# 1 Markov Chain Monte Carlo (MCMC) Algorithms

The MCMC algorithms, and in particular the Metropolis–Hastings (MH) algorithm, are powerful simulation tools to simulate samples from any complex distribution, known up to a scaling constant, at the expense of introducing dependence between the samples. The Metropolis-Hastings algorithm was first developed by Metropolis and his co-workers and then generalized by Hastings. In MCMC and also in the MH algorithms, the samples are generated as states of a special Markov chain whose limiting stationary distribution is the target posterior PDF. For this useful properties of a Markov Chain are next discussed followed by the description of the Metropolis-Hasting algorithm.

## 1.1 Markov Chains

A series of random variables $\{\underline{\theta}^{(1)},\underline{\theta}^{(2)},\cdots,\underline{\theta}^{(N)}\}$ constitute a first-order Markov Chain if the following independence property holds for the conditional probability distributions:

$$p(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)},\ldots,\underline{\theta}^{(1)}) = p(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)})$$

where $p(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)})$ is defined to be the transition probability $t(\underline{\theta}^{(j)},\underline{\theta}^{(j+1)})$ of the Markov chain which is homogeneous if the transitional probabilities are the same, independent of $j$. Using marginalization theorem, the marginal distribution of the next random variable $\underline{\theta}^{(j+1)}$ in the chain is related to the marginal distribution of the current random variable $\underline{\theta}^{(j)}$ as

$$p(\underline{\theta}^{(j+1)}) = \sum_{\underline{\theta}^{(j)}} p(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)})\,p(\underline{\theta}^{(j)}) = \sum_{\underline{\theta}^{(j)}} t(\underline{\theta}^{(j)},\underline{\theta}^{(j+1)})\,p(\underline{\theta}^{(j)})$$

For a homogeneous Markov chain with same marginal probability distribution $p(\underline{\theta})$ for each random variable in the chain, the following is true

$$p(\underline{\theta}) = \sum_{\underline{\theta}'} t(\underline{\theta}',\underline{\theta})\,p(\underline{\theta}') \qquad (1)$$

In this case the probability distribution $p(\underline{\theta})$ is said to be an invariant distribution for the Markov chain.

A probability distribution $p(\underline{\theta})$ is invariant if the transition probability $t(\underline{\theta}^{(j)},\underline{\theta}^{(j+1)})$ from state $j$ to state $j+1$ of the Markov process satisfies the **detailed balance** equation given by

$$p(\underline{\theta})\; t(\underline{\theta},\underline{\theta}') = p(\underline{\theta}')\; t(\underline{\theta}',\underline{\theta})$$

or equivalently

$$p(\underline{\theta}'\,|\,\underline{\theta})\; p(\underline{\theta}) = p(\underline{\theta}\,|\,\underline{\theta}')\; p(\underline{\theta}')$$

This can be shown by starting with the right hand side of (1) and using detailed balance to arrive at the left hand side of (1) as follows

$$\sum_{\underline{\theta}'} t(\underline{\theta}',\underline{\theta})p(\underline{\theta}') = \sum_{\underline{\theta}'} t(\underline{\theta},\underline{\theta}')p(\underline{\theta}) = p(\underline{\theta})\sum_{\underline{\theta}'} t(\underline{\theta},\underline{\theta}') = p(\underline{\theta})\sum_{\underline{\theta}'} p(\underline{\theta}' \,|\, \underline{\theta}) = p(\underline{\theta})$$

A Markov chain with distribution that satisfies the detailed balance is called reversible.

To sample from a distribution $p(\underline{\theta})$ using Markov chain we need to select a transition distribution which satisfies the detailed balance equation.

## 1.2 Metropolis-Hasting Algorithm

Consider the posterior distribution $p(\underline{\theta} \,|\, D,I)$ of a set of uncertain parameters $\underline{\theta}$. The objective in the MCMC is to generate a Markov chain (MC), i.e. a sequence of points $\{\underline{\theta}^{(1)},\underline{\theta}^{(2)},\cdots,\underline{\theta}^{(N)}\}$ in the parameter space that are samples of the posterior distribution $p(\underline{\theta} \,|\, D,I)$. The next sample $\underline{\theta}^{(j+1)}$ is generated from the current sample $\underline{\theta}^{(j)}$ using a proposal distribution $q(\underline{\xi} \,|\, \underline{\theta}^{(j)})$ which depends on the current state $\underline{\theta}^{(j)}$. The Metropolis-Hasting algorithm generates these points by applying the following steps.

- The first sample point $\underline{\theta}^{(1)}$ in the parameter space $\underline{\theta}$ is set at an arbitrary initial value.
- The following procedure is repeated for $j = 1,\ldots,N-1$,

  1. the next sample $\underline{\theta}^{(j+1)}$ is generated/simulated from the current sample $\underline{\theta}^{(j)}$ by simulating a candidate state $\underline{\xi}$ from a *proposal PDF* $q(\underline{\xi} \,|\, \underline{\theta}^{(j)})$ and evaluating the probability density quotient

  $$Q(\underline{\xi},\underline{\theta}^{(j)}) = \frac{p(\underline{\xi} \,|\, D,I)}{p(\underline{\theta}^{(j)} \,|\, D,I)} \frac{q(\underline{\theta}^{(j)} \,|\, \underline{\xi})}{q(\underline{\xi} \,|\, \underline{\theta}^{(j)})} \tag{2}$$

  2. The candidate state $\underline{\xi}$ is then accepted with probability

  $$a(\underline{\xi},\underline{\theta}^{(j)}) = \min\{1, Q(\underline{\xi},\underline{\theta}^{(j)})\} \tag{3}$$

  and rejected with probability $1 - a(\underline{\xi},\underline{\theta}^{(j)}) = \max\{0, 1 - Q(\underline{\xi},\underline{\theta}^{(j)})\}$. Specifically, the next sample in the Markov chain is

  $$\underline{\theta}^{(j+1)} = \begin{cases} \underline{\xi} & \text{with probability} \quad a(\underline{\xi},\underline{\theta}^{(j)}) \\ \underline{\theta}^{(j+1)} & \text{with probability} \quad 1\text{-}a(\underline{\xi},\underline{\theta}^{(j)}) \end{cases} \tag{4}$$

  This means that if the candidate sample is accepted, the next sample in the Markov chain is $\underline{\theta}^{(j+1)} = \underline{\xi}$. Otherwise, the current state remains as the next sample $\underline{\theta}^{(j+1)} = \underline{\theta}^{(j)}$ in the Markov chain.

If the proposal PDF is symmetric, that is, $q(\underline{\theta} \,|\, \underline{\xi}) = q(\underline{\xi} \,|\, \underline{\theta})$ for every $\underline{\xi}$ and $\underline{\theta}$, then the acceptance ratio (probability) becomes

$$a(\underline{\xi},\underline{\theta}^{(j)}) = \min\left\{1, \frac{p(\underline{\xi}\,|\,D,I)}{p(\underline{\theta}^{(j)}\,|\,D,I)}\right\} \tag{5}$$

which is the ratio of the posterior PDF values at the candidate state and the current state. Note that the value of the proposal PDF does not play a role on the acceptance ratio and the sample generation. Note that if the candidate state $\underline{\xi}$ is more probable than the current state $\underline{\theta}^{(j)}$, i.e. $p(\underline{\xi}\,|\,D,I) \geq p(\underline{\theta}^{(j)}\,|\,D,I)$, then the candidate state $\underline{\xi}$ is always accepted.

The second step is implemented by generating a sample $u$ from a uniform distribution in the interval [0.1]. If $u \leq a(\underline{\xi},\underline{\theta}^{(j)})$ then the move is accepted and the candidate state $\underline{\xi}$ is accepted as the new state in the Markov chain, i.e. $\underline{\theta}^{(j+1)} = \underline{\xi}$. If $u > a(\underline{\xi},\underline{\theta}^{(j)})$ then the move is not allowed and the new state in the Markov chain is same with the current state, i.e. $\underline{\theta}^{(j+1)} = \underline{\theta}^{(j)}$.

The samples of the Markov chain are dependent and they are used for statistical averaging as they were independent, causing some reduction in the efficiency of the estimator. Specifically, the posterior robust prediction integral (3), given by

$$I = E[h(\underline{\theta})] = \int h(\underline{\theta})\, p(\underline{\theta}\,|\,D,I)\, d\underline{\theta}$$

can be approximated by the sample estimate

$$E[h(\underline{\theta})] \approx \hat{I}_N = \frac{1}{N}\sum_{j=1}^{N} h(\underline{\theta}^{(j)})$$

where $\underline{\theta}^{(j)}$, $j = 1,\ldots,N$, are the MC samples drawn from the posterior PDF $p(\underline{\theta}\,|\,D,I)$ using the MH algorithm. Alternatively, practically independent samples can be obtained by retaining every $m$-th sample in the sequence, where $m$ is selected to be sufficiently large.

Graphical Demonstration of Acceptance\Rejection of Candidate State: Figures 1 to 3 show contour plots of the posterior PDF in 2-dim parameter space. The acceptance of a candidate state depends on the location of candidate state $\underline{\xi}$ in the parameter space in relation to the current sample $\underline{\theta}^{(j)}$. In all the three figures, the contour values scale from low (blue ) to high posterior pdf values (red)

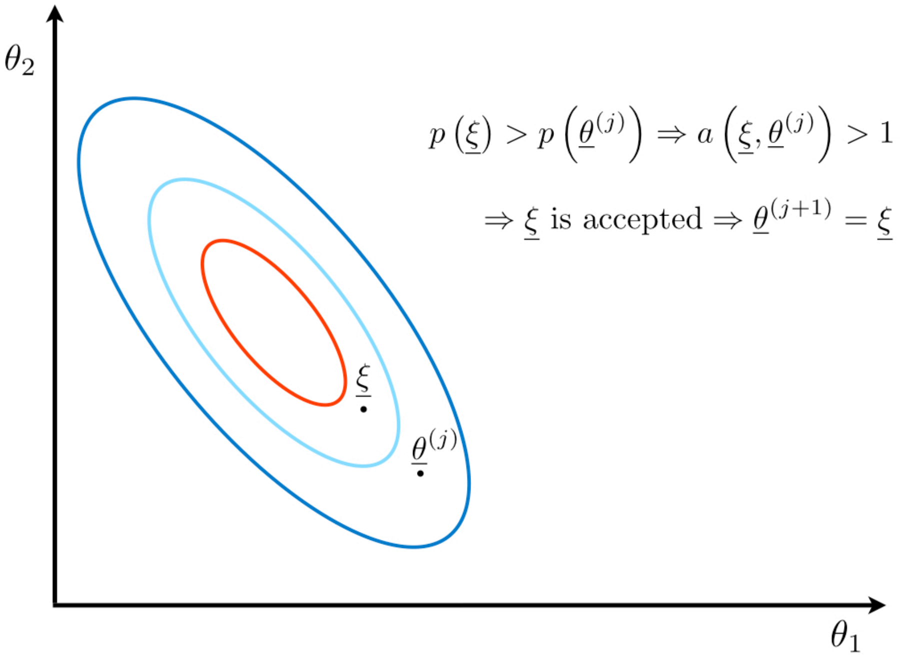
$\theta_2$
$p(\underline{\xi}) > p\left(\underline{\theta}^{(j)}\right) \Rightarrow a\left(\underline{\xi}, \underline{\theta}^{(j)}\right) > 1$
$\Rightarrow \underline{\xi}$ is accepted $\Rightarrow \underline{\theta}^{(j+1)} = \underline{\xi}$
$\underline{\xi}$
$\underline{\theta}^{(j)}$
$\theta_1$

**Figure 1:** Always accepted candidate sample

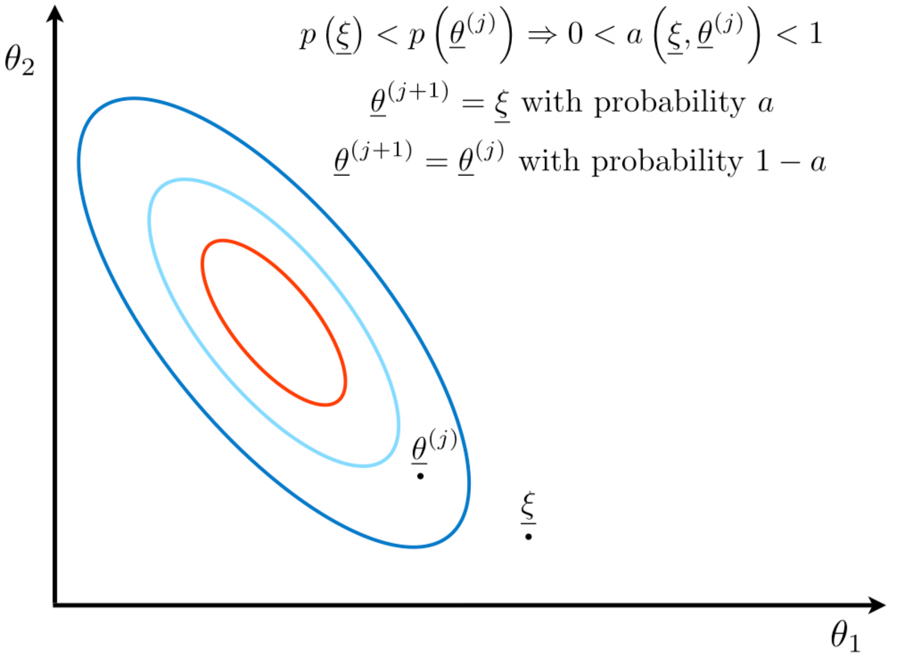
$\theta_2$
$p(\underline{\xi}) < p\left(\underline{\theta}^{(j)}\right) \Rightarrow 0 < a\left(\underline{\xi}, \underline{\theta}^{(j)}\right) < 1$
$\underline{\theta}^{(j+1)} = \underline{\xi}$ with probability $a$
$\underline{\theta}^{(j+1)} = \underline{\theta}^{(j)}$ with probability $1 - a$
$\underline{\theta}^{(j)}$
$\underline{\xi}$
$\theta_1$

**Figure 2**: Candidate sample accepted with acceptance probability $a(\underline{\xi}, \underline{\theta}^{(j)})$

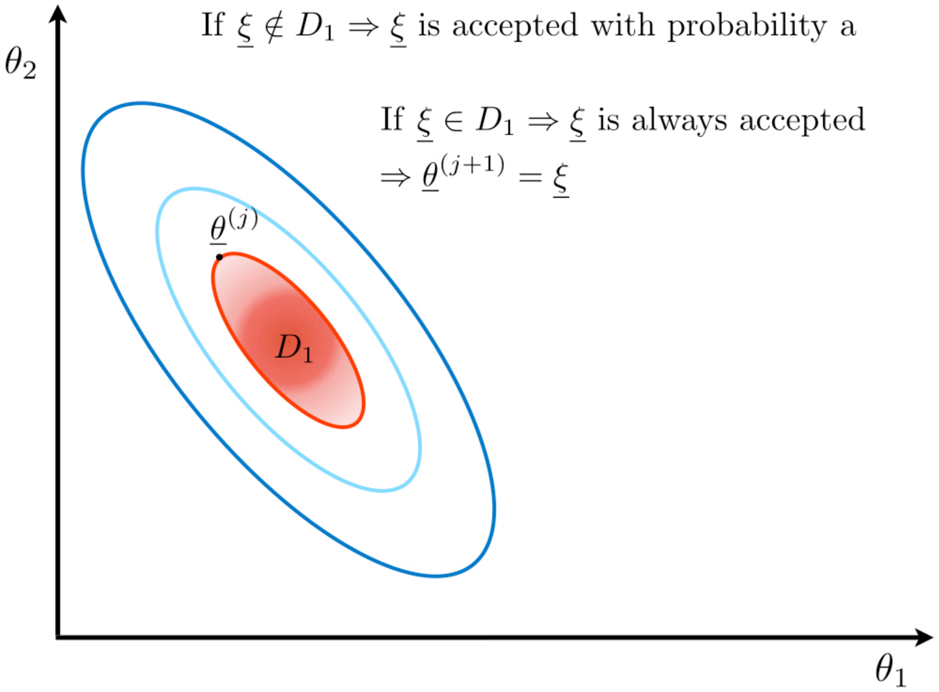
If $\underline{\xi} \notin D_1 \Rightarrow \underline{\xi}$ is accepted with probability a
If $\underline{\xi} \in D_1 \Rightarrow \underline{\xi}$ is always accepted
$\Rightarrow \underline{\theta}^{(j+1)} = \underline{\xi}$
$\underline{\theta}^{(j)}$
$D_1$
$\theta_2$
$\theta_1$

**Figure 3:** Candidate samples falling within the domain $D_1$ in the parameter space are always accepted. Candidate sample that fall in the contour plot $C_1$ are accepted with the same acceptance probability $a(\underline{\xi},\underline{\theta}^{(j)})$

To show that the Markov chain samples are drawn from the distribution $p(\underline{\theta}\,|\,D,I)$ we need to show that the distribution $p(\underline{\theta}\,|\,D,I)$ is the invariant distribution of the Markov chain defined by the MH algorithm. Specifically, it can be shown that the next sample $\underline{\theta}^{(j+1)}$ in the Markov chain is distributed according to the target posterior PDF, provided that the current sample is also distributed according to the same PDF. That is, the posterior PDF is the stationary PDF of a Markov chain. By construction the transition probability of the MH samples is

$$t(\underline{\theta}^{(j)},\underline{\theta}^{(j+1)}) = q(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)})\; a(\underline{\theta}^{(j+1)},\underline{\theta}^{(j)})$$

To show that the detailed balance equation is satisfied by this transition probability note that

$$
\begin{aligned}
p(\underline{\theta}^{(j)}\,|\,D,I)\; t(\underline{\theta}^{(j)},\underline{\theta}^{(j+1)}) &= p(\underline{\theta}^{(j)}\,|\,D,I)\; q(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)})\; a(\underline{\theta}^{(j+1)},\underline{\theta}^{(j)}) \\
&= p(\underline{\theta}^{(j)}\,|\,D,I)\; q(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)})\; \min\left\{1, \frac{p(\underline{\theta}^{(j+1)}\,|\,D,I)}{p(\underline{\theta}^{(j)}\,|\,D,I)} \frac{q(\underline{\theta}^{(j)}\,|\,\underline{\theta}^{(j+1)})}{q(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)})}\right\} \\
&= \min\left\{ p(\underline{\theta}^{(j)}\,|\,D,I)\; q(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)})\;, p(\underline{\theta}^{(j+1)}\,|\,D,I)\; q(\underline{\theta}^{(j)}\,|\,\underline{\theta}^{(j+1)}) \right\} \\
&= \min\left\{ p(\underline{\theta}^{(j+1)}\,|\,D,I)\; q(\underline{\theta}^{(j)}\,|\,\underline{\theta}^{(j+1)}), p(\underline{\theta}^{(j)})\; q(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)}) \right\} \\
&= p(\underline{\theta}^{(j+1)}\,|\,D,I)\; q(\underline{\theta}^{(j)}\,|\,\underline{\theta}^{(j+1)}) \min\left\{1, \frac{p(\underline{\theta}^{(j)}\,|\,D,I)\; q(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)})}{p(\underline{\theta}^{(j+1)}\,|\,D,I)\; q(\underline{\theta}^{(j)}\,|\,\underline{\theta}^{(j+1)})}\right\} \\
&= p(\underline{\theta}^{(j+1)}\,|\,D,I)\; q(\underline{\theta}^{(j)}\,|\,\underline{\theta}^{(j+1)})\; a(\underline{\theta}^{(j)},\underline{\theta}^{(j+1)}) \\
&= p(\underline{\theta}^{(j+1)}\,|\,D,I)\; t(\underline{\theta}^{(j+1)},\underline{\theta}^{(j)})
\end{aligned}
$$

Note that the choice of the proposal distribution $q(\underline{\theta}^{(j+1)}\,|\,\underline{\theta}^{(j)})$ affects the performance of the algorithm.

## 1.3 Remarks

1. Initially, the Markov chain started at an arbitrary state $\underline{\theta}^{(1)}$ will be in a transient state, with the sample distribution differing from the target posterior PDF. The Markov chain will converge to the stationary state and the sample $\underline{\theta}^{(j)}$ will tend to the target posterior PDF as $j \to \infty$.

2. The MCMC methods apply to any distribution that is known up to a scaling constant. Herein it was applied to the posterior distribution encountered in Bayesian analysis. Note that the scaling constant (the evidence in Bayesian analysis) is not required to be known. The evidence is given by a multi-dimensional integral and it is not known. Thus, the MCMC-MH algorithm is very powerful technique for generating samples from the posterior PDF, since it bypasses the problem of computing the evidence. Specifically, substituting the target posterior PDF in (5) the acceptance probability $a(\underline{\xi},\underline{\theta}^{(j)})$ becomes

$$
a(\underline{\xi},\underline{\theta}^{(j)}) = \min\left\{1, \frac{f(D\,|\,\underline{\xi},I)\; f(\underline{\xi}\,|\,I)}{f(D\,|\,\underline{\theta}^{(j)},I)\; f(\underline{\theta}^{(j)}\,|\,I)}\right\}
$$

   which depends only on the ratio of the product of the likelihood and the prior distribution between the candidate state and the current state.

3. The selection of the form of the proposal PDF is arbitrary. However, the proposal PDF should be selected so that it is easy to sample from. The MH method replaces the difficult sample generation from the posterior PDF by the many generations from the proposal PDF. A Gaussian proposal PDF, centered at the current state $\underline{\theta}^{(j)}$ and with appropriately selected covariance matrix, can be a useful choice. The selection of the covariance matrix of the proposal PDF is also crucial since it affects the moves that the candidate sample makes in the parameter space. These moves have to cover the important domain of high probability in the parameter in reasonable computing time for all samples in the chain. The proposal PDF is usually selected to be in a form that samples are readily generated from the proposal PDF. For example it can be chosen to be Gaussian, uniform, etc.

4. Acceptance rates from 20% to 50% are reported as reasonable rates. For Gaussian distributions with Gaussian proposals theoretically results have been obtained about optimal acceptance rates. For one-dimensional distributions, the scale of the proposal distribution should be tuned to give an acceptance rate of 45%. For high-dimensional distributions the scale should be tuned to achieve an acceptance

rate of 24%, while for six and infinite-dimensional distributions the optimal acceptance rate is 25% and 23%, respectively.

5. The MCMC method is general and applies to any distribution, not just posterior PDF arising in Bayesian analysis.
6. The first sample point in the Markov chain can be an arbitrary point in the parameter space generated by a PDF that approximates the posterior $f(\underline{\theta} \mid D,I)$ or it can be chosen deterministically, for example, to be the most probable value of the posterior PDF obtained by solving an optimization problem. The first sample in the Markov chain can be selected to be the mode of the posterior distribution, estimated by minimizing the minus logarithm of the posterior distribution.
7. Other sampling methods are:
    a. **Gibbs Sampling**: is a special case of MH sampling
    b. **Slice Sampling**: with adaptive step size adjusting automatically to match the characteristics of the explored distribution
    c. **Hybrid Monte Carlo sampling**

### 1.4 Demonstration Examples

A series of movies shown in class, demonstrate the **effectiveness of the MH algorithms for various choices of the proposal PDFs**. For simplicity, the likelihood is chosen to be a Gaussian distribution in $\underline{\theta} \in R^2$ and the prior distribution is chosen to be uniform with large enough bound so that they do not affect the form of the posterior. That is the posterior PDF is proportional to the two-dimensional Gaussian likelihood:

$$f(\underline{\theta} \mid D,I) = N(\underline{\theta};\underline{\mu},\Sigma)$$

where the mean is selected to be $\underline{\mu} = \underline{0}$ and the covariance matrix is diagonal $\Sigma = diag(\sigma_1^2,\sigma_2^2)$ with variances $\sigma_1^2$ and $\sigma_2^2$. In the example, the standard deviations are chosen to be $\sigma_1 = 5$ and $\sigma_2 = 1$.

The proposal PDF determines the distribution of the candidate state $\underline{\xi}$ given the current state $\underline{\theta}^{(j)}$. As a result, the proposal PDF will also determine the convergence rate of the estimator $\hat{I}_N$.

Note that the Markov chain may remain at the same state for many consecutive samples. The percentage of candidate states that are accepted constitutes a useful monitoring index. This index is the *acceptance rate* of candidate states that should be computed and reported in practical applications. For the method to be successful, the acceptance rate should not be very low. Obviously, in order to increase the acceptance rate one needs to make very small moves around the candidate sample. In this way, the acceptance probability will be very close to one and the candidate samples/states will have a very high acceptance rate. However, in order to converge to the "equilibrium" posterior distribution, the Markov chain must be capable to transverse/explore the whole parameter space with significant probability volume. Very small moves will require a very large number of iterations for the chain to converge. For insufficiently number of samples in the Markov chain, the region visited by the Markov chain samples will be small compared to the important region of high probability in the parameter space, slowing down convergence significantly and also leading to substantial bias in the estimate $\hat{R}$ of robust prediction integrals.

**MOVIES**: Report also 1. acceptance rate, 2. Time History, 3. Correlation. For 200 samples (live animation) and 2000 (or 10000) samples (end result).

MOVIE 1: Isotropic normal proposal with size of **proposal PDF** $N(\underline{\xi};\underline{\theta}^{(j)},s^2 I)$ to be **one half** of the size of the **posterior PDF**. Starting point **outside (with burn-in samples)** posterior PDF. $s = \sigma_2 / 2$

MOVIE 2: Isotropic normal proposal with size of **proposal PDF** $N(\underline{\xi};\underline{\theta}^{(j)},s^2I)$ to be **one half** of the size of the **posterior PDF**. Starting point **inside** posterior PDF. $s=\sigma_2/2$

MOVIE 3: Isotropic normal proposal with size of **proposal PDF** $N(\underline{\xi};\underline{\theta}^{(j)},s^2I)$ to be **one twentieth** of the size of the **posterior PDF**. Starting point **inside** posterior PDF. $s=\sigma_2/20$

MOVIE 4: Isotropic normal proposal with size of **proposal PDF** $N(\underline{\xi};\underline{\theta}^{(j)},\sigma^2I)$ to be **30 times** of the size of the **posterior PDF**. Starting point **inside** posterior PDF. $s=10\sigma_1$

MOVIE 5: Non-isotropic normal proposal with size of **proposal PDF** $N(\underline{\xi};\underline{\theta}^{(j)},diag(s_1^2,s_2^2))$ to be **one third** of the size of the **posterior PDF**. Starting point **outside** posterior PDF. $s_1=\sigma_1/3$ and $s_2=\sigma_2/3$.

MOVIE 6: Non-isotropic normal proposal with size of **proposal PDF** $N(\underline{\xi};\underline{\theta}^{(j)},diag(s_1^2,s_2^2))$ to be **one third** of the size of the **posterior PDF**. Starting point **outside** posterior PDF. $s_1=\sigma_2/3$ and $s_2=\sigma_1/3$.

In contrast, too large moves/displacements are likely to always fall in the tails of the posterior distribution, causing very low acceptance rate due to small values of the acceptance probability. As a result, the Markov chain will contain a large number of repeated samples, causing large correlation among samples and slowing down the convergence of the estimator $\hat{I}_N$.

The choice of the appropriate proposal PDF is a critical issue to accelerate convergence with the least number of samples in the Markov chain. However, such choice is problem dependent. A number of techniques, such as Adaptive Metropolis, have been devised to adapt the proposal PDF based on the samples obtained up to the current state in the Markov chain.

## OTHER PROBLEMS WITH THE USE OF MCMC-MH algorithm

MOVIE 7 (Very peaked posterior PDF): Normal proposal with size of **proposal PDF** $N(\underline{\xi};\underline{\theta}^{(j)},\sigma^2I)$ to be **one third** of the size of the **posterior PDF**. Starting point **outside** posterior PDF. $s=20\sigma_1$, $\sigma_1=0.5$, $\sigma_2=0.1$. Demonstrate that MCMC-MH is unable to find the important support of the posterior PDF.

MOVIE 8 (**Multi-modal posterior PDF**): Non-isotropic normal proposal with size of **proposal PDF** $N(\underline{\xi};\underline{\theta}^{(j)},\sigma^2I)$ to be **one third** of the size of the **posterior** **bimodal** **PDF**. Starting point **outside** posterior PDF. Demonstrate that MCMC-MH converges/explores only one of the two regions of high probability content in this case. Cannot populate both regions of high probability content.

MOVIE 9 (**Multi-modal posterior PDF**): Isotropic normal proposal with size of **proposal PDF** $N(\underline{\xi};\underline{\theta}^{(j)},\sigma^2I)$ to be **significantly larger than the size of the support** of the **posterior** **bimodal** **PDF**. Starting point **outside** posterior PDF. Demonstrate that MCMC-MH converges very slowly to both regions of high probability content.