

MTH 511a - 2020: Lecture 29

Instructor: Dootika Vats

The instructor of this course owns the copyright of all the course materials. This lecture material was distributed only to the students attending the course MTH511a: “Statistical Simulation and Data Analysis” of IIT Kanpur, and should not be distributed in print or through electronic media without the consent of the instructor. Students can make their own copies of the course materials for their use.

From last week, it should be evident that accept-reject sampling is very difficult to implement in most practical Bayesian models. This makes it really difficult to estimate quantities of the posterior distribution without being able to get samples from it!

It is here where Markov chain Monte Carlo (MCMC) will be credibly useful and powerful.

1 Markov chain Monte Carlo

We want to sample from a distribution $\pi(x)$ which we may or may not know completely. If iid samples (like using accept-reject) are not available or not possible, we often use Markov chain Monte Carlo (MCMC).

Let $\pi(x)$ be the target distribution defined on a space \mathcal{X} . We will choose a starting value $X_1 \sim \pi(x)$.

In MCMC, we sample a *Markov chain* X_1, X_2, \dots with $\pi(x)$ as the *stationary distribution*. The Markov chain is determined by *Markov transition kernel*, where for a set $A \subseteq \mathcal{X}$, $P(x_1, \cdot)$ describes how the next step in the Markov chain will be obtained:

$$P(x_1, A) := \Pr(X_2 \in A \mid X_1 = x_1).$$

The kernel P satisfies

$$\int P(x_1, x_2) \pi(x_1) dx_1 = \pi(x_2).$$

That is, the next value (X_2) is chosen based on my current value, such that it is also a draw from π . In practice, it is difficult to sample $X_1 \sim \pi$, so a *reasonable starting value* is chosen.

In this way, we keep generating more and more samples. The sample obtained: X_1, X_2, \dots, X_T satisfies the following

- The samples are *all correlated*. So every new sample carries far less information than an iid sample
- Since we don't start from π , we actually do not sample exactly from π . In fact, the distribution of X_t converges to π . That is $X_\infty \sim \pi$. But usually, if X_1 is chosen well, we are very close to π reasonably quickly.

Since, $X_\infty \sim \pi$, a Markov chain law of large numbers holds, for any function g ,

$$\frac{1}{T} \sum_{t=1}^T g(X_t) \xrightarrow{p} E_\pi[g(X)] \quad \text{as } n \rightarrow \infty.$$

So even though, we have far from iid samples, we will still converge to the right quantity.

Question is, how can we construct such a Markov chain?

2 Metropolis-Hastings

The Metropolis-Hastings algorithm is the most common one to construct such a Markov chain. And is similar to accept-reject in principle, although very different. The Metropolis-Hastings algorithm functions as following.

To construct a Markov chain with target distribution π and given that the current sample drawn is x , choose a proposal distribution Q_x with density $q(y|x)$.

1. Choose a starting value x_1 that is in \mathcal{X} . For any iterate t do the following
2. Draw a proposed value $y^* \sim q(y^*|x_t)$.
3. Calculate

$$\alpha(x_t, y^*) = \min \left\{ 1, \frac{\pi(y^*) q(x_t|y^*)}{\pi(x_t) q(y^*|x_t)} \right\}$$

4. Draw $U \sim U[0, 1]$. If $U < \alpha(x_t, y^*)$. Set $x_{t+1} = y^*$.
5. Else, set $x_{t+1} = x_t$.

Remember that if we reject a value, then our next step is our previous step. The proposal distribution is usually taken so that it is symmetric in both x and y . That is, $q(y|x) = q(x|y)$. This means that the MH ratio is

$$\alpha(x_t, y^*) = \min \left\{ 1, \frac{\pi(y^*)}{\pi(x_t)} \right\}$$

Example 1 (Sampling from a circle). Recall our example of sampling from a circle and the problem set problem of sampling from a p -dimensional sphere. As the dimension of the sphere increases, the acceptance probability of the iid accept-reject algorithm reduces (exponentially). Although demonstrating this in p -dimensions will be difficult, we will demonstrate this for sampling from a circle.

The target density is

$$f(x, y) = \frac{1}{\pi} I(x^2 + y^2 < 1).$$

We consider the proposal distribution to a small box around around our current step. That is, given a step (x_t, y_t) , the proposal distribution is $U[x_t - h, x_t + h] \times [y_t - h, y_t + h]$. The density of the proposal distribution is:

$$g(w, z|x_t, y_t) = \frac{1}{4h} I(x_t - h < w < x_t + h) I(y_t - h < z < y_t + h).$$

Since the proposal distribution is symmetric, to implement the MH algorithm, we need the ratio Consider

$$\frac{f(w, z)}{f(x_t, y_t)} = I(w^2 + z^2 < 1) \leq 1$$

So, the MH algorithm here can be implemented in the following way:

1. Set (X_0, Y_0) to be any coordinate inside the circle.
2. Generate $(W, Z) \sim U[x_t - h, x_t + h] \times [y_t - h, y_t + h]$.
3. If $I(W^2 + Z^2 < 1)$, then set $(X_{t+1}, Y_{t+1}) = (W, Z)$
4. Else set $(X_{t+1}, Y_{t+1}) = (X_t, Y_t)$

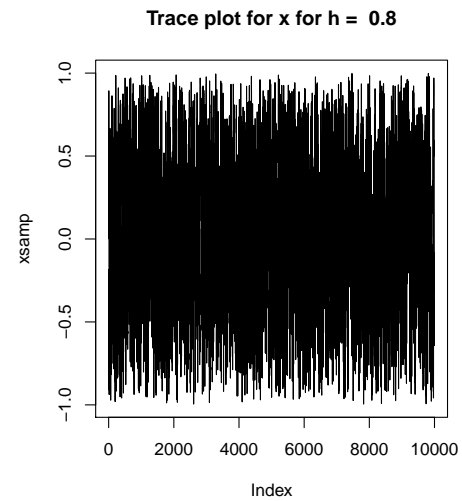
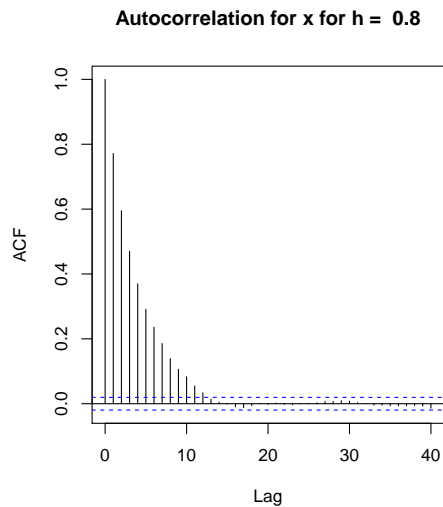
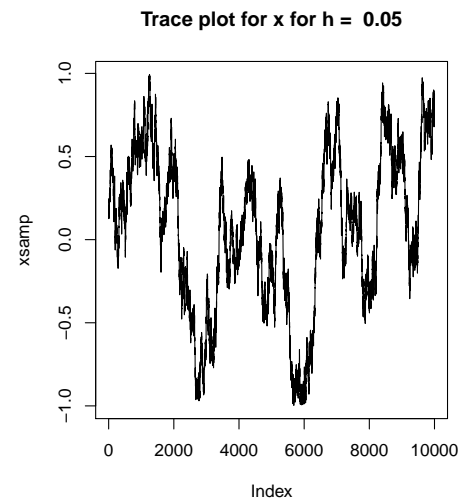
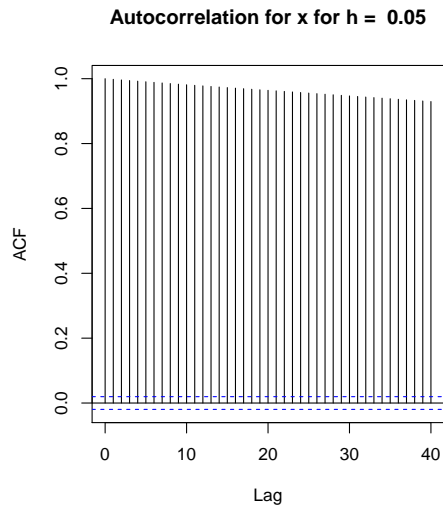
[Run simulation separately to show animation from this.](#)

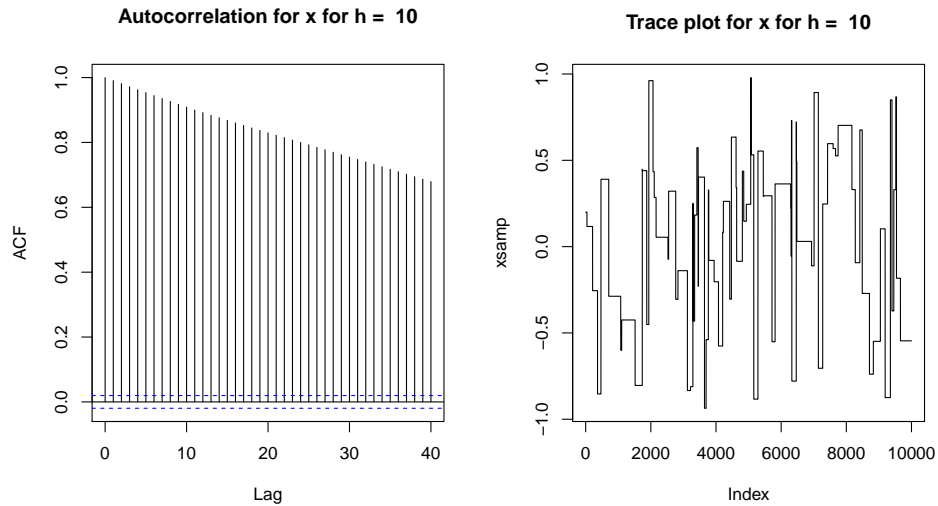
From the above example it is clear that the quality of the MCMC sampler is determined by the choice of h .

- If h is too large, then we will propose values potentially too far away, which in principle would be good, but this means a lot of the values will be rejected. That means we will stay where we are quite often, increasing the correlation in the sample.
- If h is too small, then we will make tiny moves implying we may accept a lot of them, but the samples obtained will be quite dependent.

Two visual aides to assess the performance of the sampler are

- Autocorrelation plots (ACF plots): these tell us upto how many steps (on average) is there correlation in the Markov chain. The higher the ACF correlation, the worst it is.
- Trace plot: Trace plots indicate how well the Markov chain moves. We want to see a white noise (random behavior).





Clearly the performance of the MCMC sampler can be highly impacted by the variance in the proposal distribution. A good rule of thumb is:

Aim for 45% acceptance probability for 1-dimensional target distributions and 23% acceptance probability for high-dimensional target distributions (> 5). Anything in between you can choose between 23-45%.

3 Questions to think about

- What would happen if you choose the starting point to be outside the support of the target distribution?
- In general, how can you choose a starting values?
- Although a law of large numbers holds, do you think a central limit theorem can also hold here?
- Implement the MCMC algorithm for the p -dimensional sphere problem and compare with the efficiency of the iid accept-reject sampler.