the variance for independent samples. The two proposal distributions give almost similar estimates on the second part, with a small preference to g_1.

One can also look at the effective sample size which is defined as

$$\frac{L}{1 + 2\sum_{k=1}^{\infty} \rho(k)}$$

where L is the number of samples used.

$$1 + 2\sum_{k=1}^{\infty} \rho(k) = \begin{cases} 1 + 2(3.36 - 1) = 5.72 & \text{for case 1} \\ 1 + 2(3.73 - 1) = 6.46 & \text{for case 2} \end{cases}$$

An alternative to the use of the Metropolis-Hastings algorithm in this case is the rejection sampling method. Assume $g_1(x)$ again is used as proposal distribution.

(f) Does this proposal distribution meet the requirements needed for constructing a proper rejection sampling algorithm?

(f) We have that

$$\frac{\pi(x)}{g_1(x)} = c\frac{\sin^2(x)\sin^2(2x)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}}{\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{1}{2\sigma_1^2}x^2}}$$
$$= c\sigma_1\sin^2(x)\sin^2(2x)e^{-\frac{1}{2}[1-\frac{1}{\sigma_1^2}]x^2} \leq c\sigma_1$$

showing that the ratio has a finite maximum. The requirements needed are then fulfilled.

(g) $N = 3000$ values of x were generated by this algorithm, giving a mean number of proposals equal to 16.9.

In order to estimate $E[x^2]$, would you prefer to use one of the Metropolis-Hastings algorithms or the rejection sampling algorithm. Justify your answer.

(g) For the M-H with q_1, the effective number of samples is estimated to be $10000/5.72 = 1748$ samples which is less than the number of samples generated by the rejection method. This means that we will get almost two times reduction in variance using rejection sampling

However, while M-H needed $10000 + 1000 = 11000$ samples to be generated, the rejection sampling required 16.9*3000=50700 samples, indicating that the computational effort with rejection sampling was much larger.

An argument towards rejection sampling compared to MCMC is however that the former is exact!