ISSN No. 0976-5697

Volume 3, No. 2, March-April 2012

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Fundamental Concepts of MCMC Methods Proved on Metropolis-Hastings Alghorithm

GH. Gholami
Department of Mathematics
Faculty of Sciences, Urmia University
Urmia, Iran
Gh.gholami@urmia.ac.ir

A. Fayyaz Movaghar
Department of Statistics, Faculty of
Mathematics, University of Mazandaran
Babolsar, Iran
A_fayyaz@umz.ac.ir

A. Etemadi
Department of Statistics, Faculty of
Mathematics, University of Mazandaran
Babolsar, Iran
A.etemadi@umz.ac.ir

H. Rasi
Department of Statistics, Faculty of
Mathematical Sciences, Tabriz University
Tabriz, Iran
Rasi_stat@yahoo.com

E. Fayyazi*
Department of Statistics, Science and Research Branch
Islamic Azad University
Fars, Iran
E.fayyazi@fsriau.ac.ir

S. Panahi
Department of Statistics, Faculty of
Basic Sciences, Payam-e Noor University
Tehran, Iran
S_panahi_stat@yahoo.com

Abstract: Monte Carlo Markov Chain methods have been used extensively in Bayesian methods of inference. Metropolis-Hastings algorithm is one of the most known methods in this regards. In this work we will introduce mathematical bases of this algorithms and show why these algorithms work and their outputs are trustable.

Keywords: Markov Chain; MCMC algorithms; Bayesian computation; Metropolis-Hastings algorithm; Invariant measures

I. INTRODUCTION

The Bayesian paradigm is based on specifying a probability model for the observed data, given a vector of unknown (but non-constant) parameters, leading to the likelihood function, and a probability model for this unknown vector of parameter which is called prior distribution. Inference concerning the model parameter is then based on the posterior distribution which is obtained by Bayes' theorem. In most of the cases, the posterior distributions do not have an analytical closed form [1]. Complex posterior distribution leads to problem (like calculating integrals or optimizing a function) that does not admit on analytical solution [4]. These problems can be solved by finding an adequate approximation for their problem at hand by sampling from the posterior distribution. This dilemma leads to the following question: How do we sample from the multivariate posterior distribution when no closed form is available for this posterior distribution? This question has led to an enormous literature on computational methods for sampling from a given multivariate distribution (posterior distribution) as well as on methods for approximating integrals.

II. BAYESIAN COMPUTATION

Although the Bayesian recipe for inference is conceptually simple, a practical problem with this approach is the difficulty associated with exploring and summarizing realistically complex posterior distributions. In most practical problems, the integral resulting from those inference procedures, do not admit an analytical solution and computational techniques are required to approximate them. Furthermore, in most models and applications, $m(x)$, marginal distribution of data, does not have a closed form and the posterior is represented by an unnormalized density, and thus the Bayes factor will not be in closed form and should be approximated [5]. The introduction to the statistical literature of the techniques known as Markov Chain Monte Carlo (MCMC) methods in the late 1980's overcomes this problem and greatly simplifies the Bayesian analysis of even the most complex models.

A. The Bayesian Approach to Statistical Inference:

Let $D = \{x_1,...,x_n\}$ be an independent and identically distributed sample from a density $f(.|\tilde{\theta})$, with an unknown parameter $\tilde{\theta} \in \tilde{\Theta}$, where $\tilde{\Theta}$ denotes the parameter space of $\tilde{\theta}$. Then the associated likelihood function is

$$L(\tilde{\theta} \mid D) = f_{\tilde{\theta}}(x_1,...,x_n \mid \tilde{\theta}) = \prod_{i=1}^{n} f(x_i \mid \tilde{\theta}).$$

This quantity is a fundamental entity for the analysis of the information provided about $\tilde{\theta}$ by the sample D in both frequentist and Bayesian approaches.

In a Bayesian approach, we assume that $\tilde{\theta}$ is random and has a *prior* distribution denoted by $\pi(\tilde{\theta})$. This approach is based upon Bayes' theorem [Bayes (1763)], which, states that

$$\pi(\tilde{\theta} \mid D) = \frac{L(\tilde{\theta} \mid D)\pi(\tilde{\theta})}{\int_{\tilde{\Theta}} L(\tilde{\theta} \mid D)\pi(\tilde{\theta})d\tilde{\theta}}.$$

The posterior distribution represents the exhaustive summary of our beliefs about the model parameters after having observed the data D. The prior is concerned with our original beliefs about $\tilde{\theta}$ before observing the data. Thus, Bayes' theorem allows us to update our prior beliefs on the basis of the observed data in order to obtain our posterior beliefs, which constitute the basis for our inference.

It is clear that $\pi(\tilde{\theta} \mid D)$ is proportional to the likelihood multiplied by the prior,

$$\pi(\tilde{\theta} \mid D) \propto L(\tilde{\theta} \mid D)\pi(\tilde{\theta}),$$

and thus it involves a contribution form the observed data through $L(\tilde{\theta} \mid D)$, and a contribution from the prior distribution quantified through $\pi(\tilde{\theta})$.

The quantity

$$m(D) = \int_{\tilde{\Theta}} L(\tilde{\theta} \mid D)\pi(\tilde{\theta})d\tilde{\theta}$$

is the *normalizing constant* of $\pi(\tilde{\theta} \mid D)$, and is often called the *marginal* distribution of the data or the prior predictive distribution.

B. The Integration Problem:

We are given a density function $\pi(.)$ on some state space X, which is possibly unnormalized. We want to estimate expectations under π of some functions $g : X \to R$, i.e.

$$E_\pi[g(X)] = \int_X g(x)\pi(x)dx.$$

If X is high-dimensional and if $\pi(.)g(.)$ is a complicated function, the direct (either analytical or numerical) resolution of this integral is infeasible.

The classical Monte Carlo solution to this problem is to simulate *iid* random variables $x_1, x_2, ..., x_N \sim \pi(.)$, and then to estimate $E_\pi[g(X)]$ by

$$\hat{E}_\pi[g(X)] = \frac{1}{N}\sum_{i=1}^{N} g(x_i).$$

This approximation gives an unbiased estimate of $E_\pi[g(X)]$ with standard deviation of order $O(1/\sqrt{N})$. Furthermore, if $E_\pi[g^2(X)] < \infty$, then by the classical Central Limit Theorem, the error $\hat{E}_\pi[g(X)] - E_\pi[g(X)]$ enjoys a limiting normal distribution. The problem, however, is that if π is complex, then it is very difficult to directly simulate *iid* random variables from $\pi(.)$. A possible solution is to apply other methods like Importance Sampling, which allow generating random values from the distribution of interest by simulating from an instrumental distribution.

C. Importance Sampling:

The key idea used in importance sampling is to generate random values through an instrumental distribution $q(x)$, called *importance distribution*, and to correct the resulting simulated function by means of the ratio between the true density and the instrumental density

$$E_\pi[g(X)] = \int g(x)\pi(x)dx = \int g(x)\frac{\pi(x)}{q(x)}q(x)dx$$

$$= E_q[g(X)w(X)] \approx \frac{1}{N}\sum_{i=1}^{N} g(x_i)w(x_i),$$

where

$$w(x_i) = \frac{\pi(x_i)}{q(x_i)}, i = 1,2,..., N$$

are called *importance weights*. The resulting Monte Carlo estimator is unbiased and converges to $E_\pi[g(X)]$ as $N \to \infty$, whatever the choice of the instrumental distribution q and as long as $Supp(q) \supset Supp(\pi)$. Note however that a good choice of the distribution q may reduce the variance of the estimator, which is

$$E_q[g^2(X)\frac{\pi^2(X)}{q^2(X)}] = E_q[g^2(X)\frac{\pi(X)}{q(X)}].$$

Therefore instrumental distributions with tails lighter than those of π are not appropriate for importance sampling. Moreover if the ratio π/q is unbounded then importance weights vary widely giving too much importance to a few values x_i.

In order to improve the efficiency of the Monte Carlo estimator an alternative is to use the following importance sampling estimator

$$E_\pi[g(X)] \approx \frac{\frac{1}{N}\sum_{i=1}^{N} g(x_i)w(x_i)}{\frac{1}{N}\sum_{i=1}^{N} g(w_i)}$$

where

$$w(x_i) = \frac{\pi(x_i)}{q(x_i)},$$

is the importance weight. Note however that the estimator is no more unbiased because the quantity at the denominator is a random variable. See Robert and Casella (2004) for more details.

Instead the Markov chain Monte Carlo (MCMC) solution is to simulate sequences that are neither independent nor identically distributed, but converge in distribution to $\pi(.)$. This can be done by constructing a Markov chain on X which has $\pi(.)$ as a stationary distribution.

D. Markov Chain Monte Carlo Methods:

Markov Chain Monte Carlo (MCMC) methods date from the original work of Metropolis et al [10], who were interested in methods for the efficient simulation of the energy levels of atoms in a crystalline structure. The original idea was

subsequently generalized by Hastings [6], but its true potential was not fully realized within the statistical literature until Gelfand and Smith [3] demonstrated its application to the estimation of integrals commonly occurring in the context of Bayesian statistical inference. MCMC methods work by constructing a Markov Chain [11], which is essentially a series of random variables generated one after the other so that conditional on the past the distribution at any time in the sequence depends only upon the proceeding value. The key to MCMC methods is to construct a generation method for the chain that has the target distribution (posterior distribution) as a stationary distribution. Thus the working principle of MCMC algorithms is as follows:

For an arbitrary starting value x_0, a chain (X_n) is generated using a transition kernel with stationary distribution f, which ensures the convergence in distribution of (X_n) to a random variable from f.

Definition 1: A Markov Chain Monte Carlo (MCMC) method for the simulation of a distribution f is any method producing an Ergodic Markov chain (X_n) whose stationary distribution is f.

The use of a chain (X_n) produced by a Markov Chain Monte Carlo algorithm with stationary distribution f is fundamentally identical to the use of an *iid* sample from f in the sense that the ergodic theorem guarantees the (almost sure) convergence of the empirical average

$$\eta_N = \frac{1}{N}\sum_{i=1}^{N} h(X_i)$$

to the quantity $E_f[h(X)]$. A sequence (X_n) produced by a Markov Chain Monte Carlo algorithm can thus be employed just as an *iid* sample [2].

E. The Gibbs Sampler:

The Gibbs sampler may be one of the best known MCMC sampling algorithms in the Bayesian computational literature. The Gibbs sampler is found on the ideas of Grenander [9], while the formal term is introduced by Geman and Geman [12]. The primary bibliographical landmark for Gibbs sampling in problems of Bayesian inference is Gelfand and Smith [18]. A similar idea termed as data augmentation is introduced by Tanner and Wong [22].

F. The Slice Sampler:

The slice sampler is a special type of Markov chain Monte Carlo (MCMC) auxiliary variable method [20] , Edwards and Sokal [8] , Besag and Green [7], that has been popularized by Neal [16], Fishman [8], and Damien et al [14].

G. The Reversible Jump Technique:

Some statistical inference procedures consist of comparing competitive models to fit the data. The parameter spaces of these models usually have different dimensions. The simulation methods introduced before cannot take into account this extra dimension variability and we need more advanced techniques that are capable moving between parameter spaces of different dimensions to explore the whole parameter space. The standard Metropolis-Hastings algorithm described earlier is indeed incapable of such movements, whereas the Reversible Jump algorithm of Green [19] is an extension of the standard Metropolis-Hastings algorithm to allow exactly for this possibility.

III. BASIC IDEAS

A Markov Chain is composed of a sequence which is possibly includes a set of random variables which are said to be growing over time [17]. Furthermore, a Markov chain has a probability of a transition which is dependent on the particular set in which the chain is located. Based on the afore mentioned comments on Markov Chain, the most straightforward way and mathematically the most lucid method of defining chain is in terms of its transition kernel. Moreover, transition kernel is the function which determines these transitions [13].

A. Fundemental Concepts:

Definition 2: A transition kernel is naturally a function K which is defined on $\chi \times B(\chi)$ in a way that

(i) $\forall x \in \chi, K(x,.)$ is a probability measure.

(ii) $\forall A \in B(x), K(.,A)$ is measurable.

Definition 3: If we suppose K a transition kernel, a sequence $X_1, X_2, ..., X_n, ...$ of random variables can be regarded as a Markov Chain, portrayed by (X_n), if for any number of t, the conditional distribution of X_t given $x_{t-1}, x_{t-2}, ..., x_0$ is like the distribution of X_t given x_{t-1}; that is,

$$P(X_{k+1} \in A \mid x_0, x_1, x_2, ..., x_k) = P(X_{k+1} \in A \mid x_k) = \int_A K(x_k, dx). \quad (1)$$

Example 1 : $AR(1)$ models provide a simple illustration of Markov chains on continuous state-space. If

$$X_n = \theta X_{n-1} + \varepsilon_n, \theta \in R$$

with $\varepsilon_n \sim N(0, \sigma^2)$, and if the ε_n's are independent, X_n is indeed independent from $X_{n-2}, X_{n-3}, ...$ conditionally on X_{n-1}. The characteristics of Markov of an $AR(q)$ process can be derived by considering the vector $(X_n, ..., X_{n-q+1})$. In addition, $ARMA(p,q)$ models do not have the characteristics of Markov [13].

Definition 4: Assume that $A \in B(\chi)$. The initial n for which the chain goes into the set A is shown by:

$$\tau_A = \inf\{n \geq 1; X_n \in A\} \quad (2)$$

and is called the *stopping time* at A with, by agreement $\tau_A = \infty$ if $X_n \notin A$.

Definition 5: If we refer to $K^1(x,A) = K(x,A)$, the kernel for n transitions is defined by $(n > 1)$

$$K^n(x,A) = \int_\chi K^{n-1}(y,A) K(x,dy) \quad (3)$$

B. Irreducibility, Atoms and small sets:

The first measure of sensitivity of Markov chain to its first condition is the property of irreducibility [15]. It is vital in the setup of Markov Chain Monte Carlo algorithms owing to the fact that it guarantees the convergence of the generated chain through MCMC algorithms.

Definition 6: Assuming a measure ψ, the Markov Chain X_n with transition kernel $K(x,y)$ is ψ-irreducible, if for any $A \in B(\chi)$ with $\psi(A) > 0$, there is n in a way that $K^n(x,A) > 0$ for all $x \in \chi$.

Definition 7: The Markov Chain X_n has an atom $\alpha \in B(\chi)$ if there is an connected nonzero measure ν in a way that

$$K(x;A) = \nu(A) \qquad \forall x \in \alpha, \forall A \in B(\chi) \qquad (4)$$

If (X_n) is ψ-irreducible, the atom is accessible when $\psi(\alpha) > 0$.

a. Minorizing condition:

There is a set $C \in B(\chi)$, a constant $\varepsilon > 0$, and a probability measure ν in a way that

$$K(x,A) \geq \varepsilon \nu(A) \quad \forall x \in C, \forall A \in B(\chi).$$

The minimizing condition is said to be a significant technique of proving the renewal theory.

Definition 8: A set C is small if there is $m \in N^*$ and a nonzero measure ν_m in a way that

$$K^m(x,A) \geq \nu_m(A) \quad \forall x \in C, \forall A \in B(\chi) \qquad (5)$$

Definition 9: A renewal time (or regeneration time) is a stopping rule τ with the characteristics which $(X_r, X_{r+1}, ...)$ does not depend on $(X_{r-1}, X_{r-2}, ...)$.

Occasionally the behavior of (X_n) might be confined by determistic constraints on the moves from X_n to X_{n+1}. For the time being, we organize these constraints and suggest that the chains produced by Markov Chain Monte Carlo algorithms do not display this behavior [21].

Definition 10: A ψ-irreducible chain (X_n) has a cycle of length d if there exists a small set C, an related integer M, and a probability distribution ν_m in a way that d is the *g.c.d* of

$$\{m \geq 1; \exists \delta_m > 0 \text{ such that } C \text{ is small for } \nu_m \geq \delta_m \nu_M\}. \qquad (6)$$

The period of (X_n) can be therefore described as the largest integer d which proves to be true about (6) and if $d = 1$ then (X_n) is a periodic.

C. Transience and Recurrence:

Looking from an algorithmic perspective, a Markov Chain is compulsory to possess stability properties to guarantee an acceptable approximation of the stimulated model. Moreover, irreducibility confirms that every set A will be accessed by the Markov Chain (X_n), though, this property is too weak to confirm that the behavior of (X_n) will enter A often enough. Formalizing this stability of the Markov Chain leads to different notions of recurrence.

Definition 11: For any $x \in A$, A set A is considered recurrent if $E_x[\eta_A] = \infty$ [13].

Definition 12: if there is a constant M such that $E_x[\eta_A] < M$ for every $x \in A$ then set A is uniformly transient [13].

Definition 13: If there exists a covering of χ by uniformly transient sets then set A is transient which is a countable collection of uniformly transient sets B_i in a way that

$$A = \cup_i B_i.$$

Definition 14: A Markov Chain (X_n) is recurrent if:

i. There exists a measure ψ such that (X_n) is ψ-irreducible, and
ii. for every $A \in B(\chi)$ such that $\psi(A) > 0$, $E_x[\eta_A] = \infty$ for every $x \in A$. The chain is transient if it is ψ-irreducible and if χ is transient [13].

Theorem 1: Assume that (X_n) is ψ-irreducible Markov chain with an accessible atom α.

i. If α is recurrent, every set A of $B(\chi)$ in a way that $\psi(A) > 0$ is recurrent
ii. If α is transient, χ is transient.

The quality of (*i*) is the most appropriate in the Markov Chain Monte Carlo setup [13].

A ψ-irreducible chain (X_n) is recurrent if there exists a small set C with $\psi(C) > 0$ such that $P_x(\tau_c < \infty) = 1$ for every $x \in C$ [13].

a. Harris Recurrence:

It is actually possible to build up the stability properties of a chain (X_n) by requiring not only an infinite average number of visits to every small set but also an infinite number of visits for every path of the Markov Chain [13].

Definition 15: A set A is Harris recurrent if $P_x(\eta_A = \infty) = 1$ for all $x \in A$.

The chain (X_n) is Harris recurrent if there exists a measure ψ such that (X_n) is ψ-irreducible and for every set A with $\psi(A) > 0$, A is Harris recurrent [13].

Theorem 2: If (X_n) is a ψ-irreducible Markov Chain with a small set C such that $P_x(\tau_c < \infty) = 1$ for all $x \in \chi$, then (X_n) is a Harris recurrent [13].

D. Invariant Measures :

An elevated level of stability for the chain (X_n) is obtained if the marginal distribution of X_n depends on n. More formally, this is an obligatory occasion for the existence of a probability distribution π in a way that $X_{n+1} \sim \pi$ if $X_n \sim \pi$, and Markov Chain Monte Carlo methods are based on the fact that this requirement, which defines a particular kind of recurrence called positive recurrence, can be met. The Markov

Chains constructed from Markov Chain Monte Carlo algorithms enjoy this greater stability property.

Definition 16: A σ -finite measure π is invariant for the transition kernel $K(.,.)$ (and for the related chain) if

$$\pi(B) = \int_{\chi} K(x,B)\pi(dx) \quad \forall B \in B(\chi) \tag{7}$$

When there is an invariant probability measure for a ψ -irreducible chain, the chain is positive [13].

a. Stationary Distribution:

The invariant distribution is also mentioned to as stationary if π is a probability measure, since $X_0 : \pi$ denotes that $X_n : \pi$ for every n; thus, the chain is stationary in distribution [8].

The relation between positivity and recurrence is specified by the following results: which formalizes the intuition that the being of an invariant measure avoids the probability mass from "fleeing to infinity".

b. Proposition:

If the chain (X_n) is positive, it is recurrent.

Theorem 3: Assume (X_n) be ψ -irreducible with an atom α . The chain is positive if and only if $E_{\alpha}[\tau_{\alpha}] < \infty$. In this case, the interval distribution π for (X_n) satisfies [13].

The chains created by Markov Chain Monte Carlo methods are, by core, certain to possess an invariant distribution.

The Markov chains created from Markov Chain Monte Carlo algorithms has positive recurrence things.

E. Reversibility and Detailed Balance Condition:

The stability stuff basic to stationary chains can be linked to another stability stuff called Reversibility, which statuses that the instructions of time dose no modification in the dynamics of the chain [12].

Definition 17: A stationary Markov Chain (X_n) is reversible if the distribution of X_{n+1} conditionally on $X_{n+2} = x$ is the same as the distribution of X_{n+1} conditioning on $X_n = x$ [13].

a. Detailed Balance Condition:

A Markov Chain with transition kernel K placates the detailed balance condition if there is a function f sufficient

$$K(y,x)\,f(y) = K(x,y)\,f(x) \tag{8}$$

Though this state is not necessary for f to be a stationary measure related with the transition kernel K , it offers a sufficient state that is frequently easy to check and that can be used for greatest MCMC algorithms.

The detailed balance state express on balance in the flow of the Markov chain, specifically that the probability of existence in x and moving to y is the similar as the probability in y and moving back to x .

F. Ergodicity and Convergence:

Since the Markov chain (X_n) from a sequential view, it is natural and important to create the limiting activities of X_n ; that is to what is the chain converging? the being and uniqueness of an invariant distribution π creates that distribution a usual candidate for the limiting distribution, and we now try to finding sufficient settings on (X_n) for X_n to be asymptotically distributed allowing to π . The next theorems and definitions are essential convergence consequences for Markov chains and they are at the essential of the incentive for Markov chain Monte Carlo algorithms [9].

Definition 18: For a Harris positive chain (X_n) , with invariant distribution f , an atom α is ergodic if [13]

$$\lim_{n\to\infty} |K^n(\alpha,\alpha) - f(\alpha)| = 0 \tag{9}$$

In the countable situation, the being of an ergodic atom is, infact, sufficient to create convergence giving to the total variation norm

$$\| \mu_1 - \mu_2 \|_{TV} = \sup_A |\mu_1(A) - \mu_2(A)| \tag{10}$$

a. Geometric Convergence:

An additional exact report of convergence properties includes the study of the speed of convergence of k^n to f . An assessment of this is key for Markov Chain Monte Carlo algorithms in the sense that it relays to stopping rules for these algorithms; minimal convergence speed is also a requirement for the application of the Central Limit Theorem [7].

Definition 19: A chain (X_n) is geometrically h-ergodic, with $h \geq 1$ on χ , if (X_n) is Harris positive, with stationary distribution f , if (X_n) satisfies $\mathsf{E}^{\pi}[h] < \infty$, and if there exists $r_h > 1$ such that

$$\sum_{n=1}^{\infty} r_h^n \| K^n(x,.) - f \|_h < \infty$$

For every $x \in X$. The case $h = 1$ corresponds to the geometric ergodicity of (X_n) [13].

Definition 20: An accessible atom α is geometrically ergodic if there exists $r > 1$ such that

$$\sum_{n=1}^{\infty} |K^n(\alpha,\alpha) - f(\alpha)| r^n < \infty$$

and α is a Kendall atom if there exists $K > 1$ such that $\mathsf{E}_{\alpha}[k^{\tau_{\alpha}}] < \infty$.

If α is a Kendal atom, it is thus geometrically ergodic and ensures geometric ergodicity for (X_n) [13].

b. Uniform Ergodicity:

The stuff of uniform ergodicity is robust than geometric ergodicity in the sense that the rate of geometric convergence must be unchanging over the entire space. It is used in the Central Limit Theorem.

Definition 21: The chain (X_n) is uniformly ergodic if [13]

$$\lim_{n\to\infty} \sup_{x\in\chi} \| K^n(x,.) - f \|_{TV} = 0. \tag{11}$$

G. Limit Theorems:

Assumed observation $X_1,...,X_n$ of a Markov chain, we at this time survey the limiting behavior of the incomplete sums

$$S_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

When n goes to limitlessness, getting back to the *iid* case through renewal when (X_n) has an atom. Think over first the idea of harmonic functions, which is linked to ergodicity for Harris recurrent Markov chains.

Definition 22: A measurable function h is harmonic for the chain (X_n) if

$$\mathsf{E}[h(X_{n+1}) \mid x_n] = h(x_n).$$

These functions are invariant for the transition kernel and they characterize Harris recurrence as follows [13].

Theorem 4: For a positive Markov chain, if the only restricted harmonic functions are the constant functions, the chain is Harris recurrent.

Proof. Frist, the probability of an infinite number of returns, $Q(x,A) = P_x(\eta_A = \infty)$, as a function of x, $h(x)$, is clearly a harmonic function. This is because

$$E_y[h(X_1)] = E_y[P_{x1}(\eta_A = \infty)] = P_y(\eta_A = \infty)$$

And thus, $Q(x,A)$ is constant (in x).

The function $Q(x,A)$ describes a tail event, an occasion whose occurrence does not be contingent on $X_1, X_2, ..., X_m$, for any finite m. Such occasions usually follow a 0-1 laws are classically recognized in the independence case, and, inappropriately, extension lead to cover Markov chains are outside our choice. For the sake of our proof, we will just state that $Q(x,A)$ conforms a 0-1 law and proceed.

If π is the invariant measure and $\pi(A) > 0$, the case $Q(x,A) = 0$ is impossible. To see this, assume that $Q(x,A) = 0$. It then surveys that the chain almost surely appointments A only a finite number of times and the average

$$\frac{1}{N}\sum_{i=1}^{N} I_A(X_i)$$

Will not converge to $\pi(A)$, opposing the low of Large Numbers. So, for some x, $Q(x,A) = 1$, creating that the chain is a Harris chain [13].

a. Ergodic Theorem:

If (X_n) has a σ-finite invariant measure π, the following two statements are equivalent:

a) If $f, g \in L^1(\pi)$ with $\int g(x)d\pi(x) \neq 0$ then

$$\lim_{n\to\infty} \frac{S_n(f)}{S_n(g)} = \frac{\int f(x)d\pi(x)}{\int g(x)d\pi(x)}.$$

b) The Markov chain (X_n) is Harris recurrent.

The law of large numbers for Markov chains which is customarily called the Ergodic theorem guarantees the convergence of $S_n(h)$ [13].

By way of key part of ergodic theorem is that π does not want to be a probability measure and, that there can be similar kind of strong stability even if the chain is null recurrent. In the format of a Markov Chain Monte Carlo algorithm this effect is occasionally entreated to defend the use of improper posterior measures.

b. Central Limit Theorems:

There is a regular development from the low of large numbers to the Central Limit Theorem. We offer alternative settings for the Central Limit Theorem to apply in different settings.

(The Discrete Case)

Theorem 5: If X_n is Harris positive with an atom α such that

$$\mathsf{E}_\alpha[\tau_\alpha^2] < \infty \quad , \quad \mathsf{E}_\alpha[(\sum_{n=1}^{\tau_\alpha} | h(X_n) |)^2] < \infty$$

and

$$\gamma_h^2 = \pi(\alpha)\mathsf{E}_\alpha[(\sum_{n=1}^{\tau_\alpha}\{h(X_n) - \mathsf{E}^\pi[h]\})^2] > 0$$

the Central Limit Theorem applies; that is

$$\frac{1}{\sqrt{N}}(\sum_{n=1}^{N}(h(X_n) - \mathsf{E}^\pi[h])) \xrightarrow{L} \mathsf{N}(0, \gamma_h^2) .$$

Proof. If $\bar{h}$ denotes $h - E^\pi[h]$, we get

$$\frac{1}{\sqrt{l_N}}\sum_{i=1}^{l_N} S_i(\bar{h}) \xrightarrow{L} N(0, E_\alpha[\sum_{n=1}^{\tau_\alpha}\bar{h}(X_n)]^2),$$

Following from Central Limit Theorem for the independent variables $S_i(\bar{f})$, while N / l_N converges almost surely to $E_\alpha[S_0(1)] = 1/\pi(\alpha)$. Since

$$|\sum_{i=1}^{l_{N-1}} S_i(\bar{h}) - \sum_{k=1}^{N}\bar{h}(X_k)| \leq S_{l_N}(|\bar{h}|)$$

and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n} S_i(|\bar{h}|)^2 = E_\alpha[S_0(|\bar{h}|)^2],$$

we get

$$\limsup_{N\to\infty}\frac{S_{l_N}(|\bar{h}|)}{\sqrt{N}} = 0,$$

and the reminder goes to 0 almost surely [13].

This result indicates that an extension of the Central Limit Theorem to the nonatomic case will be more delicate than for the Ergodic Theorem [13].

(Reversibility)

Theorem 6: If (X_n) is aperiodic, irreducible, and reversible with invariant distribution f the Central Limit Theorem applies when

$$0 < \gamma_g^2 = \mathsf{E}_f[\bar{g}^2(X_0)] + 2\sum_{k=1}^{\infty}\mathsf{E}_f[\bar{g}(X_0)\bar{g}(X_k)] < \infty$$

The key fact here is that even however reversibility is a very limiting statement in overall, it is frequently easy to levy in Markov Chain Monte Carlo algorithms by introducing additional simulation steps [13].

(Geometric Ergodicity)

There is so far additional method to the Central Limit Theorem for Markov chains, it depend on on geometric ergodicity.

Theorem 7: If (X_n) is aperiodic, irreducible, positive Harris recurrent with invariant distribution f and geometrically ergodic, and if, in addition,

$$\mathsf{E}^f[|h(X)|^{2+\varepsilon}] < \infty$$

for some $\varepsilon > 0$, then

$$\sqrt{n}(S_n(h)/n - \mathsf{E}^f[h(X)]) \xrightarrow{L} \mathsf{N}(0, \gamma_h^2)$$

where Y_h^2 is defined as in theorem 7 ,[13].

IV. THE METROPOLIS-HASTINGS ALGORITHM

In this unit was talking the topic of theoretic validity of the Metropolis-Hastings algorithms which are a kind of MCMC methods for simulation. The MCMC sampling plan sets up an irreducible, aperiodic Markov chain for which the stationary distribution equals the posterior distribution of interest [3].

Definition 23: The Metropolis-Hastings algorithm begins with the objective (target) density f. This is the distribution from which we are inclined to produce. Really, we request to create from target distribution f indirectly.

To do so, we would put on supposed instrumental distribution, presented by $q(y \mid x)$. To create a choice with respect to instrumental distribution, some settings are to be in use into explanation;

i. Generation should be simply complete.
ii. It must be either accessible or symmetric $(q(x \mid y) = q(y \mid x))$.

The act of a Metropolis-Hastings algorithm rest on on the choice of a proposal density q(.). As discoursed in Chib and Greenberg [1], the spread of the proposal density q(.) affects the behavior of the chain in at least two dimensions : one is the "acceptance rate" (the percentage of times a move to a new point is made), and the other is the region of the sample space that is covered by the chain. If the spread is extremely large, some of the generated candidates will have a low probability of being accepted. On the other hand, if the spread is chosen to be too small, the chain will take longer to explore the support of the density. Both these situations are likely to be reflected in a high autocorrelation across sample values. In the context of the random walk proposal density, Roberts et al [15] show that if the target proposal densities are Normal, then the scale of the proposal should be tuned so that the acceptance rate is approximately 0.45 in one-dimensional problems and approximately 0.23 as the dimension of the problem approaches infinity, with the optimal acceptance rate being around 0.25 in six dimensions. For the independence chain, in which we take $q(x \mid y) = q(y)$, it is important to ensure that the tails of the proposal density $q(y)$ dominate those of the target density $q(y \mid x)$ as in importance sampling [3].

A. Metropolis-Hastings Algorithm:

Step 1 : Initialization: choose an arbitrary starting value $x^{(0)}$; Iteration $t, (t \geq 1)$.

Step 2 : Generate $Y_t : q(y \mid x^{(t)})$.

Step 3 : Compute

$$R = \frac{f(y)}{f(x)} \frac{q(x \mid y)}{q(y \mid x)}.$$

Step 4 : Compute the acceptance probability $\mathsf{P}(x, y) = \min\{R, 1\}$.

Step 5 : With probability P ; $X^{(t+1)} = Y_t$; otherwise $X^{(t+1)} = x^{(t)}$ [13].

Theorem 8: Assume that $(X^{(t)})$ be present the Metropolis-Hastings chain. For each conditional distribution q whose contains ξ, the support of f ,

a. The kernel of the chain satisfies the detailed balance condition with f ;
b. f is a stationary distribution of the chain.

Proof. The transition kernel associated with Metropolis-Hastings algorithm is

$$K(x, A) = P(X_{t+1} \in A \mid X_t = x)$$
$$= P(Y \in A, X_{t+1} = Y \mid X_t = x) + P(x \in A, X_{t+1} = x \mid X_t = x)$$
$$= \int_A q(y \mid x)\, p(x, y)\, dy + \int_Y \mathsf{I}_{(x \in A)}(1 - p(x, y))\, q(y \mid x)\, dy$$

$A = \{y\}, q(y \mid x)$ is instrumental density

, $p(x, y) = p(X_{t+1} = y \mid X_t = x)$

$$\Rightarrow K(x, y) = p(x, y)q(y \mid x) + (1 - \int p(x, y)q(y \mid x)dy)\delta_x(y)$$

$$\delta_x(y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}, \quad \delta_y(x) = \begin{cases} 1 & y = x \\ 0 & y \neq x \end{cases}$$

$$\sigma_x(y) = \sigma_y(x) \Rightarrow \sigma_y(x) f(y) = \sigma_x(y) f(x)$$

$$r(x) = \int p(x, y)q(y \mid x)dy$$

$$\Rightarrow K(x, y) = p(x, y)q(y \mid x) + (1 - r(x))\delta_x(y)$$

Show that

$$p(x, y)q(y \mid x) f(x) = p(y, x)q(x \mid y) f(y)$$

And

$$(1 - r(x))\delta_x(y) f(x) = (1 - r(y))\delta_y(x) f(y) \tag{12}$$

Which together establish detailed balance for the Metropolis-Hastings chain:

$$\Rightarrow p(x, y) = 1 \Rightarrow p(y, x) = \frac{f(x)}{f(y)} \frac{q(y \mid x)}{q(x \mid y)} \tag{13}$$

Replacement in (13)

$$1 \times q(y \mid x) f(x) = \frac{f(x)}{f(y)} \frac{q(y \mid x)}{q(x \mid y)}.q(x \mid y) f(y)$$
$$\Rightarrow q(y \mid x) f(x) = f(x)q(y \mid x)$$

$$\Rightarrow p(x, y) = \frac{f(y)}{f(x)} \frac{q(x \mid y)}{q(y \mid x)} \Rightarrow p(y, x) = 1 \tag{14}$$

Replacement in (13)

$$\frac{f(y)}{f(x)} \frac{q(x \mid y)}{q(y \mid x)}.q(y \mid x) f(x) = 1 \times q(x \mid y) f(y)$$
$$\Rightarrow f(y)q(x \mid y) = q(x \mid y) f(y)$$

Giving to (12)

$$\delta_x(y) f(x) - r(x)\delta_x(y) f(x) = \delta_y(x) f(y) - r(y)\delta_y(x) f(y)$$

Giving to definition $\delta_x(y)$ and $\delta_y(x)$, the proof of above equation is evident [13].

Note: We have now shown that Metropolis chain is reversible and f distribution (purpose distribution) is the stationary distribution of metropolis chain.

B. Measuring the Irreducibility State of Metropolis-Hastings Chain:

One can cover the irreducibility quality of Metropolis-Hastings chain if the instrumental distribution of q is positive, that is to say

$$\forall (x, y) \in \xi \times \xi \quad q(y \mid x) > 0$$

Then the chain is irreducible, since

$$\forall x, y \in \chi \Rightarrow p_x(\tau_y < \infty) > 0 \Rightarrow \int q(y \mid x)dy > 0 \Rightarrow q(y \mid x)dy > 0$$

(definition of irreducibility)

The above your head condition conditions that each set of ξ (f range) with positive Lebesgue measure, is achievable in one level [13].

C. Measuring the Positive and Recurrent State of Metropolis-Hastings Chain:

In invariant probability measure for a chain we have if there is an invariant probability measure for chain, then the chain will be positive. Here, since f is an invariant measure for the chain, we can complete that metropolis chain is positive and then, Metropolis-Hastings chain is recurrent [13].

D. Measuring the Aperiodicity Quality of Metropolis-Hastings Chain:

One enough state for Metropolis-Hastings chain to be aperiodic is that Metropolis-Hastings provides possible events such $\{X^{(t+1)} = X^{(t)}\}$. In other words, probability of these events be not equal to zero. So,

$$p[f(x^{(t)})q(Y_t \mid X^{(t)}) \le f(Y_t)q(X^{(t)} \mid Y_t)] < 1$$

$$\Rightarrow p[\frac{f(Y_t)}{f(X^{(t)})}\frac{q(X^{(t)} \mid Y_t)}{q(Y_t \mid X^{(t)})} \ge 1] < 1$$

In other words, the propability of accepting formed sample be less than 1. Obviously, all created samples of instrumental distribution are not accepted, i.e. the chain remains repeated in some cases.

Lemma 1: If the Metropolis-Hastings chain $(X^{(t)})$ is f -irreducible, it is Harris recurrent [13].

Theorem 9: Assume that the Metropolis-Hastings Markov Chain $(X^{(t)})$ is f -irreducible

- If $h \in L^1(f)$, then

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}h(X^{(t)}) = \int h(x)f(x)dx \quad \text{a.e.} \quad f$$

- If, in addition, $(X^{(t)})$ is aperiodic, the

$$\lim_{n\to\infty} \| \int K^n(x,.)\mu(dx) - f \|_{TV} = 0$$

For every initial distribution on μ, where $K^n(x,0)$ denotes the kernel for n transitions.

Proof. 1 If $(X^{(t)})$ is f -irreducible, it is Harris recurrent by Lemma 1 and part (i) then follows from ergodic theorem [13].

In point, above theorem is a convergence effect for Metropolis-Hastings Markov Chain.

Example 2 : Assume $AR(1)$ models. If

$$Y_{n+1} = \theta Y_n + \varepsilon_{n+1}, \theta \in \mathbb{R}$$

And ε_n are independent normal variables, the chain is irreducible. The situation measure existence the Lebesgue measure. We can determine settings for irreducibility.

In practice, the condition most likely to be adopted is that the innovation process ε has a distribution Γ with an everywhere positive density [13].

Example 3 : Assume $AR(1)$ models. If

$$Y_{n+1} = \theta Y_n + \varepsilon_n, \theta \in \mathbb{R}$$

And ε_n are independent uniform $[-1,1]$ variables. It is yet not continuously sufficient for irreducibility to have density only positive in a region of zero. If $|\theta| \le 1$ the chain will be irreducible under such a density condition; but if $|\theta| > 1$, then once we have an initial state larger than $(|\theta| - 1)^{-1}$, the chain will monotonically go off near infinity and will not be irreducible [13].

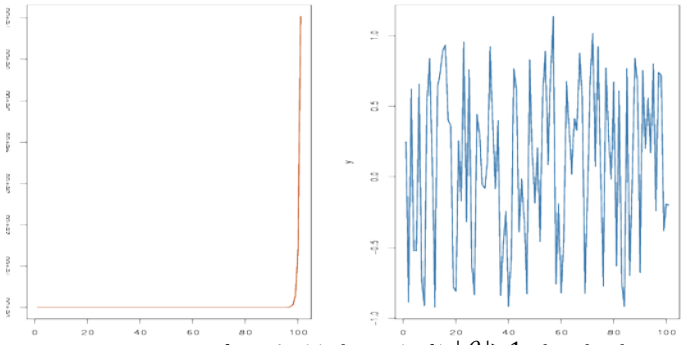

Figure 1. Trajectories of two AR(1) chains. (Left) $|\theta| > 1$, then the chain is not irreducible, (right) $|\theta| \le 1$, and the chain is irreducible

Example 4 : To show the result of proposal distribution, we use two distributions $Ga(4,7)$ and $Ga(5,6)$ to estimate the mean of a gamma distribution with parameters $\alpha = 4.3$ and $\beta = 6.2$.

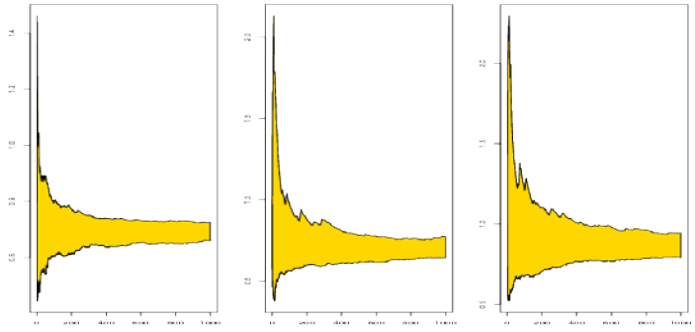

Figure 2. Range of 3 sampling for estimating the mean of $Ga(4.3,6.2)$.

Figure 2 shows tree sampling path from gamma distribution. The first graph is genereted by using an *iid* sampler. The two other graphs are genetrated by output of a Metropolis-Hastings algorithm with proposal distributions $Ga(4,7)$ and $Ga(5,6)$. Since output of the third sampling is mostly similar to that of *iid* sampling output, Metropolis-Hastings algorithm is partly biased based on $Ga(4,7)$ proposal distribution which can indicates inconvergency of stationary distribution [2].

V. REFERENCES

[1] A. Gelfand and A. F. M, Smith. "Sampling Based Approaches to Calculating Marginal Densities". Journal of The American Statistical Association, 85 :398–409, 1990.

[2] G. Fishman. "An analysis of Swendsen-Wang and Related Sampling Methods". Technical report, Department of Operations Research, University of North Carolina, 1996.

[3] GH. Gholami, "Change-point Problems in Regression: A Bayesian Approach". Ph. D. Thesis, 2008.

[4] G. O, Roberts, A. Gelman and W. R, Gilks. "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms". Annals of Applied Probabilities, 7 :110–120, 1997.

[5] J. Besag, and P. J. Green. "Spatial Statistics and Bayesian Computation (with discussion)". The Journal of the Royal Statistical Society, Series B, 55 :25–38, 1993.

[6] J. Geweke. "Bayesian inference in econometric models using Monte Carlo integration". Econometrica, 57 :1317–1340, 1989.

[7] M. James, M. Haran and G. Jomes, "Markov Chain Monte Carlo: Can We Trust The Third Significant Figure?". Statistical Science, 23, 250-260, 2008.

[8] M. Tanner and W. Wong, "The Calculation of Posterior Distributions by Data Augmentation". The Journal of the American Statistical Association, 82 :528–550, 1987.

[9] N. Metropolis, A. W, Rosenbluth, M. N, Rosenbluth, A. H, Teller and E. Teller. "Equations of State Calculations by Fast Computing Machines". Journal of Chemical Physics, 21 :1087–1092, 1953.

[10] P. Damien, J. Wakefield and S. Walker. "Gibbs Sampling for Bayesian Non Conjugate and Hierarchical Models by Using Auxiliary Variables". Journal of the Royal Statistical Society. Series B, 61(2) : 331–344, 1999.

[11] P. D. Hoff. "A First Course in Bayesian Statistical Methods". Springer, 2009.

[12] P. J, Green. "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination". Biometrika, 82(4) : 711–732, 1995.

[13] P. R, Christian and G. Casella, "Monte Carlo Statistical Methods". The Annals of Mathematical Statistics, 49: 327–335, 2004.

[14] P. R, Christian and H. M, Marin, "Bayesian Core: A Practical Approach to Computational Bayesian Statistics". Springer Texts in Statistics, 2006.

[15] R. G, Edwards and A. D, Sokal. "Generalization of the Fortuin-KasteleynSwendsen-Wang Representation and Monte Carlo algorithm". Physical Review D, 38(6): 2009–201, 1989.

[16] R. H, Swendson and J. S, Wang. "Nonuniversal Critical Dynamics in Monte Carlo Simulations". Physical Review Letters, 58 :86–88, 1978.

[17] R. M, Neal, (1997). "Markov Chain Monte Carlo Methods Based on "Slicing" the Density Function". Technical report, Univ. of Toronto, 1997.

[18] S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images". IEEE Trans. Pattern Anal. Mach. Intell., 6 :721–741, 1984.

[19] S. P, Brooks, "Markov Chain Monte Carlo Method and Its Application". The Statistician, 47, part 1, pp. 69-100, 1998.

[20] S. P, Meyn and R. L, Tweedie. "Markov Chains and Stochastic Stability ". Springer-Verlag, 1993.

[21] U. Grenander. "Tutorial in Pattern Theory. Technical Report, Providence". Division of Applied Mathematichs, Brown University, 1983.

[22] W. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Application". Biometrika, 57: 97–109, 1970.