1 Stochastic Simulation Algorithms for Bayesian computation

1.1 Introduction

So far we discussed asymptotic methods for Bayesian computations, that is (a) representing the posterior PDF

$$p(\underline{\theta}|D,I) = \frac{p(D|\underline{\theta},I)p(\underline{\theta}|I)}{p(D|I)} \tag{1}$$

by a Gaussian distribution using the Bayesian Central Limit Theorem, where

$$p(D|I) = \int_{\Theta} p(D|\underline{\theta},I)p(\underline{\theta}|I)d\underline{\theta} \tag{2}$$

is the evidence, and (b) carrying out robust predictions of an output quantity of interest (QoI) $h(\underline{\theta})$, formulated by multi-dimensional integrals of the form

$$E\left[h(\underline{\theta})\right] = \int_{\Theta} h(\underline{\theta})p(\underline{\theta}|D,I)d\underline{\theta} \tag{3}$$

with Θ being the domain of definition in the parameter space of $\underline{\theta}$. The integral in (3) is one of the measures of uncertainty of the output QoI $h(\underline{\theta})$. It provides the expectetion of the output QoI with respect to the posterior distribution $p(\underline{\theta}|D,I)$.

Asymptotic methods involve solving optimization problems as well as computing the Hessian of the log of the posterior PDF. We have discussed in detail the problems that arise and also the fact that the asymptotic methods are local methods and thus provide approximate estimates.

Stochastic simulation methods such as variants of Monte Carlo algorithms are powerfull tools in numerically representing the posterior PDF in (1) with samples drawn from the posterior distribution and also using these samples to compute probability integrals of the type (2) and (3). These integrals are of the general form

$$E\left[h(\underline{\theta})\right] = \int_{\Theta} h(\underline{\theta})p(\underline{\theta})d\underline{\theta} \tag{4}$$

In the general case these integrals cannot be evaluated using analytical techniques. Sampling methods provide useful techniques to approximate the value of the integral, where the samples are drawn from the distribution $p(\underline{\theta})$. The problem of sampling from a distribution is thus important. This problem is presented first and then it is used to approximate the value of the integral by the **sampling estimate**.

1.2 Sampling from Distribution

[Reference Book: Rubenstein(1981) - Monte Carlo Methods]

Standard Uniform Distribution (SUD)

Let U be a uniformly distributed random variable. The PDF is given by

$$p_U(u) = \begin{cases} 1, & \text{if } 0 \leq u \leq 1. \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Let $u^{(i)}$, $i = 1, \cdots, N$ be the random samples drawn from the SUD. Any commercial software program can generate samples (pseudo-random number) from a SUD $p_U(u)$. It is assumed herein that such samples are available. Details for generating samples or psuedo-random numbers can be found in a number of reference (e.g. Rubenstein 1981). The SUD samples are useful for generating samples from arbitrary distributions as it is duiscussed next.

Sampling from Arbitrary Distributions

Let X be a random variable (RV) that follows a PDF $p_X(x)$ and cumulative density function (CDF) $F_X(x)$. Note that

$$F_X(x) = \int_{-\infty}^{s} p_X(s)ds \tag{6}$$

Samples from the PDF $p_X(x)$ are generated using the **Inverse Transform Sampling** Method, which makes use of the samples $u^{(i)}$, $i = 1, \cdots, N$. The sample generation is based on the following considerations. A transformation between the values of the random variables X and U is introduced as follows:

$$x = g(u) \tag{7}$$

and we seek the function $g(u)$ such that the random variable X follows the desired distribution $p_X(x)$. Given $x = g(u)$, it is known that the PDFs $p_X(x)$ and $p_U(u)$ are given by

$$p_X(x)dx = p_U(u)du \tag{8}$$

or, equivalently, the PDF $p_X(x)$ is

$$p_X(x) = p_U(u)\left|\frac{du}{dx}\right| = p_U(u)\left|\frac{dg(x)}{du}\right|^{-1} \tag{9}$$

Since U is uniform, one has that $p_U(u) = 1$ for $u \in [0, 1]$ and integration of (8) yields

$$\int_{-\infty}^{x} p_X(s)ds = \int_{0}^{u} p_U(s)ds = u \tag{10}$$

which, by making use of (6), one derives that

$$F_X(x) = u \tag{11}$$

Thus the following transformation is true

$$x = F_X^{-1}(u) \tag{12}$$

which means that the values of x are given by the inverse of the CDF $F_X(x)$ which also this inverse represents exaclty the function $g(u)$. This transformation allows to draw samples $x^{(i)}$, $i = 1, \cdots, N$ from the PDF $p_X(x)$ as

$$x^{(i)} = F_X^{-1}(u^{(i)}) \tag{13}$$

where $u^{(i)}$ are samples from the SUD.

Example 1: Use the inverse transform sampling method to sample from the exponential distribution

$$p(x) = \lambda e^{-\lambda x}, x > 0 \tag{14}$$

The CDF of the exponential distribution is

$$F(x) = \int_0^x p(x)dx = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x} \tag{15}$$

Thus the transformation $x = g(u)$ which is obtained from $x = F^{-1}(u)$ or equivalently $F(x) = u$ is derived by solving

$$1 - e^{-\lambda x} = u \tag{16}$$

with respect to x to yield

$$x = -\frac{1}{\lambda} ln(1 - u) \tag{17}$$

The transformation (17) between the random variable X and the standard uniform variable U defines an exponentially distributed random variable X and it allows to draw samples $x^{(i)}$, $i = 1, \ldots, N$ from the PDF $p(x)$ as

$$x^{(i)} = -\frac{1}{\lambda} ln(1 - u^{(i)}) \tag{18}$$

where $u^{(i)}$, $i = 1, \ldots, N$ are samples drawn from the standard uniform distribution.

Example 2: Use the inverse transform sampling method to sample from the standard Gaussian distribution.

1.3 Remarks

(1) The inverse transform sampling method requires the inversion of the CDF $F_X(x)$. This may be time consuming for cases where $F_X^{-1}(u)$ is not known in closed form as, for example, the Normal distribution.

(2) Alternatively, we can use the **Rejection Sampling** method or the **Importance Sampling** method to generate samples from a distribution.

(3) In particular, for Normal Distribution, the inverse transform is not efficient. Instead, **Box-Muller Transformation** is an exact method that uses the inverse transform method to convert two independent uniform random variables

into two independent normally distributed random variables. This is possible by using the generalization of the relation between the PDF of two random vectors $\underline{Y}$ and $\underline{X}$. If the values of $\underline{Y}$ and $\underline{X}$ are related by the transformation $\underline{x} = \underline{g}(\underline{y})$, then the distributions $p_X(\underline{x})$ and $P_Y(\underline{y})$ are given by

$$p_X(\underline{x}) = p_Y(\underline{y}) \left| \frac{d\underline{g}(\underline{y})}{d\underline{x}} \right| \tag{19}$$

where $\frac{d\underline{g}(y)}{d\underline{x}}$ is the Jacobian and $|\cdot|$ denotes determinant. Suppose that $\underline{U} = (U_1, U_2)$ is a set of two standard uniformly distributed variables with values defined in the interval $[0, 1]$. Using (19) for $\underline{y} = \underline{u}$, it can readily shown that the transformation

$$x_1 = u_1 \left(\frac{-2ln(u_1)}{r^2} \right)^{1/2} \tag{20}$$

$$x_2 = u_2 \left(\frac{-2ln(u_2)}{r^2} \right)^{1/2} \tag{21}$$

where $r^2 = u_1^2 + u_2^2$, results in independent and standard Gaussian variables X_1 and X_2 since the joint distribution of X_1 and X_2 is

$$p(x_1, x_2) = p(u_1, u_2) \left| \frac{d\underline{g}(\underline{u})}{d\underline{x}} \right| \tag{22}$$

which, after computing the Jacobian, it can be shown that

$$p(x_1, x_2) = \frac{1}{\sqrt{2\pi}} exp\left[-\frac{x_1^2}{2} \right] \frac{1}{\sqrt{2\pi}} exp\left[-\frac{x_2^2}{2} \right] \tag{23}$$

(4) Generation of samples from multivariate Gaussian distribution

Let $\underline{X}$ be a vector of independent identically distributed (iid) Gaussian random variables, i.e. the k-th component X_k follows

$$X_k \sim N(0, 1) \tag{24}$$

Then a Gaussian random vector $\underline{Y}$ with mean $\underline{\mu}$ and covariance matrix C is obtained from the transformation

$$\underline{Y} = \underline{\mu} + \Phi\sqrt{\Lambda}\underline{X} \tag{25}$$

where $\Lambda = diag(\lambda_i)$, $\Phi = [\phi_1, \ldots, \phi_n]$, λ_i and ϕ_i are the eigenvalues and orthonormal eigenvectors of the covaraince matrix C, respectively, satisfying

$$C\Phi = \Phi\Lambda \tag{26}$$

with $\Phi\Phi^T = \Phi^T\Phi = I$ and $\Phi^T C\Phi = \Lambda$. Them N samples from any multivariate Gaussian variable with mean $\underline{\mu}$ and covariance matrix C can thus be obtained from the transformation

$$\underline{y}^{(i)} = \underline{\mu} + \Phi\sqrt{\Lambda}\underline{x}^{(i)} \tag{27}$$

where $\underline{x}^{(i)}$ are iid samples drawn from the Standard Gaussian Distribution $N(\underline{0}, I)$. That is, these samples are generated independently for each component x_k of $\underline{x}$ from a standard Gaussian distribution: $x_k^{(i)} \sim N(0,1)$, $i = 1, \ldots, N$.

(5) Generation of samples from a uniform distribution

Samples drawn from a uniform distribution X with PDF

$$p_X(x) = \begin{cases} 1/(b-a), & \text{for } x \in [a,b] \\ 0, & \text{otherwise} \end{cases} \tag{28}$$

are obtained from the transformation

$$x = a + (b-a)u \tag{29}$$

where u is a SUD.

(6) For a large number of often used PDFs, there are software programs available to generate iid samples.

1.4 Monte Carlo Integration

Both integrals (2) and (3) are probability integrals of the type

$$I = \int h(\underline{\theta}) p(\underline{\theta}) d\underline{\theta} \tag{30}$$

where $h(\underline{\theta})$ is a general function of $\underline{\theta}$ and $p(\underline{\theta})$ is the PDF of $\underline{\theta}$. Note that

$$I = E[h(\underline{\theta})] \tag{31}$$

is the expected value of $h(\underline{\theta})$. Using the **Law of Large number**, the expected value of a variable is approximated by the sample estimate

$$I = E\left[h(\underline{\theta})\right] \approx I_N = \frac{1}{N} \sum_{i=1}^{N} h(\underline{\theta}^{(i)}) \tag{32}$$

where $\underline{\theta}^{(i)}$ $i = 1, \cdots, N$ are random samples, a sequence of independent and identically distributed random variables, drawn from the PDF $p(\underline{\theta})$. The sample average in (32) converges to the expected value when $\underline{\theta}^{(i)}$ $i = 1, \cdots$ is an infinite sequence of i.i.d. random variables.

The Law of Large Numbers ensures that

$$\lim_{N \to \infty} I_N = I \tag{33}$$

Using the **Central Limit Theorem**, one has that $\sqrt{N}(I_N - I)$ converges in distribution to the normal distribution $N(0, \sigma^2)$, where $\sigma^2 = Var[h(\underline{\theta})]$ is the variance of $h(\underline{\theta})$. That is, the deviation of the sample average I_N from its

limit $E[h(\underline{\theta})] = I$, when multiplied by $\sqrt{N}$, approximates a normal distribution with mean 0 and variance $\sigma^2 = Var[h(\underline{\theta})]$. Equivalently $I_N - I$ converges in distribution to the normal distribution $N(0, \frac{\sigma^2}{N})$ or

$$I_N - I \sim N(0, \frac{\sigma^2}{N}) \tag{34}$$

which means that the error $\sigma/\sqrt{N}$ of the sample estimate is $O\left(\frac{1}{\sqrt{N}}\right)$, of the order of $\frac{1}{\sqrt{N}}$.

In practice the variance

$$\sigma^2 \equiv Var[h(\underline{\theta})] = E\left[\left(h(\underline{\theta}) - E[h(\underline{\theta})]\right)^2\right] = \int \left(h(\underline{\theta}) - E[h(\underline{\theta})]\right)^2 p(\underline{\theta}) d\underline{\theta} \tag{35}$$

is unknown and can be replaced by the sample variance using the unbiased estimate of the variance

$$\sigma_N^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(h(\underline{\theta}^{(i)}) - E[h(\underline{\theta})]\right)^2 \tag{36}$$

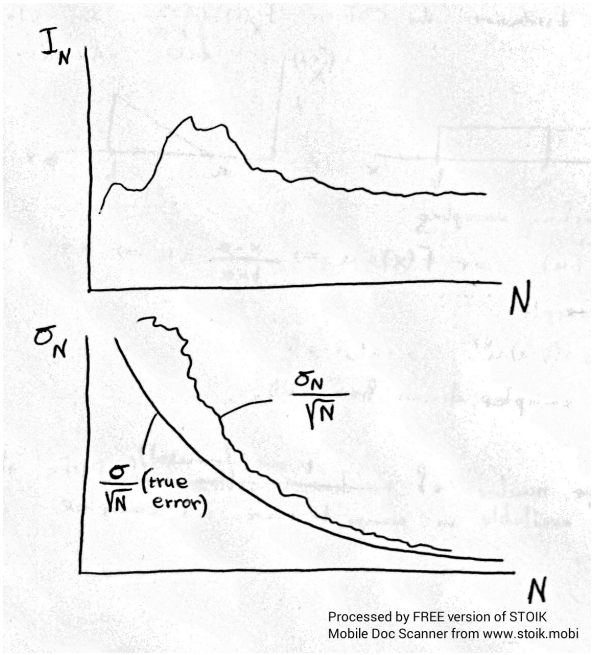
where $E[h(\underline{\theta})]$ is replaced by the sample mean I_N, so that

$$\sigma_N^2 \approx \frac{1}{N-1} \sum_{i=1}^{N} \left(h(\underline{\theta}^{(i)}) - I_N\right)^2 \tag{37}$$

The error of the sample estimate I_N can thus be replaced by

$$\sqrt{Var(I_N)} \approx \frac{\sigma_N}{\sqrt{N}} \tag{38}$$

which is of $O\left(\frac{1}{\sqrt{N}}\right)$. However, the accuracy of σ_N (how close σ_N is to σ) depends also on the accuracy of the sample estimate I_N and σ_N may overestimate σ. In practice, one monitors I_N and σ_N as a function of N and terminates the sampling after the σ_N falls below a specified thresshold.



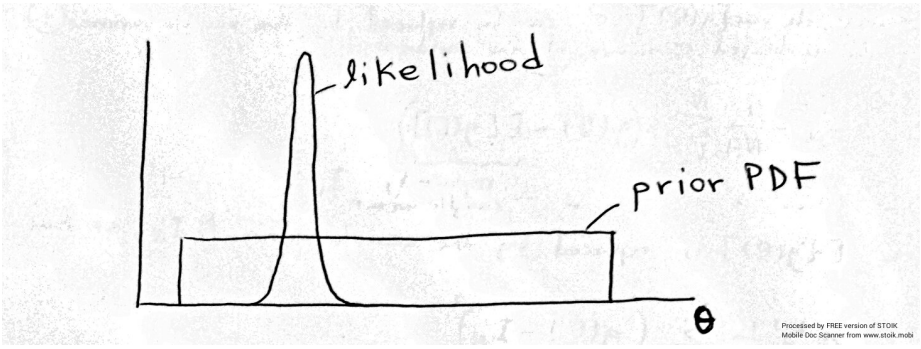
1.5 Estimation of Evidence

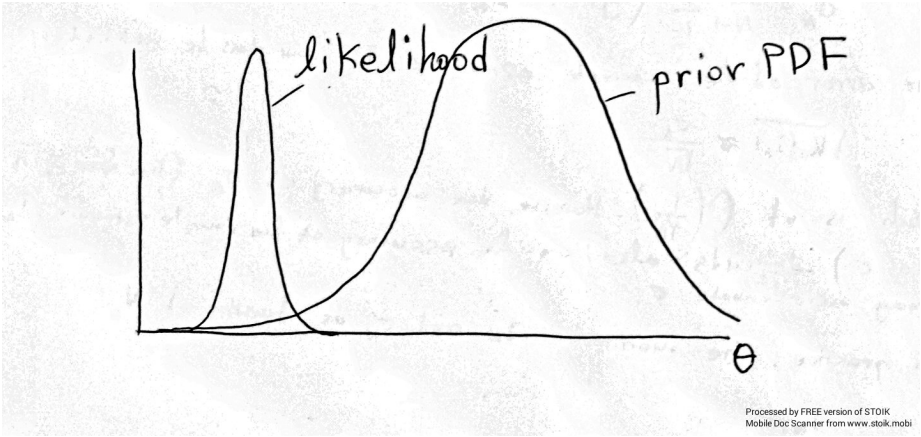
Returning to the probability integrals arising in the Bayesian formulation, let us first compute the evidence $p(D|I)$, given in (2), which is the normalizing factor in the posterior PDF in (1). The estimate can be approximated by the sample estimate, provided that the samples $\underline{\theta}^{(i)}$ are drawn from the prior distribution $p(\underline{\theta}|I)$. The sample estimate of integral (2) is:

$$p_N(D|I) = \frac{1}{N}\sum_{i=1}^{N} p(D|\underline{\theta}^{(i)}, I) \tag{39}$$

Prior to data, the components of $\underline{\theta}$ are usually assumed to be independent and the prior distribution for each component is assumed to follow simple known distributions (e.g. uniform, Gaussian, Gamma) from which sample estimates are readily available in computer programs. However, sampling from the prior distribution will slow down convergence of the estimate significanlty due to the fact that the importance region (domain of significance) of the likelihood $p(D|\underline{\theta}, I)$ in the parameter space is very small compared to the domain over which the support of the prior occupies. Even worse, this importance region might fall at the tails of the prior PDF. As a result, only a small fraction of the samples drawn from the prior distribution may fall within the importance region so that in order to get a sufficiently accurate result, a very large number of samples is required. In practical application, one cannot usually afford large number of samples since each sample usually requires a model run which can

be computationally tedious. The problem can be overcome by trying to sample in the important regions. The problem deteriorates as the dimension of the parameter space increases.





1.6 Estimation of Robust Prediction Integral

Consider now the integral (3)

$$E\left[h(\underline{\theta})\right] = \int_{\Theta} h(\underline{\theta}) p(\underline{\theta}|D, I) d\underline{\theta} \tag{40}$$

The posterior PDF is in most cases a complicated multi-variate distribution defined within an unknown normalizing constant, the evidence $p(D|I)$. So methods for generating i.i.d. samples from the posterior PDF are in most cases not available.

One can write the integral (3) as:

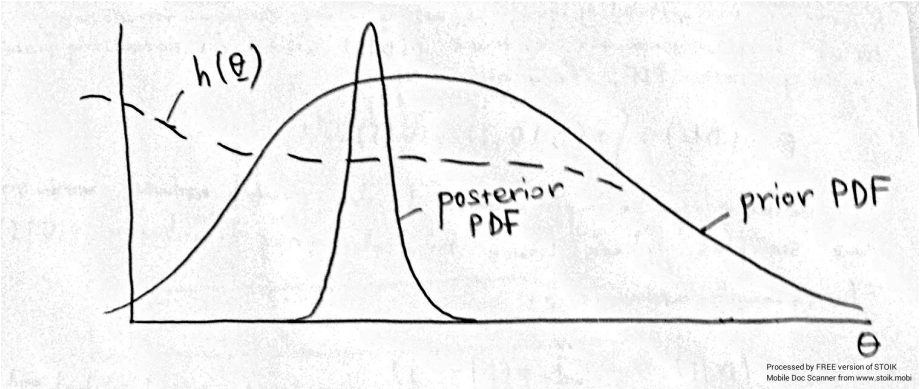
$$E[h(\underline{\theta})] = \frac{1}{p(D|I)} \int h(\underline{\theta}) p(D|\underline{\theta}, I) p(\underline{\theta}|I) d\underline{\theta} \tag{41}$$

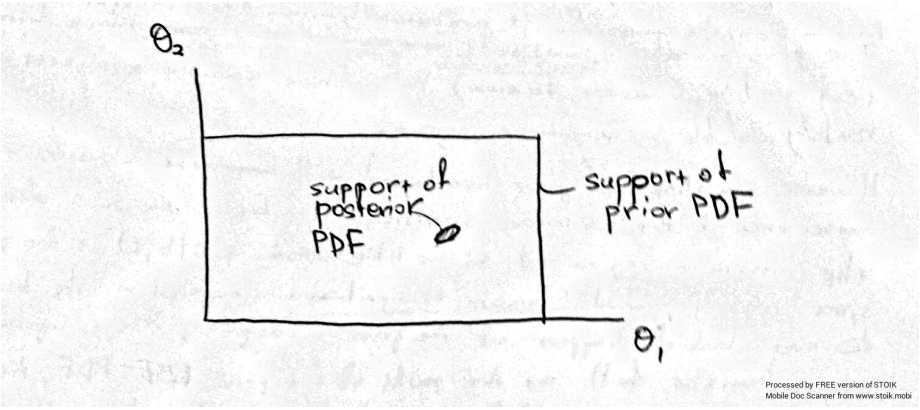
and use i.i.d. samples from the prior PDF to obtain a sample estimate of the integral in the numerator and also for the evidence $p(D|I)$. The sample estimate

$$E[h(\underline{\theta})] \approx \frac{1}{p_N(D|I)} \frac{1}{N} \sum_{i=1}^{N} h(\underline{\theta}^{(i)}) p(D|\underline{\theta}^{(i)}, I) \tag{42}$$

suffers from the same problems, i.e. very slow convergence, already discussed for the evidence.

Variance reduction techniques such as the importance sampling method can be used to generate samples in parts of the region that are most important, instead of covering a much larger region. However, it is not trivial to identify the importance region. The Markov Chain Monte Carlo (MCMC) algorithms are powerful methods for generating samples from an arbitrary PDF that is known up to a scaling constant. The samples generated are dependent. However, these samples are used for statistical averaging as they were independent, accepting a reduced efficiency of the sample estimate.


h(θ)
posterior PDF
prior PDF
θ
Processed by FREE version of STOIK
Mobile Doc Scanner from www.stoik.mobi


θ2
support of posterior PDF
support of prior PDF
θ1
Processed by FREE version of STOIK
Mobile Doc Scanner from www.stoik.mobi