Adaptive MCMC: Background, Theory, and Simulations [*]

Supervisor: Professor Jeffrey S. Rosenthal[†]
Student: Shuheng Zheng[‡]

Summer 2006

[*] Research is supported by the NSERC Undergraduate Student Research Award

[†] Department of Statistics, University of Toronto, Toronto, Ontario, Canada. M5S 3G3. Email: *jeff@math.toronto.edu*. http://www.probability.ca/jeff

[‡] Department of Statistics, University of Toronto, Toronto, Ontario, Canada. M5S 3G3. Email: *john.zheng@utoronto.ca*

Abstract

This research report was written during the summer of 2006 under Professor Jeffrey Rosenthal with the support of NSERC USRA. There are two major purposes to this research report. The first is to summarize the theoretical foundations of adaptive MCMC, including necessary backgrounds in measure theory, general state-space Markov chains, and non-adaptive MCMC so that a strong undergraduate math student with only undergraduate (non-measure theoretic) probability can understand adaptive MCMC on continuous state-spaces. The second is to report on the results of my simulation findings during those four months, provide connections between the simulation and theory, and raise further questions based on empirical trends.

The report begins by providing an overview of the essential background for understanding general state-space Markov chains. Most of these are taken from my reading notes while reading classic graduate probability texts [7], [6], and [4]. The second part deals with the theory of ordinary MCMC and are mainly based papers by [25], [29], [26], [23]. The final are recent theoretical results from [24] and my simulation results and analysis.

I owe my deepest gratitude to Prof. Rosenthal who took time out of his extremely busy schedule and provided careful guidance and insights which allowed me to stay on track within the tight time schedule of summer projects. My thanks also goes to the Department of Statistics at University of Toronto who provided me with valuable office space and computing facilities.

Shuheng John Zheng
University of Toronto, 2006

Contents

List of Figures

Chapter I

Measure Theory and Markov Chains

I.1 Measure-Theoretic Probability

This section will reveal how concepts of measure theory is applied to probability and Markov chains.

I.1.1 Foundations

Probability Space Any standard construction of measure-theoretic probability will begin by defining the sample space, events, and probability measure. The sample space Ω is the set of all possible experimental outcomes. For example, the n-repetition coin tossing sample space will be the set $\{H, T\}^n$. Events are a collection of possible measurable subsets on the sample space denoted $\mathscr{F}$. If we are to denote the event where the first toss is a head in a 3-repetition coin toss sample space, the event will be $\{HHH, HHT, HTH, HTT\}$. For technical reason, we would like this collection "events" to satisfy the *σ-field* property so that unions of events would still be proper events.

Probabilities must be assigned to all these events in a way that is consistent with natural intuition. That is, the probability of anything happening is 1 and it is countably additive.

Definition I.1.1. A measure $\mathbb{P}$ is called a probability measure if $\mathbb{P}(\Omega) = 1$

Definition I.1.2. A sample space Ω, an σ-field $\mathscr{F}$ containing Ω, and a probability measure $\mathbb{P}$ on $\mathscr{F}$ together forms the probability triplet: $(\Omega, \mathscr{F}, \mathbb{P})$.

Of course, the usual measure continuity properties are carried over.

Theorem I.1.3. *If (measurable) sets $A_n \to A$ monotonically, then* $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.

Random Variables A random variable is not a variable but a function defined from Ω to $\mathbb{R}$. The function usually reports an aspect of the experimental outcome thus causing its value to change each time.

Definition I.1.4. Random Variable
A measurable function $X : \Omega \to \mathbb{R}$ is called a random variable.

These random variables induce probability measures on $\mathbb{R}$ and these measures are called laws.

Definition I.1.5. Probability Laws
Given a random variable X on $(\Omega, \mathscr{F}, \mathbb{P}$, its law $\mathscr{L}(X)$ is defined by the following Borel probability measure on $(\mathbb{R}, \mathscr{B})$:

$$\mathscr{L}(X) = \mathbb{P} \circ X$$

Laws are also called the *distributions*.

Independence Independence is a probability concept not present in measure theory. Intuitively, from the rules of combinatorics, two events are independent if their probabilities "multiply out".

Definition I.1.6. A collection $\{A_\alpha\}, \alpha \in I$ is independent if for each sequence $\alpha_1, \alpha_2, \ldots, \alpha_n \in I$, we have

$$\mathbb{P}(A_{\alpha 1} \cap A_{\alpha 2} \cap \ldots \cap A_{\alpha n}) = \mathbb{P}(A_{\alpha 1})\mathbb{P}(A_{\alpha 2}) \ldots \mathbb{P}(A_{\alpha n})$$

Random variables are independent if the events that they generate through the $\mathbb{R}$ Borel sets are all independent.

I.1.2 Expectations and Conditional Expectations

Expectation Expectation should satisfy the natural notion of average. Or a sum of all the real numbers weighted by the probability measure.

Definition I.1.7. Expectation
The expected value of X, $\mathbb{E}[X]$ is:

$$\mathbb{E}[X] = \int_{\omega\in\Omega} X(\omega)\mathbb{P}(d\omega)$$

where the integral is the Lebesgue integral. $\mathbb{E}[X]$ is often symbolized by μ.

This definition of the expectation allows the carrying over of three standard integral convergence theorems.

Theorem I.1.8. *Monotone Convergence Theorem*
If a sequence $\{X_n\}$ converges to X monotonically almost everywhere, then $\lim_{n\to\infty}\mathbb{E}[X_n] = \mathbb{E}[X]$

Theorem I.1.9. *Dominated Convergence Theorem*
If $X_n \to X$ almost everywhere, and

- $\exists Y s.t. |X_n| \leq Y$ *almost everywhere*
- $\mathbb{E}[Y] < \infty$

then $\lim_{n\to\infty}\mathbb{E}[X_n] = \mathbb{E}[X]$

Theorem I.1.10. *Fatou's Lemma*
If $X_n \geq 0$ almost everywhere, then $\mathbb{E}[\liminf_{n\to\infty} X_n] \leq \liminf_{n\to\infty}\mathbb{E}[X_n]$

Theorem I.1.8 requires that $X_n \to X$ with probability 1. This condition can be weakened to convergence in probability.

Theorem I.1.11. *Probabilist's Dominated Convergence Theorem*
The sufficient condition of dominated convergence theorem only requires $X_n \to X$ in probability as opposed to with probability 1.

Proof. The key to the proof is to use the fact that if every subsequence of a sequence has a further subsequence that converges to a constant c, then the sequence itself must converge to c.

Given any subsequence of $\{\mathbb{E}[X_n]\}$, the associated $\{X_{n(k)}\}$ still converges to X in probability. From this subsequence, we can find a further subsequence that converges to X with probability one. By assumption, this sub-subsequence is bounded above by a random variable Y with finite expectation. Applying theorem I.8, this sub-subsequence's expectation converges to $\mathbb{E}[X]$. Hence every subsequence of $\{\mathbb{E}[X_n]\}$ has a further subsequence that converges to $\mathbb{E}[X]$; therefore, $lim_{n\to\infty}\mathbb{E}[X_n] = \mathbb{E}[X]$ □

Conditional Probability Conditioning is one of the most powerful tool in probability theory. A solid understanding of it will be necessary to understand stochastic processes such as random walk, Poisson process, martingales, and of course, Markov chains

The elementary definition of conditional probability and conditional expectation is as follows

Definition I.1.12. Elementary Conditional Probability and Conditional Expectation

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \qquad \mathbb{E}[X|B] = \frac{\mathbb{E}[X * \mathbb{I}_B]}{\mathbb{P}(B)}$$

This definition requires that $\mathbb{P}(B) > 0$; however, this is often too restrictive. One common problem is if $B = \{X = c\}$ where X is a random variable that is absolutely continuous with respect to the Lebesgue measure. In this case, $\mathbb{P}(B) = 0$.

Conditional Expectation To remedy the above difficulty, we define conditional expectation using measure theory and Hilbert space theory.

Theorem I.1.13. *L^2 space*
Let space $L^2(\Omega, \mathscr{G}, \mathbb{P})$ be the space of $\mathscr{G}$-measurable random variables defined on Ω that have finite variances. This space with the inner-product $< X, Y >= \mathbb{E}[XY]$ is a Hilbert space.

Proof. It's simple to verify that this satisfies the definition of an indder-product. We then just show that this is a linear space. If X and Y are $\mathscr{G}$-measurable then $aX + bY$ is obviously $\mathscr{G}$-measurable. Furthermore, expanding $(aX + bY)^2$ and using the fact that $\mathbb{E}[abXY] \leq ab\sqrt{Var[X]Var[Y]}$ will show that $aX + bY$ has finite variance.

Completeness is obvious as this is just the standard L^2 space because the expectation is merely an integral. □

The original concept of conditional expectation came out of the search for the best predictor. "Best" is meant in the L^2 sense.

Definition I.1.14. L^2 Conditional Expectation
Assume X and Y are random variables with finite variances on the sample space Ω. $\mathbb{E}[X|Y]$ is a random variable in $L_2(\Omega, \sigma(Y), \mathbb{P})$ such that

$$\|X - \mathbb{E}[X|Y]\|_{L2} \leq \|X - Z\|_{L2} \qquad \forall Z \in L_2(\Omega, \sigma(Y), \mathbb{P})$$

Theorem I.1.15.

- *The above conditional expectation exists and is unique up to a set of probability zero.*
- *The above condition is equivalent to requiring that $X - \mathbb{E}[X|Y]$ is orthogonal to all random variables from $L_2(\Omega, \sigma(Y), \mathbb{P})$*

Proof. These are just results from the projection theorem from Hilbert space theory and can be found in chapter 6 of [9]. □

Now it's just a simple extension to define $\mathbb{E}[X|\mathscr{G}]$ where $\mathscr{G}$ is a sub σ-field of $\mathscr{F}$ as the unique projection of X onto $L_2(\Omega, \mathscr{G}, \mathbb{P})$.

The requirement for finite variance of X is often too strong because there are large number of random variables (i.e. Cauchy) that do not have finite variance (or in this case, no mean!). We thus resort to the analog of the orthogonality condition as the definition.

Definition I.1.16. General Conditional Expectation
Given X defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and a sub σ-field $\mathscr{G}$, we define $\mathbb{E}[X|\mathscr{G}]$ to be the $\mathscr{G}$-measurable random variable such that

$$\mathbb{E}[(X - \mathbb{E}[X|\mathscr{G}]) * \mathbb{I}_G] = 0 \qquad \forall G \in \mathscr{G}$$

Theorem I.1.17. *The conditional expectation defined as above always exist and is unique a.e.*

Proof. This proof is a semi-constructive proof that builds upon the intuition of non-measure-theoretic conditional expectation.

Recall that $\mathbb{E}[X|A] = \frac{\mathbb{E}[X*\mathbb{I}_A]}{\mathbb{P}(A)}$. Now this division presents problems on null measure sets. To alleviate the issue, the measure-theoretic "division" must be used, namely the Radon-Nikodym derivative.

For every $A \in \mathscr{G}$, denote $M_X(A) \triangleq \mathbb{E}[\mathbb{I}_A X]$. This is a signed (bounded) measure which is absolutely continuous with respective to $\mathbb{P}|\mathscr{G}$ (measure restricted to $\mathscr{G}$). By the Radon-Nikodym theorem, there exists a $\mathscr{G}$-measurable random variable $\hat{X}$ such that

$$\mathbb{E}[\mathbb{I}_A\hat{X}] = \int_A \hat{X} d\mathbb{P} = M_X(A) = \mathbb{E}[\mathbb{I}_A X] \qquad \forall A \in \mathscr{G}$$

and this $\hat{X}$ is unique almost everywhere. It satisfies the above definition of conditional expectation. □

Finally we can define conditional probability.

Definition I.1.18. General Conditional Probability

$$\mathbb{P}(A|\mathscr{G}) = \mathbb{E}[\mathbb{I}_A|\mathscr{G}]$$

Corollary I.1.19. *The abstract definition of conditional probability is equivalent to the elementary one (Definition I.1.12).*

Proof. Given $\mathbb{P}(B) > 0$, we will examine $\mathbb{P}(A|\sigma(B, B^c))$. This is equal to, by definition, $\mathbb{E}[\mathbb{I}_A|\{\emptyset, B^c, B, \Omega\}]$. We hypothesize that this is equal to the $\{\emptyset, B^c, B, \Omega\}$-measurable random variable

$$\mathbb{E}[\mathbb{I}_A|\{\emptyset, B^c, B, \Omega\}](\omega) = \begin{cases} \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} & \omega \in B \\ \frac{\mathbb{P}(A\cap B^c)}{\mathbb{P}(B^c)} & \omega \in B^c \end{cases}$$

This function is piecewise constant so it is definitely $\{\emptyset, B^c, B, \Omega\}$-measurable. We also need to make sure that

$$\mathbb{E}[\mathbb{E}[\mathbb{I}_A|\{\emptyset, B^c, B, \Omega\}] * \mathbb{I}_G] = \mathbb{E}[\mathbb{I}_A * \mathbb{I}_G] \qquad \text{where } G \in \{\emptyset, B^c, B, \Omega\}$$

The $\emptyset, \Omega$ cases are trivial. If $G = B^c$

$$\mathbb{E}[\mathbb{E}[\mathbb{I}_A|\{\emptyset, B^c, B, \Omega\}] * \mathbb{I}_{B^c}] = \int_{\omega\in B^c} \frac{\mathbb{P}(A \cap B^c)}{\mathbb{P}(B^c)} d\mathbb{P} = \mathbb{P}(A \cap B^c) = \mathbb{E}[\mathbb{I}_A * \mathbb{I}_{B^c}]$$

the other case is very similar.

So our conditional probability $\mathbb{P}(A|B)$ is merely the case where the B event from the σ-field has "occurred". □

Lemma I.1.1. *Conditional Density*
Given two random variables X, Y defined on the same probability space with Lebesgue $\mathbb{R}^2$ density $f(x,y)$. Then if $\mathbb{E}[g(Y)] < \infty$

$$\mathbb{E}[g(Y)|X] = \int_{y\in\mathbb{R}} g(y) f_{Y|X}(y;X)\lambda(dy)$$

where $f_{Y|X}$ is any function (with random parameter X) that satisfies

$$f_{Y|X}(y;X=x)\int_{t\in\mathbb{R}} f(x,t)\lambda(dt) = f(x,y)$$

Proof. The process of verification is identical to the corollary above and is left to the reader. The answer can also be found in [7]. □

I.1.3 Weak Convergence and Characteristic Functions

One important question that is often asked if whether a sequence of Borel probability measures $\mu, \mu_1, \mu_2, \ldots$ converges to a particular Borel measure. One way to define is convergence that is of importance in applied probability is weak convergence.

Definition I.1.20. Weak Convergence
A sequence of Borel probability measures $\{\mu_n\}$ converges to μ weakly if

$$\int_{\mathbb{R}} f d\mu_n \to \int_{\mathbb{R}} f d\mu \qquad \forall f \text{ bounded \& continuous}$$

Now a sequence of random variables $X_1, X_2, X_3 \ldots$ converges weakly to X if their respective laws converge weakly.

One theorem that can be proven is the following equivalence between the cumulative distribution function convergence and weak convergence. The proof is quite complicated and is omitted. It can be found in Rosenthal.[27]

Theorem I.1.21. *The following are equivalent*

1. *$\{\mu_n\}$ converges to μ weakly*

2. *$\mu_n(A) \to \mu(A)$ for all measurable sets A with measure zero boundaries*

3. *$\mu_n((-\infty, x]) \to \mu((-\infty, x])$ for all $x \in \mathbb{R}$ such that $\mu(\{x\}) = 0$*

Tightness The notion of tightness of probability laws is important because it has a sequential compactness consequence.

Definition I.1.22. A set $\mathscr{P}$ of probability laws is tight if for every $\epsilon > 0$, there exists $a < b$ such that $\mathbb{P}([a, b]) > 1 - \epsilon$ for all $\mathbb{P} \in \mathscr{P}$.

Intuitively, a collection of tight laws place almost all of their probability mass on some compact interval in $\mathbb{R}$. One powerful consequence is the following.

Theorem I.1.23. *Let $\{\mu_n\}$ be a sequence of tight probability laws. Then there is a subsequence such that $\mu_{n(k)} \to \mu$ for some μ.*

In another words, tight sequence of probability laws will imply that they are sequentially compact in the weak convergence sense. There are several different approaches to the proof of this theorem. One of the easier ones [27] involves looking at the sequence $F_n(x) = \mu_n((-\infty, x])$ of cumulative distribution functions (cdfs). The *Helly selection principle* can show that for any sequence of cdfs, there exists a convergent subsequence. After that, tightness of the probability laws can generalize the convergence of subsequence of cdfs to convergence of their respective probability laws. Other approaches uses abstract tools such as Stone-Weierstrass, metrization, and Tychnoff's theorem. [6].

The following theorem allows for empirical verification of weak convergence.

Theorem I.1.24. *Glivenko-Cantelli*
Let $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$ be the empirical distribution function when $X_1, X_2, \ldots, X_n$ are coming from cdf F. Then $F_n \to F$ uniformly almost everywhere.

Proof. Proof is given in [6]. □

Metrization Weak convergence can be stated in terms of convergence in a certain metric space.

Definition I.1.25. Lévy-Prohorov Metric
For any subset A of $\mathbb{R}$ and an $\epsilon > 0$, let A^ϵ be the ϵ-neighborhood of A.

$$A^\epsilon = \{y \in \mathbb{R} : |x - y| < \epsilon \text{ for some } x \in A\}$$

Then for any two probability laws μ and ν, the Prohorov metric is

$$\rho(\mu, \nu) = \inf\{\epsilon > 0 : \mu(A) \leq \nu(A^\epsilon) + \epsilon \qquad \forall A \in \mathscr{B}\}$$

Theorem I.1.26. *ρ is a metric on the set of all probability laws, and for probability laws $\{\mu_n\}$ and μ, the following are equivalent:*

(1) $\int f d\mu_n \to \int f d\mu \qquad \forall f$ bounded & continuous

(2) $\rho(\mu_n, \mu) \to 0$

The proof [6] is quite technical and will be omitted here.

Characteristic Functions Weak convergence is closely related to the convergence of the respective characteristic functions.

Definition I.1.27. Characteristic function
A characteristic function ϕ of a random variable X with law μ is just its Fourier transform

$$\phi_X(s) = \int_{\mathbb{R}} e^{isx} d\mu(x)$$

The nice thing about the characteristic function is that it exists as long as the Fourier transform exists. In fact, the following properties can be shown just from standard properties of the Fourier transform and can be found in [11]

Theorem I.1.28. *The characteristic function ϕ_X has the following properties*

1. $\phi_X(0) = 1, |\phi_X(s)| \leq 1$ *for all* $s \in \mathbb{R}$

2. *ϕ_X is uniformly continuous and semi-positive definite.*

There is a more important correlation between the characteristic function's existence and the moments $\mathbb{E}(|X^k|)$s' existence.

Theorem I.1.29. *Let X be a random variable with char. function $\phi(t)$.*

(1) If $\mathbb{E}(|X|^n) < \infty$ for some *positive integer n, then ϕ is C^n and $\phi^{(k)}(t) = i^k\mathbb{E}(X^k e^{itX})$ for all $k = 1, \ldots, n$.*

(2) If n is an positive even *integer and $\phi^{(n)}(0) < \infty$, then $\mathbb{E}(|X|^k) < \infty$ for $k = 1, \ldots, n$.*

Proof. (1) We know that all moments from 1 to n are finite by Hölder's inequality and we prove this theorem by iterating on k. First for the case that $k = 1$ i.e. we want to show that $\phi'(t) = i\mathbb{E}(Xe^{itX})$ and it is continuous.

Note

$$\left|\frac{e^{i(t+h)X} - e^{itX}}{h}\right| = \left|\frac{e^{ihX} - 1}{h}\right| \leq \frac{|hX|}{|h|} = |X| \text{ by the convexity of } e^x$$

$$\phi'(t) = \lim_{h\to 0}\left(\frac{\mathbb{E}(e^{i(t+h)X}) - \mathbb{E}(e^{itX})}{h}\right) = \lim_{h\to 0}\left(\mathbb{E}\left[\frac{e^{i(t+h)X} - e^{itX}}{h}\right]\right)$$

The finiteness of the first moment allows us to use the DCT and arrive at $\phi'(t) = i\mathbb{E}(Xe^{itX})$ and this is continuous also by the DCT. If we generalize the above DCT inequality into

$$\left|\frac{X^k e^{i(t+h)X} - X^k e^{itX}}{h}\right| = |X^k|\left|\frac{e^{ihX} - 1}{h}\right| \leq |X^k|\frac{|hX|}{|h|} = |X|^{(k+1)}$$

then iteration allows us to say $\phi(t)$ is continuously differentiable n times and $\phi^{(k)}(t) = i^k\mathbb{E}(X^k e^{itX})$.

(2) We prove this by induction on n. First consider the case that $n = 2$. That means ϕ has a second derivative at zero, thus the first derivative exists around zero which further implies that ϕ is differentiable in an open interval around 0. For some small $h > 0$, denote

$$\Delta^2\phi(h) = \frac{\phi(h) - 2\phi(0) + \phi(-h)}{h^2} = \frac{\mathbb{E}(e^{ihX} - 2 + e^{-ihX})}{h^2}$$

The above function $\Delta^2\phi(h)$ is defined for a open interval about zero. Take the limit as $h \to 0$ and use l'Hopital's rule twice and the uniform continuity of ϕ and ϕ' to obtain

$$\lim_{h\to 0} \Delta^2\phi(h) = \lim_{h\to 0} \frac{\phi'(h) - \phi'(-h)}{2h} = \phi''(0)$$

Notice that $X^2 = \lim_{h\to 0}(\frac{2(1-\cos(hX))}{h^2})$ and $e^{ihX} + e^{-ihX} = 2\cos(hX)$. Putting these pieces together and using Fatou's lemma gives us a upper bound on $\mathbb{E}(X^2)$ in terms of $\phi''(0)$.

$$\begin{aligned}\mathbb{E}(X^2) =& \mathbb{E}\left[\liminf_{h\to 0}\left(\frac{2(1-\cos(hX))}{h^2}\right)\right] \leq \liminf_{h\to 0} \mathbb{E}\left(\frac{2(1-\cos(hX))}{h^2}\right) \\ =& -\liminf_{h\to 0} \Delta^2\phi(h) = -\phi''(0) < \infty\end{aligned}$$

Now for the induction step, we assume that the theorem is true for all $n \leq k$ for k even and try to prove that the theorem still holds for $n = k+2$.

If $\phi^{(k+2)}(0) < \infty$, then $\phi^{(k)}(0)$ is finite as well. So $\mathbb{E}(X^k) < \infty$ by induction hypothesis. By result (1), we have $\mathbb{E}(X^k e^{itX}) = \frac{\phi^{(k)}(t)}{i^k}$ Let this function be called $\psi(t)$ so $\psi''(t) = \frac{\phi^{(k+2)}(t)}{i^k}$ and therefore exists at zero by assumption. We can repeat the above style argument with $\psi(t)$ to arrive at the conclusion that $\mathbb{E}(X^{(k+2)}) \leq -\psi''(0) < \infty$ □

Uniqueness and Continuity One important feature of characteristic functions is that they are in one-to-one correspondence with probability laws.

Theorem I.1.30. *Uniqueness of Characteristic Functions*
Characteristic functions are in 1-to-1 correspondence with probability laws.

Proof. It is trivial to show that for every law there is a characteristic function as this is just the integrability of e^{itX} with respect to a Borel probability measure. The converse direction requires the Fourier uniqueness theorem from Fourier analysis, which states

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it}\phi(t)dt = \frac{1}{2}\mu(\{a\}) + \mu((a,b)) + \mu(\{b\})$$

where the improper integral is either construed as the Lebesgue integral or the Cauchy principal value. This theorem allows us to conclude that two random variables with same characteristic function will have the same probability mass on all intervals. The Carathéodory extension theorem leads to the conclusion that their probability measures are identical. □

The second important feature is that probability measures converge weakly iff their corresponding characteristic functions converge pointwise.

Theorem I.1.31. *Lévy continuity theorem*
Probability laws $\{\mu_n\}$ converge weakly to μ iff their characteristic functions $\{\phi_n\}$ converge pointwise to the characteristic function ϕ of μ.

Proof. It is straightforward to show that weak convergence implies convergence of characterstic functions because $\cos(tX)$ and $\sin(tX)$ are a bounded continuous functions. The converse direction is the difficult part. We will give a sketch of the proof of the converse and the full detail can be found in [27].

A series of lemma is needed the first of which states that a sequence of probability measures is tight if their characteristic functions converge to a function that is continuous at zero. In our case, ϕ, being a characteristic function, is uniformly continuous everywhere, so $\{\mu_n\}$ is tight.

Suppose that a subsequence of $\{\mu_n\}$ converges to ν. By the forward result of this theorem, we know that it implies $\phi_{n(k)} \to \phi_\nu$. But ϕ_ν must be the same as ϕ by the uniqueness of limit. Which implies that $\nu = \mu$ by Fourier uniqueness.

The final lemma we need to complete the proof states that if a tight sequence of probability measures has only one possible weak limit, then that sequence actually does converge to the weak limit □

I.1.4 Existence of Stochastic Processes

A *stochastic process* is a collection $\{X_t : t \in T\}$ that is defined on the same probability space. Usually they are dependant in some way and the set T is just the natural numbers. It is difficult to explicitly define the probability measure $\mathbb{P}$ on Ω and X_t's as explicit functions; rather it's more common to specify a stochastic process via its finite-dimensional distributions (fdds). Namely a collection $\{\mu_{t_1,t_2,\ldots,t_k} : k \in \mathbb{N}, t_i \in\}$ such that $\mu_{t_1,t_2,\ldots,t_k}(A) = \mathbb{P}((X_{t_1}, X_{t_2}, \ldots, X_{t_k}) \in A)$. Under quite general conditions, the corresponding probability space with the stochastic process can be shown to exist if its fdds are specified.

Theorem I.1.32. *Kolmogorov's Existence Theorem*
If the fdds for a stochastic process satisfies the following consistency conditions*:*

(1) (Permutation invariant) Given any permuation π of (1, 2, . . . , k), then for any set $t_1, t_2, \ldots, t_k \in T$ and Borel sets $A_1, A_2, \ldots, A_k \in \mathbb{R}$, we have

$$\mu_{t_1,t_2,\ldots,t_k}(A_1 \times A_2 \times \ldots \times A_k) = \mu_{t_{\pi(1)},t_{\pi(2)},\ldots,t_{\pi(k)}}(A_{\pi(1)} \times A_{\pi(2)} \times \ldots \times A_{\pi(k)})$$

(2) (Marginal distributions) Under the same setup

$$\mu_{t_1,t_2,\ldots,t_k}(A_1 \times A_2 \times \ldots \times A_{(k-1)} \times \mathbb{R}) = \mu_{t_1,t_2,\ldots,t_{(k-1)}}(A_1 \times A_2 \times \ldots \times A_{(k-1)})$$

Then there exists a probability space $(\mathbb{R}^T, \mathscr{F}^T, \mathbb{P})$ and random variables $\{X_t : t \in T\}$ such that the random variables have the given finite-dimensional distributions.

The proof of this theorem is extremely complicated and is presented in [4].

I.2 General State-Space Markov Chains

I.2.1 Fundamentals

Many processes in nature exhibit a type of *memoryless* property, meaning the development of future events depends on the state of nature of the present (and possible a finite past) but not on the entire history of nature. Mathematically, we can translate this into the *Markov property*

Definition I.2.1. Markov Property
A stochastic process $\{X_n\}$ exhibits the *Markov property* if $\mathscr{G}_h = \sigma(X_0, X_1, \ldots, X_{h-1})$ and

$$\mathbb{E}[f(X_h)|\mathscr{G}_h] = \mathbb{E}[f(X_h)|X_{h-1}] \qquad \forall h \in \mathbb{N} \text{ and f bounded-continuous}$$

In order to make the analysis of these chains more manageable, we typically require that the Markov chain by time-homogenous.

Definition I.2.2. Time-Homogeneity
A stochastic process $\{X_n\}$ is called *time-homogenous* if

$$\mathbb{P}(X_k \in B|X_{k-1} = x) = \mathbb{P}(X_1 \in B|X_0 = x) \qquad \forall x \in \chi \text{ and } B \in \mathscr{F}$$

The random variables in Markov chains can be extended from a $\mathbb{R}$ range to a range on a topological space χ with a σ-field $\mathscr{H}$. Although in practice, χ is usually some m-dimensional Euclidean space.

Obviously the behavioral-defining key of these Markov chains is how it moves from X_n to X_{n+1}. These are called transition kernels which give the probability of moving from one point to a measurable subset of χ in one transition. They are essentially generalized probability laws.

Definition I.2.3. A function $P : \chi \times \mathscr{H} \to [0,1]$ such that
1) $P(x, \cdot) : \mathscr{H} \to [0,1]$ is a probability measure for all $x \in \chi$
2) $P(\cdot, A) : \chi \to [0,1]$ is a measurable function for all $A \in \mathscr{H}$
is called a *transition kernel.*

Now we can show that every transition kernel corresponds to one Markov chain just like the relation between transition matrices and discrete state-space Markov chains. [11]

Theorem I.2.4. *Existence of Markov Chains*
Given a measurable state-space $(\chi, \mathscr{H})$, a probability measure μ, and a transition kernel P, there exists a probability space $(\Omega, \mathscr{F}, \mathbb{P}_\mu)$ and an associated stochastic process $\{X_n\}$ such that

$$\mathbb{P}_\mu(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) =$$
$$\int_{x_0 \in A_0} \int_{x_1 \in A_1} \cdots \int_{x_{n-1} \in A_{n-1}} \mu(dx_0) P(x_0, dx_1) \ldots P(x_{n-1}, A_n) \qquad \forall n \in \mathbb{N}$$

where μ is known as an initial distribution *and the stochastic process is known as a Markov chain.*

Proof. We can let

$$\nu_{t_0,t_1,t_2,\ldots,t_k}(H) = \mathbb{P}_\mu((X_{t_0}, X_{t_1}, \ldots, X_{t_k}) \in H)$$

This probability measure satisfies the consistency condition listed in I.4. Using a more general version of Kolmogorov's existence theorem for topological spaces [6] and with some technical restrictions on the topology of χ (countable), we can show that the stochastic process exists. □

The above stochastic process satisfies the Markov property and time-homogeneity. The general-case proof of Markov property when $P(x, \cdot)$ does not have a density is quite complicated and omitted here ([7], chapter 6).

n-Step Transition Kernel To make matters more convenient, let's denote a *n-step transition kernel* P^n:

$$P^n(x, \cdot) = \mathbb{P}_{\delta(x)}(X_n \in \cdot) = \int_{y_0 \in \chi} \delta_x(dy_0) \int_{y_1 \in \chi} P(y_0, dy_1) \dots \int_{y_{n-1} \in \chi} P(y_{n-1}, \cdot)$$

where δ_x is the Dirac delta measure at x and $P^0(x, \cdot) = \delta_x(\cdot)$

One immediate consequence of this definition is the Chapman-Kolmogorov equation.

Theorem I.2.5. *Chapman-Kolomogorv*
For any m *such that* $0 \leq m \leq N$

$$P^N(x, \cdot) = \int_{y \in \chi} P^m(x, dy) P^{N-m}(y, A) \qquad \forall x \in \chi$$

I.2.2 ϕ-Irreducibility

A "irreducible" Markov chain is a Markov chain that cannot be reduced to two separate Markov chains, meaning that all states should eventually be able to reach other states. The probability of reaching a specific state if χ is uncountable is almost always zero; therefore, we define irreducibility as the property where all "interesting" sets can be reached in finite time starting anywhere.

Definition I.2.6. Time to first arrival
The time to first arrival of a set A τ_A is defined as $\inf\{n \in \mathbb{N} : X_n \in A\}$

Definition I.2.7. ϕ-irreducibility
A Markov chain $\{X_n\}$ is ϕ-irreducible for a σ-finite measure ϕ on χ if for all measurable sets $A \subset \chi$ of positive measure, we have

$$\mathbb{P}(\tau_A < \infty | X_0 = x) > 0 \qquad \forall x \in \chi$$

Note that ϕ-irreducibility does not necessarily imply ordinary irreducibility from discrete Markov chains theory, only indecomposibility.

I.2.3 Recurrence and Periodicity

Recurrence

Definition I.2.8. Harris recurrent
A set $A \in \mathscr{H}$ is called *Harris recurrent* if

$$\mathbb{P}(X_n \text{ visits A infinitely often } | X_0 = x) = 1 \qquad \forall x \in \chi$$

A Markov chain is called a Harris recurrent chain it is ϕ-irreducible and every set with positive ϕ measure is Harris recurrent

Periodicity In discrete state-space Markov chains, periods are assigned to each individual state and the chain is aperiodic if all states have period 1. The period of individual states is the greatest common divisor of all times $N_1, N_2, N_3 \ldots$ such that there is a positive probability of coming back to i.

General state-space chains have a more complicated construction because the states are usually uncountable. So we must look at periodicity of the whole chain and not just individual states.

Definition I.2.9. Periodicity
A Markov chain with stationary measure π is *periodic* if there exists disjoint subsets $A_1, A_2, \ldots, A_d \subset \chi$ (known as a *periodic decomposition*) such that $P(x, X_{i+1}) = 1 \quad \forall x \in A_i(1 \leq i \leq d-1,\ P(x, A_1) = 1 \quad \forall x \in A_d$, and $\pi(A_1) > 0$. If there does not exist such $d \geq 2$, then the chain is *aperiodic*.

I.2.4 Stationary measures

We want to know whether there is an equilibrium distribution in the Markov chain because that is the most plausible limiting distribution. An equilibrium distribution, called a stationary measure, is essentially a distribution that does not change after a one-step transition of the Markov chain.

Definition I.2.10. Stationary Measure
A probability measure $\pi(\cdot)$ on $(\chi, \mathscr{H})$ is a *stationary measure* for the Markov chain with transition kernel P if

$$\pi(A) = \int_{x \in \chi} P(x, A)\pi(dx) \qquad \forall A \in \mathscr{H}$$

Analogous to discrete-time Markov chains, there is an intricate connection between recurrence, transience, and existence of stationary measure. In general, a recurrent ϕ-irreducible chain has an stationary measure and the stationary measure of atomic sets (see I.2.6) is the reciprocal of the expected mean return time $\mathbb{E}(\tau_A)$ to that atom. More detail can be found in [19].

Reversibility If stationarity means a forward-moving equilibrium, there can also be a bidirectional equilibrium. That is the probability of going forward or backward in time to a set A is equal once the chain reaches the stationary measure. Markov chains possessing this property is known as *reversible* Markov chain.

Definition I.2.11. Reversible Markov chain
Given Markov chain is reversible with transition kernel P, suppose it is under the stationary distribution so $X_n \sim \pi\ \forall n \in (-\infty, \infty)$, then if

$$\mathbb{P}(X_{n+1} \in B | X_n \in A) = \mathbb{P}(X_n \in B | X_{n+1} \in A) \qquad \forall A, B \in \mathscr{H}$$

the chain is called *reversible*

A useful way to check this is the following

Theorem I.2.12. *Reversibility*
A Markov chain $\{X_n\}$ is reversible iff the following relation is satisfied

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx) \qquad \forall x, y \in \chi \tag{I.1}$$

Proof. Suppose that a Markov chain is reversible, using the Bayes rule for conditional probability [11]:

$$\mathbb{P}(X_n \in B | X_{n+1} \in A) = \frac{\mathbb{P}(X_{n+1} \in A | X_n \in B)\mathbb{P}(X_n \in B)}{\mathbb{P}(X_{n+1} \in A)}$$

since the Markov chain is reversible $\mathbb{P}(X_n \in B | X_{n+1} \in A) = \mathbb{P}(X_{n+1} \in B | X_n \in A)$, so

$$\begin{aligned}
\mathbb{P}(X_{n+1} \in B | X_n \in A) &= \frac{\mathbb{P}(X_{n+1} \in A | X_n \in B)\mathbb{P}(X_n \in B)}{\mathbb{P}(X_{n+1} \in A)} \\
\Rightarrow \mathbb{P}(X_n \in A, X_{n+1} \in B) &= \mathbb{P}(X_n \in B, X_{n+1} \in A) \\
\Rightarrow \int_{x \in A} \int_{y \in B} \pi(dx)P(x, dy) &= \int_{x \in A} \int_{y \in B} \pi(dy)P(y, dx) \quad \forall A, B \in \mathscr{H} \\
\Rightarrow \pi(dx)P(x, dy) &= \pi(dy)P(y, dx) \quad \forall x, y \in \chi
\end{aligned}$$

The converse verification is very direct and is left to the reader. □

Remark. It is straightforward to check that every π that satisfies the above condition is also a stationary measure.

I.2.5 Ergodicity

To properly characterize the limiting behaviour of Markov Chains, we must define the criteria for convergence. Although weak convergence is one criteria, there is a stronger notion that can be used here: convergence in *total variation distance.*

Definition I.2.13. Total Variation Distance
Given two probability measures μ and ν defined on the same probability space $(\Omega, \mathscr{F})$, the total variation distance is

$$\|\mu(\cdot) - \nu(\cdot)\| = \sup_{A \in \mathscr{F}} |\mu(A) - \nu(A)|$$

Lemma I.2.1. *Convergence in total variation distance implies weak convergence.*

Proof. Given a sequence of probability measures $\{\mu_n\}$ such that $\|\mu_n(\cdot) - \mu(\cdot)\| \to 0$ and a continuous function f bounded by M, we have:

$$\left| \int f d\mu_n - \int f d\mu \right| \leq \sup_{|g| \leq M} \left| \int g d\mu_n - \int g d\mu \right| = 2M\|\mu_n - \mu\| \to 0$$

by proposition 3.1 from [25] □

With this definition, we can state the main convergence theorem of Markov chains.

Theorem I.2.14. *Markov Chain Convergence Theorem*
Given an ϕ-irreducible, aperiodic Markov chain on $(\chi, \mathscr{H})$ with a stationary distribution π, we have

$$\|\mathbb{P}(X_n \in \cdot | X_0 = x) - \pi(\cdot)\| \to 0$$

for 1) π a.e. $x \in \chi$ or 2) $\forall x \in \chi$ if the chain is Harris recurrent

Proof. The proof of this theorem is based on a coupling argument and can be found in [19] □

I.2.6 Qualitative Convergece

Atomic Sets The idea of the atomic sets is fairly important. They are essentially a subset of χ that can be treated as one unit probability-wise.

Definition I.2.15. Atom
A set $C \in \mathscr{H}$ is called an *atom* if there exists a probability measure ν such that $\mathbb{P}(x, \cdot) = \nu(\cdot)\ \forall x \in C$. If this atom has positive irreducible measure then it's called an *accessible atom.*

Minorization and Uniform Ergodicity There are two important conditions in determining uniform and geometric ergodicity.

Definition I.2.16. Small sets
A set $C \in \mathscr{H}$ is *small* if there exists a positive integer N, an $\epsilon > 0$, and a probability measure ν such that

$$P^N(x, \cdot) \geq \epsilon\nu(\cdot) \qquad \forall x \in C$$

called the *minorization condition*

Remark. Actually if the chain is ϕ-irreducible, then any set with positive ϕ measure contains a small set.

So if a Markov chain reaches any element of a small set, the small set can almost be treated as one unit whose eventual probability measure is bounded below by ϵ. All atoms are automatically small with $N = 1$, and $\epsilon = 1$.

The importance of small sets lies in its usefulness in a coupling construction. In the discrete case, coupling two chains can be quite straightforward and done with probability one ([11], 6.4). In the general state-space case, the construction runs two initially independent chains with initial measure $\delta_{x0}(\cdot)$ and $\pi(\cdot)$. These two chains are independent until they both reach the small set C where they are coupled after N transitions with probability ϵ. After the coupling, the two chains running independently will have identical distributions (but not necessarily always the identical sample path on every realization). The nice property of the coupling construction with small sets is that the two chains still follow the identical transition probabilities as if they were independent. More detail is found in [17] and [25].

Theorem I.2.17. *Uniform Ergodicity*
If the whole state-space χ is a small set, then the Markov chain is uniformly ergodic *meaning there exists $M < \infty, \rho < 1$ such that*

$$\|\mathbb{P}(X_n \in \cdot | X_0 = x) - \pi(\cdot)\| \leq M\rho^n \qquad \forall x \in \chi$$

Proof. The proof follows from the above heuristical coupling construction and a coupling inequality. Refer to [25] □

Remark. The converse direction is actually also true. That is, if a Markov chain is uniformly ergodic, then the whole state space is small. More details can be found in [19].

Drift and Geometric Ergodicity The uniform ergodicity conditions are not satisfied very often so we require a slightly weaker notion of qualitative convergence. This is known as *geometric ergodicity* (More technically V-uniform geometric ergodicity). The requirement calls for the *drift condition*

Definition I.2.18. Drift Condition
A Markov chain satisfies the drift condition if there exists constants $0 < \lambda < 1, b < \infty$ and a function $V : \chi \to [1, \infty]$ (notice that ∞ is included), such that

$$\mathbb{E}[V(X_{n+1})|X_n = x] \leq \lambda V(x) + b\mathbb{I}_C(x) \qquad \forall x \in \chi$$

The probabilist's interpretation for this is that the average one-step predictor of V where the function's input is the next Markov chain state can never exceed a linear scaling of V with the exception of a small set C where translation is required to establish an upper bound.

Theorem I.2.19. *V-Uniform Geometric Ergodicity*
If a ϕ-irreducible, aperiodic Markov chain, with stationary measure π, possesses a small set C with associated ϵ and ν, and satisfies the drift condition for some λ, ν, V, and the same C where V is finite for at least one x, then there exists $R < \infty$ and $r < 1$ such that

$$\sup_{|f| \leq V} |\mathbb{E}[f(X_n)|X_0 = x] - \mathbb{E}_\pi[f]| \leq RV(x)r^n \qquad \forall x \in \chi$$

Proof. The proof follows either an analytical approach [19] or the more modern coupling approach which involves the use of bivariate drift condition and a quantitative bound on the probability laws of coupled chains ([25], Sec. 4). □

Remark. The converse direction is also true. More details can be found in [19].

Corollary I.2.20. *Ordinary Geometric Ergodicity*
A Markov chain satisfying the same assumptions as theorem I.2.16 converges in a geometric rate, namely

$$\|P^n(x,\cdot)-\pi(\cdot)\| \leq RV(x)r^n \qquad \forall x \in \chi$$

Proof. By Proposition 3 from [25], we have, for all $x \in \chi$

$$\begin{aligned}\|P^n(x,\cdot)-\pi(\cdot)\| &= \sup_{f:\chi\to[0,1]}\left|\int_{y\in\chi} f(y)dP^n(x,dy) - \int_{y\in\chi} f(y)d\pi(y)\right| \\ &= \sup_{f:\chi\to[0,1]} |\mathbb{E}[f(X_n)|X_0=x]-\mathbb{E}_\pi[f]| \\ &\leq \sup_{|f|\leq V} |\mathbb{E}[f(X_n)|X_0=x]-\mathbb{E}_\pi[f]| \\ &\leq RV(x)r^n\end{aligned}$$

because $V \geq 1$ □

Sub-Geometricity There is a negative condition that guarantees failure of geometric ergodicity (subgeometric ergodicity). The proof is in [26].

Theorem I.2.21. *Given a ϕ-irreducible Markov chain with stationary measure π that is not equivalent to the Dirac delta measure at some point and such that $P(x,\{x\})$ is a measurable function. Then*

$$\sup_{x\in\chi} P(x,\{x\}) = 1 \quad \pi\text{-}a.e.$$

implies that the Markov chain is not geometrically ergodic.

I.2.7 Central Limit Theorems

The most basic central limit theorem is given by the following

Theorem I.2.22. *Uniformly Ergodic CLT*
If a ϕ-irreducible, aperiodic Markov chain $\{X_n\}$ with stationary measure π is uniformly ergodic and $\mathbb{E}_\pi(f^2) < \infty$,

$$\sqrt{n}(\overline{f(X_n)} - \mathbb{E}_\pi(f)) \to N(0,\sigma^2)$$

where

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}(f(X_i) - \mathbb{E}_\pi(f))\right)^2\right]$$

for every initial measure μ

Note: The σ^2 above is the same as $\tau\mathrm{VAR}_\pi(f)$ where τ is the integrated autocorrelation time. Thus a "well-mixing" - low autocorrelation - chain has a lower asymptotic variance. This variance is also true for other forms of the Central Limit Theorem as demonstrated below.

Proof. The original proof, omitted here, is due to Cogburn [5] and is done through the use of mixing processes. □

The immediate implication of this is that all finite state-space Markov chains display the uniformly ergodic CLT.

Geometrically Ergodic CLT

Theorem I.2.23. *Geometrically Ergodic CLT*
If a ϕ-irreducible, aperiodic Markov chain $\{X_n\}$ with stationary measure π is geometrically ergodic, and $f : \chi \to \mathbb{R}$ is a Borel-measurable function that satisfies one of the following conditions:
1) $\mathbb{E}_\pi|f(\cdot)^{2+\delta}| < \infty$ for some $\delta > 0$
2) $\mathbb{E}_\pi[f^2(\cdot)\max(0, \log|f(\cdot)|)] < \infty$
3) $\{X_n\}$ is reversible and $\mathbb{E}_\pi[f^2(\cdot)] < \infty$
Then the central limit theorem as stated above is satisfied

Proof. The original proof of this central limit theorem is attributed to Ibragimov and Linnik [15] and proceeds first by proving the stationary martingale central limit theorem, which roughly says that if a sequence's conditional average does not change (*martingale*) and the unconditional mean and covariance functions do not vary with time (*stationary*), then the process exhibits a CLT. The second proof is by Hobert [13] a construction similar to coupling called *regeneration*. The most recent proof is due to Roberts & Rosenthal [25] that also uses the martingale CLT but followed up with properties of the Poisson PDE. □

Remark. Note that these conditions (uniform or geometric ergodicity plus functional finite moment) guarantees that the integrated autocorrelation time is finite therefore the variance is finite.

Negative Condition There is one important negative condition that guarantees the nonexistence of CLT even when the variance of the functional under the limiting distribution is finite. This would be the case if the convergence rate is too slow.

Theorem I.2.24. *Nonexistence of CLT*
Given a reversible Markov chain that starts in its stationary distribution π, if

$$\lim_{n\to\infty} n \int_{x\in\chi} [h(x) - \pi(h)]^2 * P^n(x, \{x\})\pi(dx) = \infty$$

then the CLT does not hold for the functional h.

Proof. The idea of the proof is to show that the limiting variance diverges if the above condition is satisfied.

$$\begin{aligned}
\sigma^2 &= \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}[h(X_i) - \pi(h(\cdot))]\right)^2\right] \\
&\geq \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}[h(X_i) - \pi(h(\cdot))]\right)^2 \mathbb{I}(X_0 = X_1 = \ldots = X_n)\right] \\
&= \lim_{n\to\infty} \frac{1}{n}\int_{x\in\chi} \mathbb{E}\left[\left(\sum_{i=1}^{n}[h(X_i) - \pi(h(\cdot))]\right)^2 \mathbb{I}(X_0 = X_1 = \ldots = X_n)|X_0 = x\right]\pi(dx) \\
&= \lim_{n\to\infty} \frac{1}{n}\int_{x\in\chi} \left(n[h(x) - \pi(h(\cdot))]\right)^2 P^n(x, \{x\})\pi(dx) \\
&= \infty
\end{aligned}$$

The third line is by the property of conditional expectation. Thus if the variance diverges, the CLT cannot possibly exist. □

Chapter II

Markov Chain Monte Carlo

II.1 The Goal and Traditional Approaches

Goal In applied statistics, many expectations and probabilities cannot be calculated analytically. This is true especially in Bayesian statistics [29] and statistical physics [20]. Most of these problems comes down to simulating a desired distribution and evaluating an expectation, so this will be the focus of this report.

The *Monte Carlo* technique is one that takes advantage of the law of large numbers, which says

Theorem II.1.1. *Weak law of large numbers*
Suppose $\{X_n\}$ is an independent and identically distributed (i.i.d.) sequence of random variables with $\mathbb{E}(X_i) = \mu < \infty$. Then

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X_n} - \mu| > \epsilon) = 0 \quad \forall \epsilon > 0$$

Proof. The proof typically proceeds by making a stronger assumption of finite variance and proving the weak law with Chebyshev's inequality. Then, the a set of truncated random variables are used

$$U_{ni} = \begin{cases} X_i & |X_i| \leq \delta n \\ 0 & |X_i| > \delta n \end{cases}$$

Where ($0 < \delta < 1$). These random variables have finite variances, and by proving the weak law for these truncated variables and taking δ to be smaller and smaller, the weak law can be proved for $\{X_n\}$ □

Remark. The weak law and it's converse holds if *one* of two weaker conditions are satisfied:
1) $n\mathbb{P}(|X_1| > n) \to 0$ and $\mathbb{E}[X_1\mathbb{I}(-n \leq X_1 \leq n)] \to \mu$
2) The characteristic function of X_n, $\phi(t)$, is differentiable at $t = 0$ and $\phi^{'}(0) = i\mu$

Given a desired expectation $\mathbb{E}_\pi[f(X)]$, we can estimate this expectation if π can be simulated and the expectation exists, in which case

$$\frac{1}{n}\sum_{i=1}^{N} f(X_i) \to \mathbb{E}_\pi[f(X)]$$

Of course any (m-dimensional) integral can be interpreted as an expectation by the following transformation

$$\int_{\vec{x}\in A} f(\vec{x})g(\vec{x})d\vec{x} = \mathbb{E}_\pi[f(\vec{X})] \qquad \text{where } \pi(A) = \int_{\vec{y}\in A} g(\vec{y})d\vec{y}$$

Note that the Glivenko-Cantelli theorem (Thm I.1.24) can give a theoretical justification for the approximation of theoretical distribution functions by empirical simulations.

II.2 Metropolis-Hastings Algorithm

Monte Carlo estimation is useful only if the desired density can be simulated; however, in practice, this is almost never the case. In fact, the distribution is usually a multi-dimensional density on some restricted set and the normalizing constant is not even known very often.

Monte Carlo Markov Chain attempts to provide a solution by using a Markov chain with the same state-space as the support of the target density π_d such that the limiting distribution is the target distribution. The nature of these problems will almost always mean that the target distribution possesses a density. The two main algorithms that generates this Markov chain are the Metropolis-Hastings algorithm and the Gibbs sampler. Other hybrid algorithms also exists [20]. The algorithm explored in this section is the Metropolis-Hastings algorithm.

II.2.1 Description

Definition II.2.1. Metropolis-Hastings Algorithm
Given a target density π_d with respect to any measure (except where specified) defined on a set S, we provide a proposal transition kernel Q with density q that is also defined on this set S, and finally choose an initial value $X_0 \in S$.
1. Generate a proposal $Y_{n+1} \sim Q(X_n, \cdot)$
2. Accept the this proposal by setting $X_{n+1} = Y_{n+1}$ with probability $\alpha(X_n, Y_{n+1})$ where

$$\alpha(x, y) = \min \left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \right)$$

3. Otherwise reject the proposal by setting $X_{n+1} = X_n$
4. Repeat
The $\{X_n\}$ chain is called the *simulation chain* and the $\{Y_n\}$ chain is called the *proposal chain*.

Remark. We do not have to worry about the denominator of $\alpha(x, y)$ being zero. $\pi_d(x)$ will not be zero because the chain would never arrive at a place with zero π-probability. Also, $q(x, y)$ would never be zero because the proposal chain would never propose a target with zero probability.

Claim II.2.1. *The Metropolis-Hastings algorithm generates a reversible Markov chain with stationary distribution π that has density π_d*

Proof. We will check if condition (I.1) is satisfied.
Assume $x \neq y$, otherwise it's trivial.

$$\begin{aligned}
\pi(dx)P(x, dy) &= [\pi_d(x)dx][q(x, y)dy\alpha(x, y)] \\
&= \pi_d(x)q(x, y) \min \left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \right) dxdy \\
&= \min(\pi_d(x)q(x, y), \pi_d(y)q(y, x)dxdy \\
&= \pi(dy)P(y, dx)
\end{aligned}$$

□

Common Proposals The common classes of proposal densities are the following:

- **Independent Sampler** If $q(x,y) = g(y) \quad \forall x \in \chi$.
- **Symmetric** If $q(x,y) = q(y,x)$
- **Random Walk** If $q(x,y) = g(|y-x|)$ i.e. $N(x,\sigma^2)$

II.2.2 Ergodic Properties of Algorithm

In order for the Markov Chain Convergence Theorem to be satisfied, we must verify that the Markov chain is ϕ-irreducible and aperiodic. This depends on the correct pairing of proposal density with target density. Usually if the proposal chain is "nice" and is ϕ-irreducible on S, the simulation chain will be as well. The following claim is a straightforward consequence of the definition of *phi*-irreducibility.

Claim II.2.2. *π-irreducibility of MH Algorithm*
The MH simulation chain is π-irreducible, where π is the target measure, if

$$\pi_d(y) > 0 \textit{ implies } q(x,y) > 0 \quad \forall x \in \chi$$

where π_d and q are the densities.

Claim II.2.3. *Aperiodicity of MH Algorithm*
The MH simulation chain is aperiodic if π_d and q are continuous and positive for all $x, y \in \chi$

Proof. Here P is the transition kernel of the simulation chain. Our goal is to prove that $P(x,A) \geq \frac{\epsilon}{d}\pi(A)$ for all x in a non-empty compact set C. The implication is that the chain is aperiodic by a simple contradiction that's left to the reader.
We know that χ includes compact subsets because χ is just Euclidean space in the MH algorithm. The positivity of π_d implies $\pi(C) > 0$.
The key to the rest of the proof is to choose $B \subset C$ and a set $R_x(B)$ for a fixed x where

$$R_x(B) = \left\{ y \in B : \frac{\pi_d(y)q(y,x)}{\pi_d(x)q(x,y)} < 1 \right\}$$

Set $\epsilon = \inf_{x,y \in C} q(x,y)$ and $d = \sup_{x \in C} \pi_d(x)$, then calculate $P(x,A)$ for any $A \subset \mathscr{H}$ by splitting the integral into $R_x(B)$ and A $R_x(B)$. The full proof can be found in [18]. □

Remark. These conditions can be weakened is presented in [26].

Markov Chain Weak Law The Markov chain weak law gives an analogous version of the weak law of large numbers for Markov chains. This will allow us to use Monte Carlo estimates with Markov chains.

Theorem II.2.2. *Markov Chain Weak Law*
Given a ϕ-irreducible and aperiodic Markov chain with stationary distribution π and a functional f such that $\pi|f| < \infty$

$$\lim_{n\to\infty} \mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \pi(f)| > \epsilon) = 0$$

Proof. The proof is a consequence of the ergodic theorem from measure theory which gives limiting properties of a general class of measure-preserving transformations. The full proof is in Chapter 4 of [22]. □

II.2.3 Example

Example II.2.1. Here we demonstrate the simulation of $N(0,1)$ random variable using proposal density $N(x,5)$. This falls under the category of Random Walk algorithms. The two above claims are both satisfied thus the Markov chain should converge to stationarity. The code used for this example can be found in the Appendix. Note the well mixing of the simulation chain and slight deviation of the tails of simulation chain when plotted against the theoretical normal distribution. The functional evaluated in the expectation is $\log(1+|X|)$.
Refer to figures II.2.1 and II.2.2

II.3 Qualitative Convergence

This section will analyze the qualitative convergence property (and subsequently the CLT property) in the general case and of different major classes of algorithms.

II.3.1 General Case

Corollary II.3.1. *If support of the target density π_d is compact, simulation chain is uniformly ergodic.*

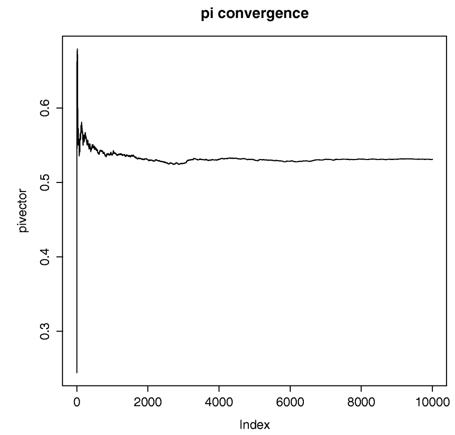


Figure II.2.1: Plot of empirical expectation - Proposal N(x,5), Target N(0,1)

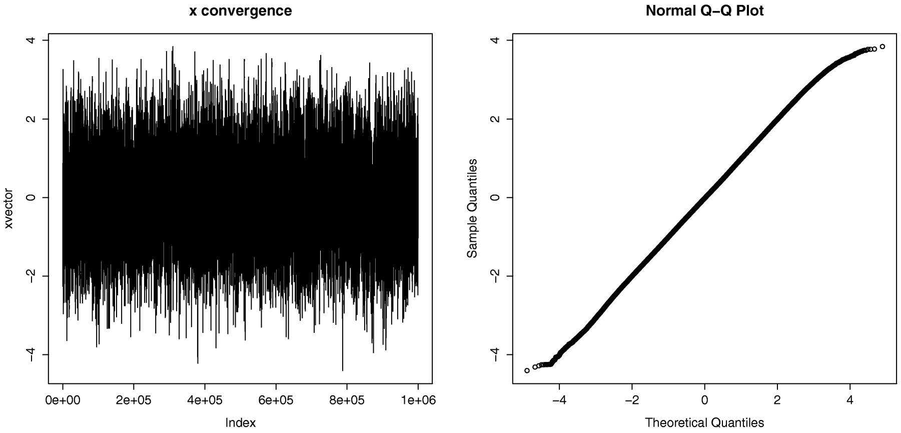


Figure II.2.2: Plots of simulation chain - Proposal N(x,5), Target N(0,1)

Proof. The state-space of the simulation chain, χ, is now compact. Following the proof of Claim II.2.3, we can take the compact set C to be χ and prove that χ is small. Therefore, by theorem I.2.17, the simulation chain is uniformly ergodic. □

Reminder that Theorem I.2.21 can be used to establish the sub-geometricity of the MH algorithm.

II.3.2 Independent Sampler

Recall that a MH algorithm is called an independent sampler if the proposal *density* is of the form $q(x, y) = g(y)$.

Theorem II.3.2. *Uniform Ergodicity of MH Independent Sampler*
The independent sampler is uniformly ergodic if there exists $\alpha > 0$ such that

$$\frac{q(x,y)}{\pi_d(y)} \geq \alpha \quad \forall x, y \in \chi$$

Proof. It is straightforward to check that the whole state-space is small. In fact, $P(x, A) \geq \alpha\pi(A) \quad A \in \mathscr{H}\ x \in \chi$. □

Corollary II.3.3. *If $\inf_{x\in\chi} \frac{q(x)}{\pi_d(x)} = 0$ π-a.e., then the independence sampler is not geometrically ergodic*

Proof. We can invoke theorem I.2.21. This is equivalent to proving that $\inf_{x\in\chi}(1 - P(x, \{x\}) = 0$.
Denote $A_x = \{y : \frac{q(x)}{\pi_d(x)}\pi_d(y) \leq q(y)\}$

$$\begin{aligned}
1 - P(x, \{x\}) &= \int_{y\in\chi\backslash\{x\}} q(x,y)\alpha(x,y)dy \\
&\leq \int_{y\in\chi} q(x,y)\alpha(x,y)dy \\
&= \int_{y\in\chi} \min\left(\frac{q(x)}{\pi_d(x)}\pi_d(y), q(y)\right) dy \\
&= \int_{y\in A_x} \frac{q(x)}{\pi_d(x)}\pi_d(y)dy + \int_{y\in\chi\backslash A_x} q(y)dy \\
&= \frac{q(x)}{\pi_d(x)} \int_{y\in A_x} \pi_d(y)dy + \int_{y\in\chi\backslash A_x} q(y)dy
\end{aligned}$$

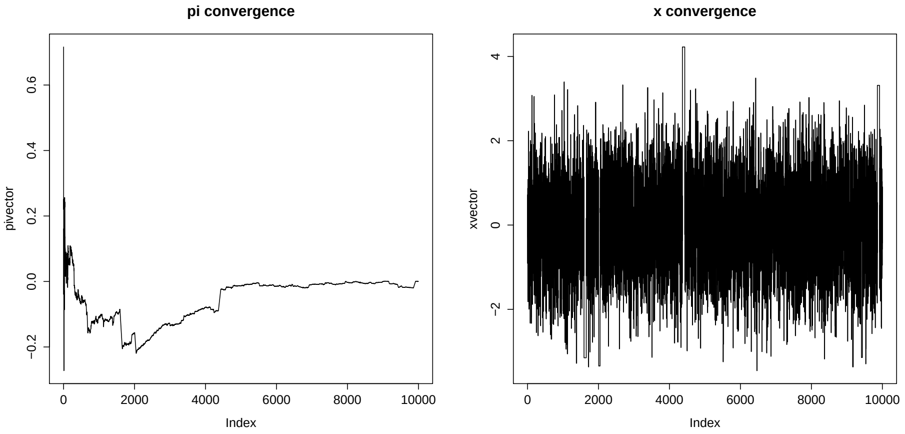


Figure II.3.3: Trace Plots - Proposal N(0,1) Target Laplace(1)

If $\inf_{x\in\chi} \frac{q(x)}{\pi_d(x)} = 0$ π-a.e., then we can make the set A_x as close to χ as possible since $q(y) > 0$ in χ. Now, $\int \pi_d(y)dy \leq 1$ so the infimum of the first integral is zero. The infimum of the second integral is also zero because $\int q(y)dy \leq 1$ and the domain of integration can be made arbitrarily close to $\emptyset$. $\square$

Example II.3.1. Proposal: N(0, 1). Target: Laplace(1). The target density is $\pi_d(x) = \frac{1}{2}e^{-|x|}$. Therefore, $\inf \frac{q(x)}{p(x)} = 0$. Here are the trace graphs of a 10000 simulation iterations (figure III.3.3) There doesn't seem to be much anomaly with the x trace, but the pi trace makes a few digressions away from 0 that was not present in the previous example. Although subgeometricity doesn't always mean the nonexistence of the CLT, in this case (and many others), it does (figure III.3.4).

Remark. There is a convergence quality that is "weaker" than geometric - polynomial. The CLT can still exist under polynomial ergodicty albeit stronger requirements on the functional. See [16].

Open Problem II.1. *What is an example of a subgeometric Markov chain that satisfies the CLT?*

II.3.3 Random Walk MH

Recall that the random walk MH algorithm is when the proposal density is of the form $q(x, y) = q(|x - y|)$. This is also a symmetric MH algorithm so

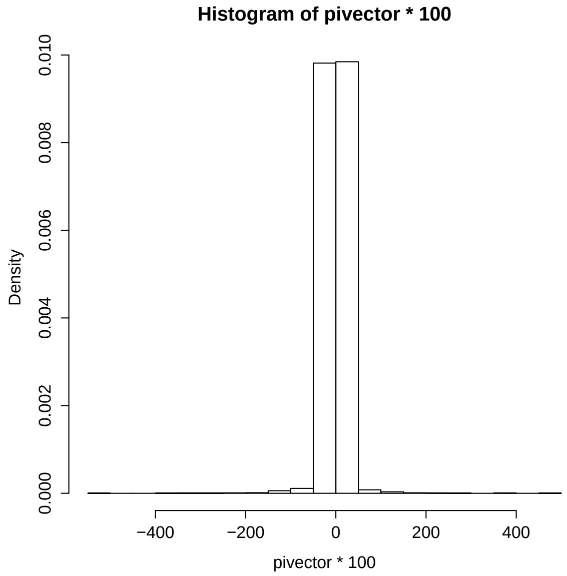


Figure II.3.4: Histogram of 10000 runs - Proposal N(0,1) Target Laplace(1)

all the theorems in the previous section applies here.
Unfortunately, most random walk MH algorithms are not uniformly ergodic.

Theorem II.3.4. *If the state-space of the random walk MH algorithm is $\mathbb{R}^k$, then it is not uniformly ergodic for any target distribution.*

Proof. Proof is found in theorem 3.1 of [18]. □

However, the following is a theorem that guarantees geometric ergodicity in one dimensional cases if the tails of the target density decreases exponentially.

Theorem II.3.5. *Geometric Ergodicity of Symmetric MH on $\mathbb{R}$*
Given a target $\pi_d(x)$ on $\mathbb{R}$ that satisfies the following

1. *Positive for all x*

2. *Continuous for all x*

3. *Symmetric*

4. *Log-concave in the tails:* $\exists \alpha > 0 \& x_0 > 0$ *s.t.* $\forall y \geq x \geq x_0, \log \pi_d(x) - \log \pi_d(y) \geq \alpha \times (y - x)$ *and* $\forall y \leq x \leq -x_0, \log \pi_d(x) - \log \pi_d(y) \geq \alpha \times (x - y)$

Then for any proposal distribution Q that is a random walk proposal with continuous density $q(x) > 0$, the algorithm is geometrically ergodic. Furthermore, if π_d is not symmetric, the same conclusion holds if there exists a finite constant b such that $q(x) \leq be^{-\alpha|x|}$.

Proof. The proof is a direct one that shows the function $V(x) = e^{s|x|}$ can satisfy the drift condition [18]. □

Remark. Analogous conditions with stronger curvature requirements exist in higher dimensional Euclidean spaces. These results can be found in [26].

The converse direction is almost true [18].

Theorem II.3.6. *Suppose that π_d is symmetric, $\theta(x) \triangleq (d/dx) \log(\pi_d(x))$ is defined for large x, and the its limit at infinity exists (possibly infinite). Suppose that the proposal density is a* continuous *random-walk density which is continuous and $\int_{z \in \mathbb{R}} |z| q(z) dz < \infty$. Then the MH algorithm is geometrically ergodic iff $\lim_{x \to \infty} \theta(x) < 0$.*

Example II.3.2. In this example, we use proposal densities $N(x, 1)$ with target distribution Student-t(4).

$$\theta(x) = \frac{d}{dx} K - (5/2) \log(1 + x^2/4) = -\frac{5/2 \cdot x/2}{1 + x^2/4}$$

So $\lim_{x \to \infty} \theta(x) = 0$ and since the Normal distribution has finite first moment so the simulation chain is not geometrically ergodic. The trace plots are given in figure II.3.5 and central limit theorem histogram in figure II.3.6.

II.4 Central Limit Theorems

The obvious way to verify the existence of CLTs is to use the theorems from II.2 to verify that a Metropolis-Hastings simulation chain is uniformly or geometrically ergodic and invoke the Markov chain CLT. However, It is not necessarily true that subgeometric Markov chains fail to satisfy the CLT. However, theorem I.2.24 can be adapted to MH algorithms.

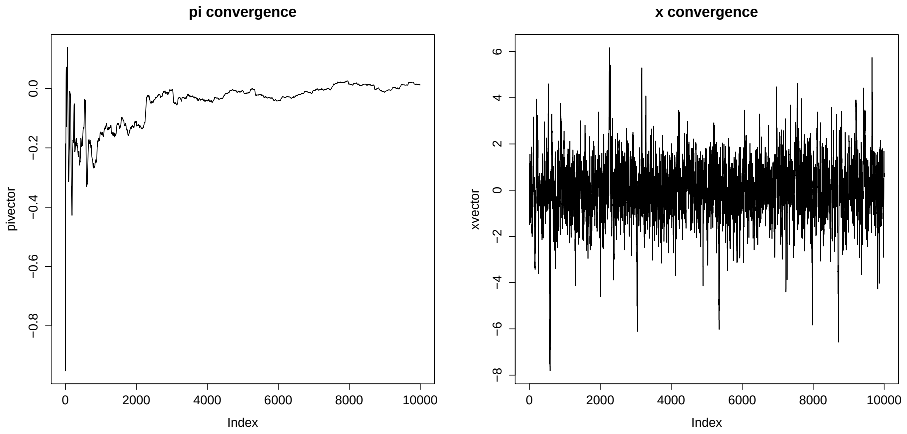


Figure II.3.5: Trace Plots - Proposal N(x,1) Target Student-t(4)

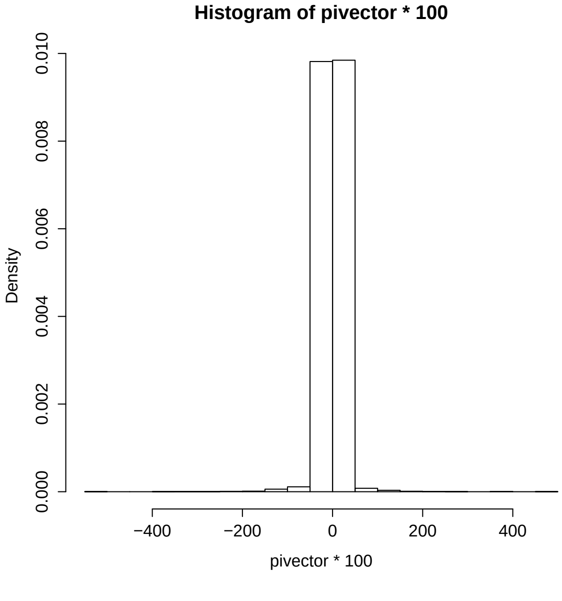


Figure II.3.6: Histogram of 10000 runs - Proposal N(x,1) Target Student=t(4)

Theorem II.4.1. *Nonexistence of CLT for MH algorithms*
Given a MH algorithm on $\mathbb{R}$ *such that the target has density* π_d with respect to the Lebesgue measure *and suppose that* $A(x) = 1 - P(x, \{x\})$ *is differentiable and for large* x

1. $A(x)$ *converges monotonically to 0*
2. $\liminf_{x\to\infty} \left|\frac{\pi_d(x)}{A'(x)}\right| = \infty$

Then the CLT does not hold for functions bounded away from zero at ∞*.*

Proof. The proof proceeds via a clever usage of the Markov's inequality and theorem I.2.24. Refer to [23]. □

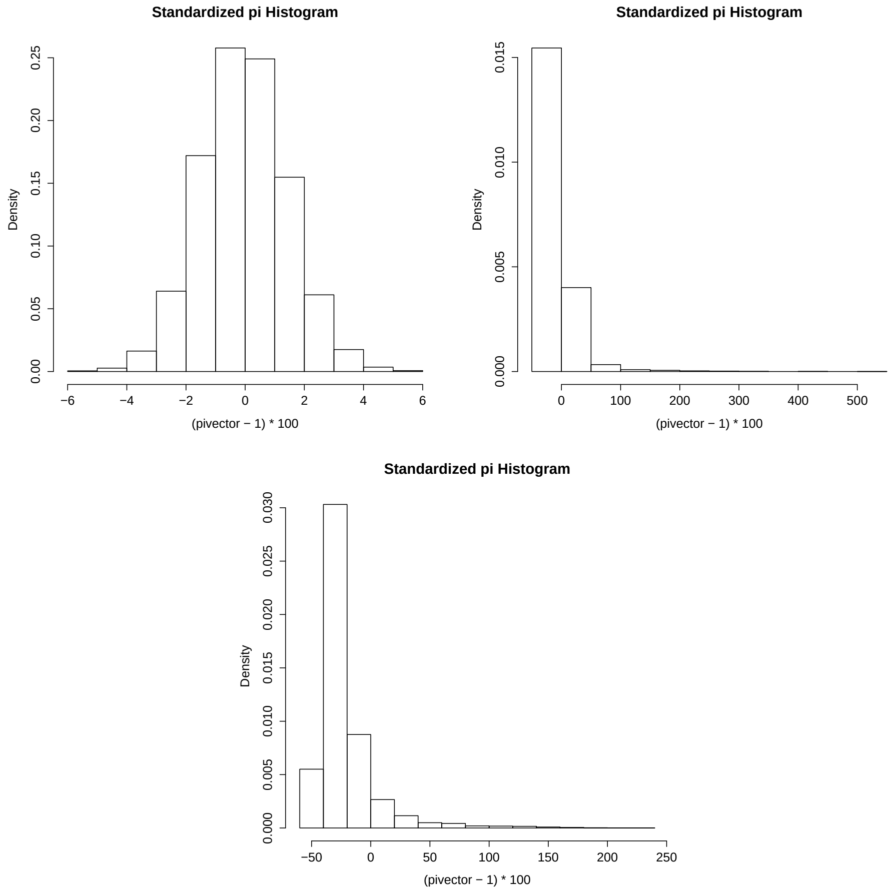
Example II.4.1. This example is based on example 1 from [23]. The target distribution is exponential(1) and the proposal distribution is exponential(θ). Direct calculation shows

$$A(x) = \theta \exp((1-\theta)x) - (\theta - 1)\exp(-\theta x)$$

$A(x)$ certainly converges monotonically to zero. Further calculation reveals

$$\left|\frac{\pi_d(x)}{A'(x)}\right| = (\theta(\theta-1)(\exp((2-\theta)x) - \exp((1-\theta)x)))^{-1}$$

which diverges for $\theta > 2$. Simulation was run for $\theta = 0.5, 3, 5$. Refer to figure (II.4.7).


Standardized pi Histogram

Standardized pi Histogram

Standardized pi Histogram

Figure II.4.7: Three Standardized pi Histograms - Proposal Exp(θ) Target Exp(1).
From left to right: $\theta = 0.5$ $\theta = 3$ $\theta = 5$

Chapter III

Adaptive MCMC

III.1 Theory

III.1.1 Description and Example

The difficulty in applying the above MH algorithms is that it is tedious to find an optimal proposal density such that the simulation chain converges the quickest. Take the following example

Example III.1.1. Let $\chi = \{1, 2, 3, 4, 5\}$ and $\pi(\{2\}) = \epsilon, \pi(\{i\}) = \frac{1-\epsilon}{4}$ for$i = 1, 3, 4, 5$ for some small $\epsilon > 0$. This target distribution has a density with respect to the counting measure. If we choose the proposal distribution to be

$$Q(x, \{y\}) = \begin{cases} 1/3 & y \in \{x-1, x, x+1\} \\ 0 & \text{o.w.} \end{cases}$$

The MH chain is π-irreducible and aperiodic so it's ergodic. In fact, it is uniformly ergodic. However, simulation results show that the simulation chain tends to get stuck on either side of $\{2\}$.

The solution to the above example's problem can be as simple as enlarging the support of the proposal distribution. However, what if the target distribution is not so visualizable, as it is often the case in higher dimensions.

The ultimate goal is to have a MCMC algorithm that can adapt to the target distribution and requires minimal human intervention. To put it formally

Definition III.1.1. Adaptive MCMC
Given a state space χ and a target distribution π. Let $\mathscr{Y}$ be some kind of an index set (*adaptation index*) for a collection $\{P_\gamma\}_{\gamma\in\mathscr{Y}}$ of Markov chain kernels that are all ϕ-irreducible, aperiodic, and stationary for π. So in another words, each $P_{\gamma 0}$ is a ergodic MCMC algorithm. Let random variable Γ_n represents a kernel used when moving from X_n to X_{n+1}, set

$$\mathscr{G}_n = \sigma(X_0, \dots, X_n, \Gamma_0, \dots, \Gamma_n)$$

The Markov chain $\{X_n\}$ behaves according to

$$\mathbb{P}[X_{n+1} \in B | X_n = x, \Gamma_n = \gamma_0, \mathscr{G}_{n-1}] = P_{\gamma_0}(x, B), \quad \forall x \in \chi, \gamma_0 \in \mathscr{Y}, B \in \mathscr{H}$$

Let's denote the overall simulation chain by $\mathbb{P}[X_n \in B | X_0 = x, \Gamma_0 = \gamma] = A^{(n)}((x, \gamma), B)$. This stochastic process is no longer necessarily a Markov chain as Γ_n can depend on the entire history of $\{X_n\}$.

There are several special cases of adaptive MCMC.

- **Independent Adaptations**: If for all n, Γ_n is independent of X_n
- **Finite Adaptations**: If there exists a stopping time τ that is finite with probability 1 such that $T_n = T_\tau \; \forall n \geq \tau$
- **Markovian Adaptations**: Define a stochastic process from Ω to $\chi \times \mathscr{Y}$ by the pair (X_n, Γ_n). If this stochastic process is Markovian, then the simulation chain $\{X_n\}$ is called Markovian adaptations.

III.1.2 Ergodic Properties

The central ergodic theorem for adaptive MCMC is from [24].

Theorem III.1.2. *Adaptive MCMC Ergodic Theorem*
Given an adaptive MCMC algorithm on χ with adaptation index $\mathscr{Y}$ such that π is stationary for each transition kernel and:

- $\forall \epsilon > 0, \exists N_\epsilon \in \mathbb{N} \; s.t. \|P_\gamma^N(x, \cdot) - \pi(\cdot)\| \leq \epsilon \quad \forall x \in \chi \,\&\, \gamma \in \mathscr{Y}$
- $\mathbb{P}(|\lim_{n\to\infty} \sup_{x\in\chi} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\| \,| > \epsilon) \to 0 \quad \forall \epsilon > 0$

Then the simulation chain is ergodic.

The proof, which uses a version of the coupling construction, can be found in [24]. There similar (usually older) versions of this ergodic theorems but they usually require that Γ_n converges to a constant with probability 1. You will see in an example below that Γ_n often oscillates *ad infinitum* [12].

There are two immediate cases where the above conditions are satisfied. If χ and $\mathscr{Y}$ are both finite, then the first condition is satisfied, and if the probability of adaptation diminishes to zero, the second condition is satisfied. A third corollary exists and deserves its own space.

Corollary III.1.3. *Ergodicity of adaptive Metropolis-Hastings algorithms Suppose an adaptive MH algorithm satisfies condition two of the above theorem with each transition kernel being ergodic for π. That is, P_γ is a MH transition kernel for each fixed γ. Also, suppose that each proposal kernel have density $f_\gamma(x, y)$ with respect to a finite reference measure λ and $\frac{d\pi}{d\lambda} = g$ exists. Finally, assume that $\{f_\gamma(x, y)\}_{\gamma \in \mathscr{Y}}$ are uniformly bounded and the mapping $(x, \gamma) \to f_\lambda(x, z)$ is continuous for each fixed y on a produce metric space topology where $\chi \times \mathscr{Y}$ is compact. Then the MCMC algorithm is ergodic*

Relaxing Condition One The first condition of theorem III.1.2, which requires that the convergence rate of all transition kernels to be strictly bounded, can be relaxed such that the convergence rate to be only bounded in probability.

Theorem III.1.4. *Define the "ϵ convergence time function".*

$$M_\epsilon(x, \gamma) = \inf_{n \in \mathbb{N}} \{n : \|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \leq \epsilon\}$$

Instead of requiring $M_\epsilon(x, \gamma) \leq N_\epsilon \quad \forall x \in \chi$ & $y \in \mathscr{Y}$ for each $\epsilon > 0$, we require that $\forall \delta > 0, \exists N \in \mathbb{N}$ such that $\mathbb{P}[M_\epsilon(X_n, \Gamma_n) \leq N | X_0 = x^*, \Gamma_0 = \gamma^*] \geq 1 - \delta \; \forall n \in \mathbb{N}$, then the algorithm is ergodic when started at $(X_0 = x^*, \Gamma_0 = \gamma^*)$.

III.1.3 Limiting Theorems

A nice weak law of large numbers for adaptive MCMC is of great practical applicability.

Theorem III.1.5. *Adaptive MCMC Weak Law Suppose that the conditions of theorem III.1.2 holds. Let g: $\chi \to \mathbb{R}$ be a* bounded *measurable function. Then the following holds for any starting value of x and γ*

$$\frac{\sum_{i=1}^{n} g(X_i)}{n} \to \pi(g)$$

in probability.

The proof builds upon the regular Markov chain weak law and uses a coupling construction. See section 9 of [24].

Remark. The strong law does not hold for Adaptive MCMC. The counterexample can be found in [24].

III.1.4 Qualitative Convergence

Theorem III.1.6. *Finite Space Uniform Ergodicity*
For any *adaptive algorithm, if it is ergodic for all starting values and χ & $\mathscr{Y}$ ae finite, then the algorithm is uniformly ergodic or namely set $A^n((x,\gamma),B) = \mathbb{P}[X_n \in B | X_0 = x, \Gamma_0, \gamma]$, we have*

$$\sup_{x\in\chi,\gamma\in\mathscr{Y}} \|A^n((x,\gamma),\cdot) - \pi(\cdot)\| \leq M\rho^n$$

for some $0 < \rho < 1$.

Proof. The proof is obvious because the sup becomes merely a max in finite space, and if the algorithm is ergodic that means the chain will eventually be less than 1/2 away from π in the above norm for each starting value. By Proposition 3e of [25] the chain will be geometrically ergodic for each starting value. Taking the max will give the appropriate M. □

III.1.5 Central Limit Theorem

In this section we focus on the Markovian adaptation algorithms where Markov chain theory can be applied to (X_n, Γ_n).

Theorem III.1.7. *Given an Markovian adaptation algorithm $(X_n\Gamma_n)$. If this Markov chain is ϕ-irreducible and aperiodic on $\chi \times \mathscr{Y}$ with some stationary distribution λ where $\lambda(\cdot,\mathscr{Y}) = \pi(\cdot)$, and if the adaptive algorithm is uniformly ergodic and $\mathbb{E}_\pi[f^2] < \infty$, then the CLT is satisfied*

$$\sqrt{n}(\overline{f(X_n)} - \mathbb{E}_\pi(f)) \to N(0,\sigma^2)$$

where

$$\sigma^2 = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^{n}(f(X_i) - \mathbb{E}_\pi(f))\right)^2\right]$$

for every initial starting point (x, γ).

Proof. This is just an application of the standard Markov Chain CLT to the process $(X_n\Gamma_n)$ with the functional $g(x, \gamma) = f(x)$. □

Example III.1.2. In this example, the target is exponential(1) and the proposal chain is an independent sampler with proposal density exponential(θ). Example II.4.1 showed how the CLT would fail in the non-adaptive case if $\theta > 2$. On the other hand, theorem II.3.2 tells us that the non-adaptive sampler would be uniformly ergodic (thus satisfying CLT) if

$$\frac{q(x, y)}{\pi_d(y)} = \theta \exp((1 - \theta)y) \geq \alpha$$

which would be the case if $0 < \theta < 1$.

Four sets of adaptive simulations were run 10000 times to check for potential normal distribution. Each simulation had 10000 iterations. Of those 10000 iterations, the θ parameter is either increased if acceptance rate is too high and decreased if acceptance rate is too low. This way the algorithm tunes itself for the optimal acceptance rate (in this case set to be 0.45). This adaptation takes place every 100 iterations and the acceptance rate counter is also reset at that time. The table in figure III.1.1 presents the histograms resulting from each simulation.

Analysis It seems that as long as θ remains in a range where the CLT for the non-adaptive chain is satisfied, the overall adaptive chain will also exhibit a CLT-like empirical behaviour. This raises an interesting open problem, and this example seems to support an affirmative answer.

Open Problem III.1. *Given an adaptive MCMC algorithm with transition kernel set* $\{P_\gamma\}_{\gamma\in\mathscr{Y}}$*, if all the transition kernels are ergodic for* π *and satisfy the* $\sqrt{n}$*-CLT, then does the adaptive MCMC algorithm satisfy the* $\sqrt{n}$*-CLT?*

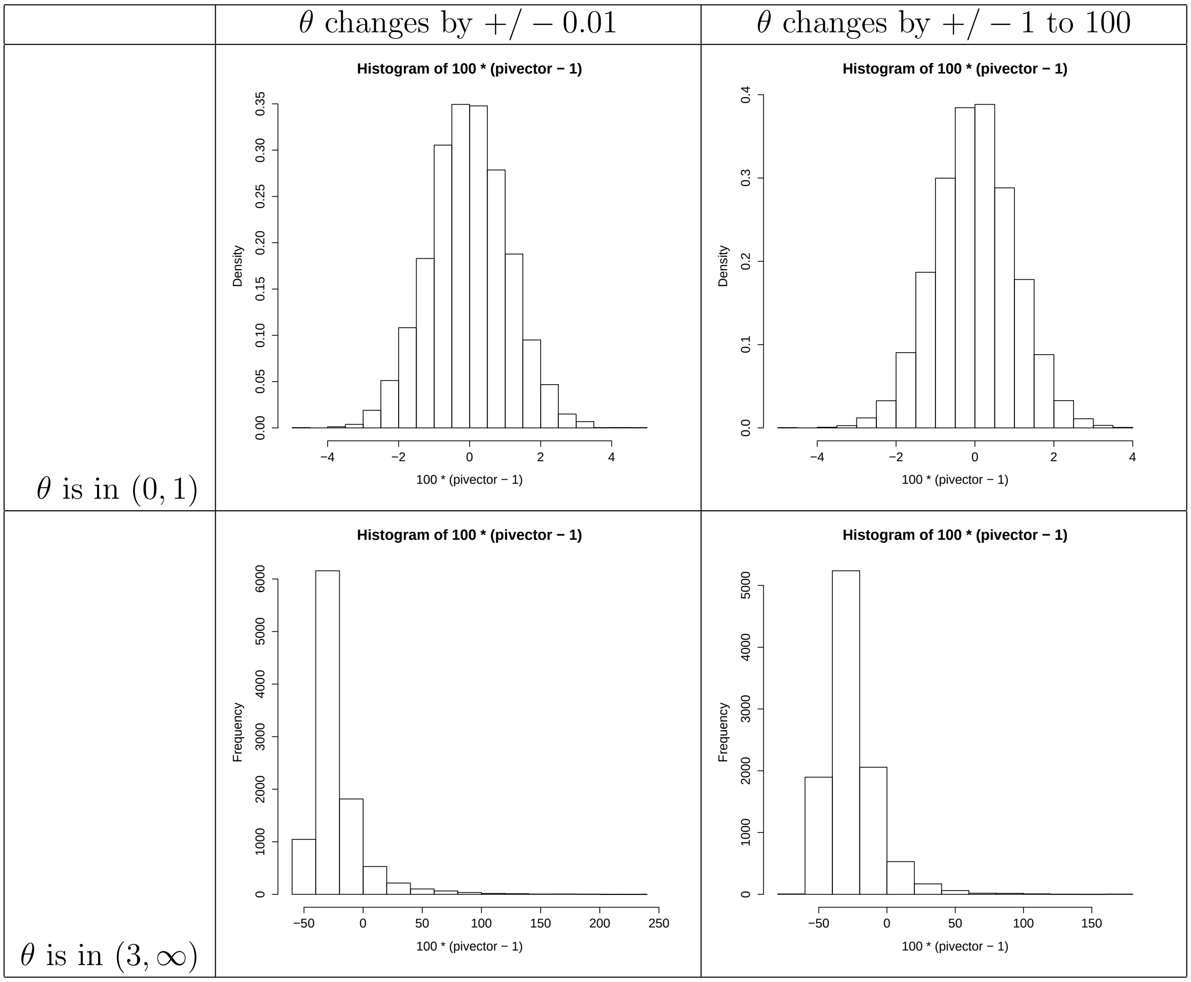


Figure III.1.1: Four Standardized pi Histograms - Adaptive Proposal Exp(θ) Target Exp(1).

III.1.6 Discussion

The conditions of the last theorem is rather hard to verify. There needs to be simpler conditions to verify the stationary of (X_n, Γ_n).

Open Problem III.2. *Does there exist a function V such that the Markovian adaptations chain (X_n, Γ_n) satisfies the drift condition?*

The issue of whether an algorithm satisfies the CLT must be addressed with greater attention. In previous simulation examples where CLT isn't satisfied, one would often find that the empirical expectation estimates from individual simulations are far off target. This defeats the practical purposes of such MCMC algorithms even if the Markov chain is ergodic in theory.

III.2 Simulation Findings

This section is a record of different computer simulations ran.

III.2.1 Simulation Setup

The simulation was programmed in C and built using the GNU C Compiler (gcc) with the math library. The computer specification of the simulators were as follows:

1. SunBlade 1000 with two 900 MHz UltraSPARC III (Cu) processors and 4 GB of memory running Solaris.
2. A RedHat Linux system with four dual core AMD Opteron processors.
3. Windows 2000 Compaq Presario laptop with P3 700 MHz processor and 384 MB of memory.

The output of the program was piped into files which were analyzed subsequently in R.

III.2.2 Location-based Exponentially Adapting Random Walk MH Algorithm

Setup

In this subsection, the adaptive MCMC algorithm operates uner the following framework, referred to as original adaptation scheme. Variations would be made to this framework in each of the following sections.

- Proposal density is $N(x, \sigma_x)$
- Target distribution is $N(0, 1)$ and target expectation is $\log(1 + |X|)$
- $\sigma_x^2 = e^a(1 + |x|)^b$
- a increases/decreases by $1/\{1 \ldots 100\}$ if the acceptance rate is too high/low (respectively, compared to the target acceptance rate)
- b increases/decreases by $1/\{1 \ldots 100\}$ if the acceptance rate when the empirical expectation estimate is increased is bigger/smaller than the acceptance rate when the empirical expectation estimate is decreased
- b is bounded by 10000 (This hardly matters because b never goes off to infinity)
- Everyone 100 iterations adaptation takes place and the acceptance rate estimator is cleared (these are called adaptation cycles or **batch sizes**)

Basic Result

The result of the simulation can be summarized by these statistics and numbers. Batch sizes are 100. Total number of iterations was 1 million. Note that the parameters change during each adaptation by 0.01 (not a decreasing sequence). Refer to figure III.2.2.

Relation between target acceptance rate, a, and b

Under the same simulation setup as described above in ”Setup”, the target acceptance rate is varied and the stable average of a & b are calculated. Refer to figure III.2.3. It’s not clear why the relation between target acceptance rate and b parameter is second order; perhaps it is due to the fact b is a “location-sensitive” parameter while a is not.

Target Acceptance Rate	Stable Average of a	Stable Average of b
0.1	3.44	2.1
0.2	2.13	2.01
0.3	1.4	1.86
0.455	0.699	1.513
0.6	0.4	0.6

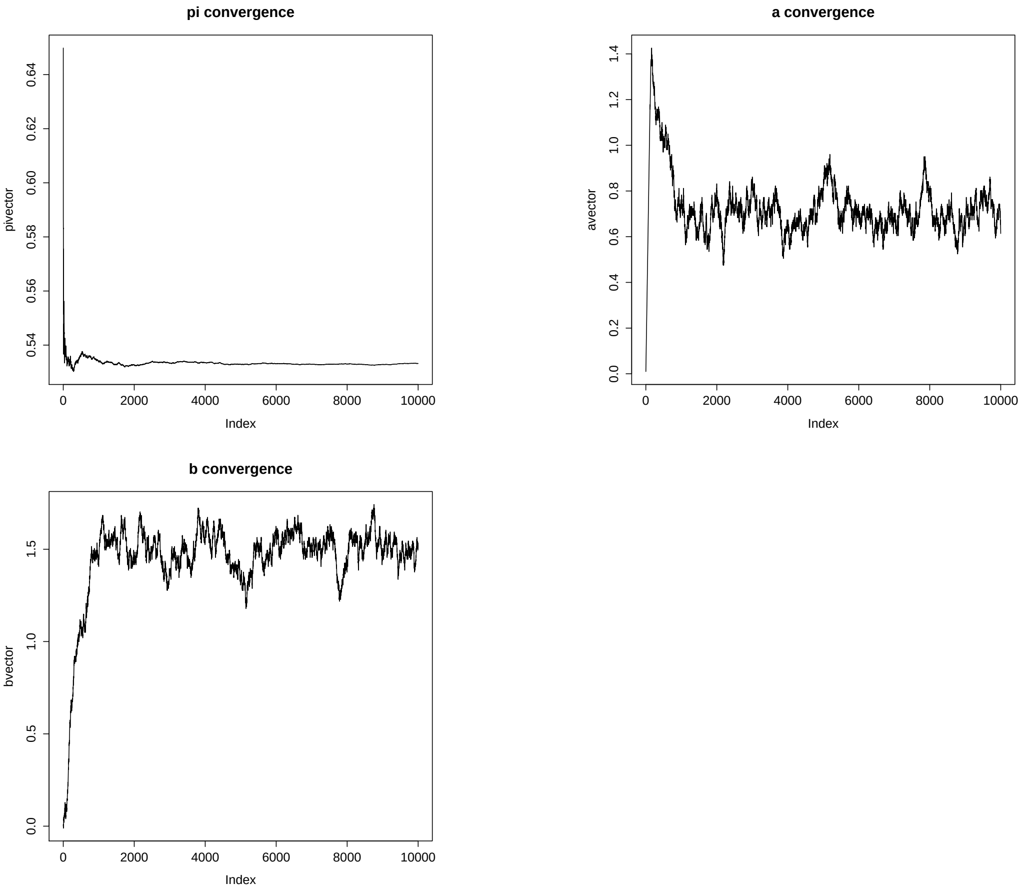


Figure III.2.2: Basic Simulation Result

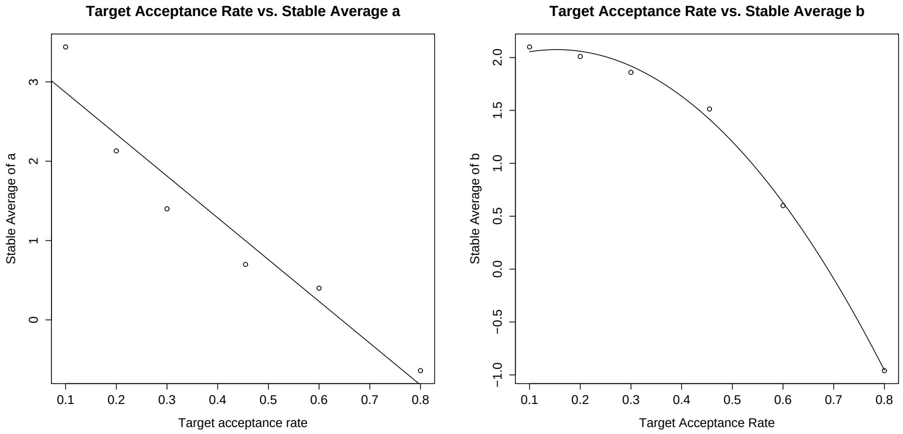


Figure III.2.3: Relation between a, b and target acceptance rate
The lines are linear regression lines under the 1st and 2nd order models respectively

Constant factors in the variance function

We now investigates whether including a constant factor in the variance function affects the stable average of a and b in a predictable way.

Let the new variance function be

$$\sigma'_x = e^a \left(\frac{1 + |x|}{C} \right)^b$$

In this simulation we chose C to be the (approximately) true value of $\mathbb{E}(\log(1+|X|))$ $X \sim N(0,1)$.

Compare the result with the original simulation.

Scheme	Int. ACT	Avg. Sq. Distance	Stable Average of a	Stable Average of b
ORIGINAL	˜2.35	0.775	0.699	1.513
NEW	2.35	0.781	-0.27	1.49

Analysis, let the a and b parameter under the new variance function be denoted α, β. We can transform the new variance function into the original form by the following

$$e^{\alpha} \left(\frac{1 + |x|}{C} \right)^{\beta} = e^{\alpha - \beta \log(C)} (1 + |x|)^{\beta}$$

So heuristically the original a should be equal to $\alpha - \beta \log(C)$, and the original b and β should be the same. This checks out with the empirical observations.

Effects of batch size

Holding total iterations constant Recall that in the original setup, batch size was 100. Note that the total number of iterations is kept constant at 1 million. The obvious result is that a batch size too small can create unstable acceptance rate and large fluctuates in a and b. On the other hand, a large batch size, under the same number of total iterations, would make it very tenuous for a and b to "converges" to the stable values. Refer to figure III.2.4.

Scaling total iterations By varying the total number of iterations with the batch size, we can allow the adaptations to stablize.

Batch Size	Total Iterations	a behaviour	b behaviour	acc. rate	avg. sq. dist.
5000	20000	rises to 1.2 and falls to 0.8 slowly	slowly rises to 1.5 and stays	mean drops from 0.7 to 0.45. fluctuation magnitude is around 0.02	inaccessible due to large size
1000	5000	similar to above	similar to above	similar to above	starts at 10 and stabilizes around 35
500	10000	rises to 1.2 and falls rather quickly to 0.8	rises quickly to 1.5	falls quickly to 0.45	rises quickly to 35
100	20000	rises to 1.2, falls even more quickly to 0.8. Larger fluctuation around 0.8	rises even more quickly to 1.5 and fluctuates	falls extremely quickly to 0.45 large fluctuations	rises extremely quickly to 35 large fluctuations

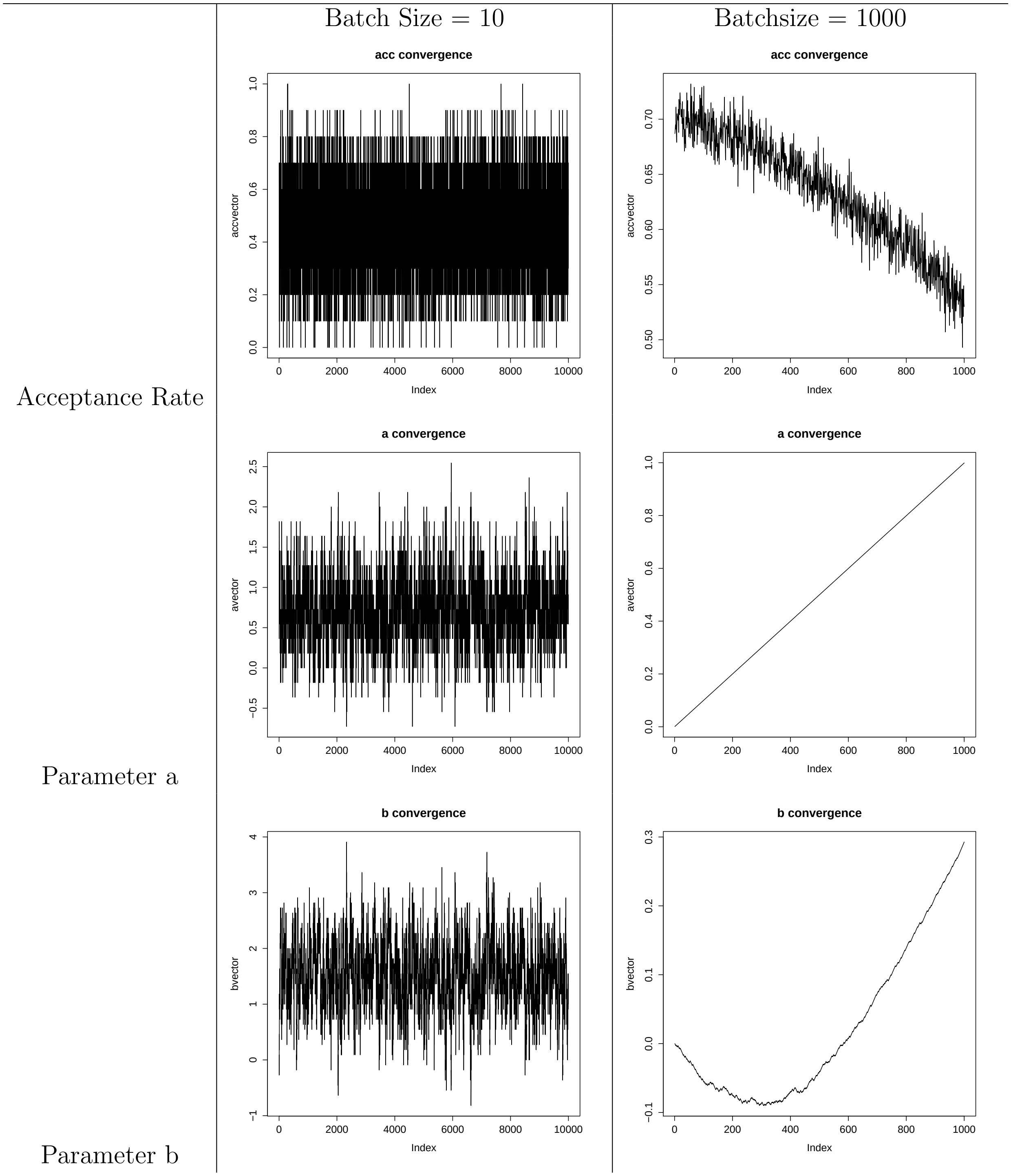


Figure III.2.4: Effects of batch size

Diminishing Adaptations

In this section, the original adaptation scheme is made to satisfy the Diminishing Adaptations property (2nd condition in theorem III.1.2. Specifically, a and/or b changes by a diminishing sequence $\{1/1, 1/2, \ldots, 1/10000\}$. The results are nearly identical to the original adaptation scheme. Here "fluctuates" denotes a parameter to oscillates around a central value that's listed and "converges" means it tends to the listed value.

Scheme	Avg.Acc.Rate	Int.ACT	Avg.Sq.Dist.	Stable avg. of a	Stable avg. of b
Original	0.455	˜2.35	0.775	(fluctuates) 0.699	(fluctuates) 1.513
a & b dim.	0.454	2.53	0.775	(converges) ˜0.68	(converges) ˜1.51

III.2.3 Non-location based exponentially adapting MH algorithm

In this section, we simplify the previous location based adaptive algorithm into a simpler one. σ_x^2 is now just e^a where a adapts. This way, we can study the necessity of diminishing adaptation and maximum bounds. The result from the simulation is as follows.

a changes by	upper bound of a	result
+/- 0.01	10	a fluctuates around 1.6
+/- 0.01	10000	a fluctuates around 1.6
+/- $\{1/1, 1/2, \ldots, 1/10000\}$	10	a converges to 1.68
+/- $\{1/1, 1/2, \ldots, 1/10000\}$	10000	a converges to 1.68

Analysis In all cases, the empirical expectation converges to the correct value at approximately the same speed. So it seems like the only affect diminishing adaptation has is on the asymptotic behaviour of the hyperparameter a. The upper bound on a has no affect if a does not achieve the upper bound during simulation. If a does achieve the upper bound, it may fail to converge to its optimal value as witnessed in other cases. **Therefore, the upper-bound on parameters is really a theoretical necessity and should not be imposed in programming unless if one parameter goes off to infinity.**

III.2.4 Relation between MCMC behaviour and expectation functional

The setup in this section is the same as the original adaptation scheme except for the target expectation and σ_x^2. The expectations are picked because $\mathbb{E}[X^2] = \mathbb{E}[Y]$ where X is standard normal and Y is $\chi^2(1)$.

Target	Adaptation	a behaviour	b behaviour
$X^2, X \sim N(0,1)$	Original Scheme	fluctuates around 0.64	fluctuates around 1.6
$X, X \sim \chi^2(1)$	Original Scheme	fluctuates around -2.7	fluctuates around 3.6
$X^2, X \sim N(0,1)$	non-location based adaptation	fluctuates around -2.7	NA
$X \sim \chi^2(1)$	non-location based adaptation	fluctuates around 1.6	NA

As you can see, the parameters behave differently for each target expectation/adaptation combination even though the expectations are really the same. This is because that the underlying target densities are different, thus the acceptance rates under the same proposal distribution are different.

III.2.5 Region-based Adaptation

A general rule regarding the efficiency of MCMC is that convergence is quicker if the proposal density closely resembles the target density. This is highly unlikely in the cases of complex target density but one way to reach this goal is to use region-based proposal densities.

Example III.2.1. Under the same setup as the Original Adaptation Scheme, we define a new variance function for the proposal density $N(x, \sigma_x^2)$

$$\sigma_x^2 = \begin{cases} e^a & |x| \leq K \\ e^b & |x| > K \end{cases}$$

a would be adapted according to the usual rules if the simulation chain remained mostly in the $|x| \leq K$ region during the batch and b would be adapted according to the usual rules if the simulation chain traversed mostly the complementary region. The total iterations is 5 million and batch size is 100.

There's no major effect on convergence rate by changing the boundary (K). The only thing that seems to be affected is the rate at which a & b reach stability. Refer to figure

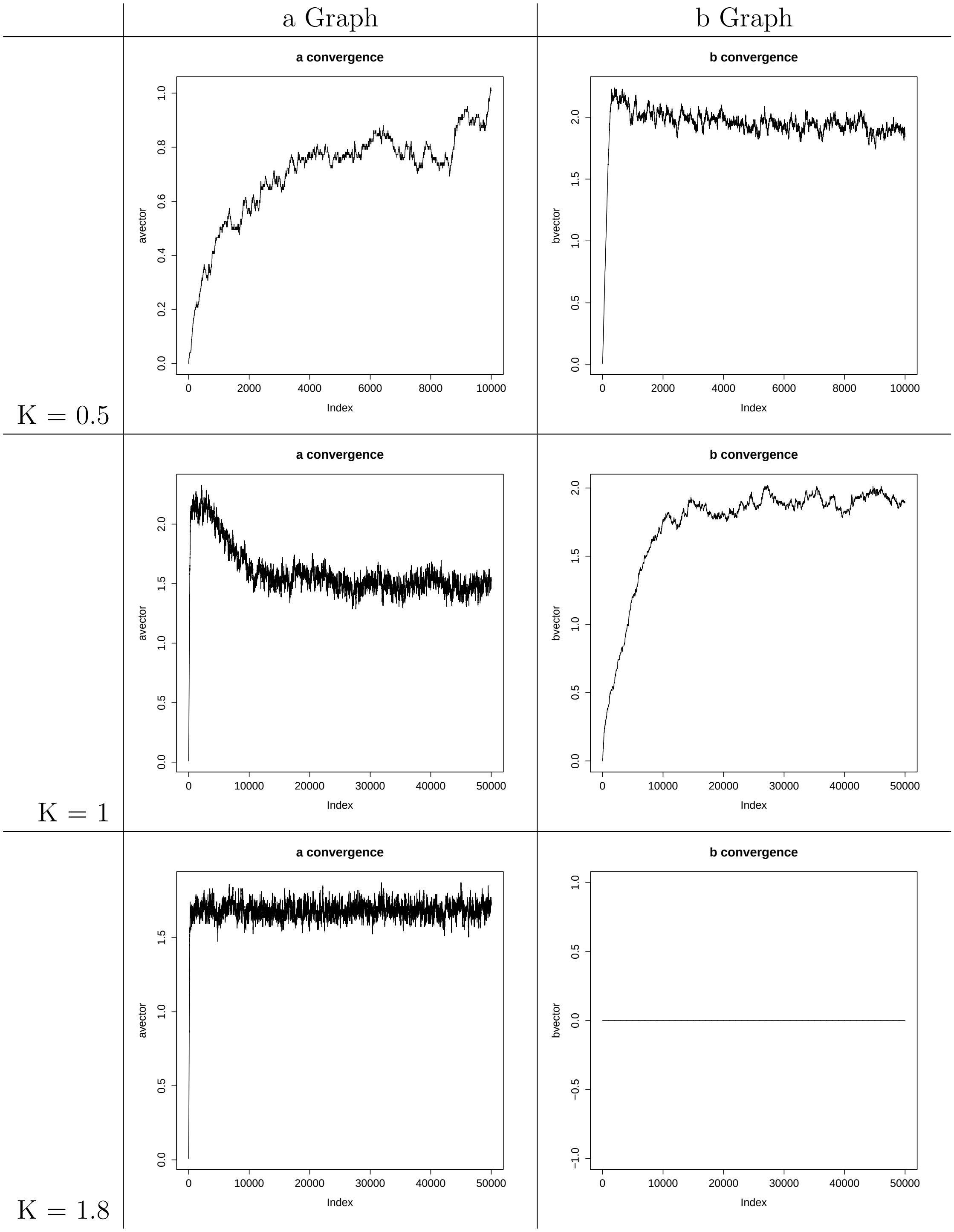


Figure III.2.5: Region-based adaptation

III.2.6 Densities with singularities

In this section, we investigate densities which have an area of low density or zero density (i.e. the support of π_d is disconnected).

Example III.1.1. is one such density which is not ergodic for the adaptive random-walk MCMC algorithm. In this section, we provide a density that has a Lebesgue density.

Example III.2.2. Define density π_d

$$\pi_d(x) = \begin{cases} \exp(-\frac{1}{x^2}) & x \in [-1.76, 0) \cup (0, 1.76] \\ 0 & \text{o.w.} \end{cases}$$

The function has a low density region around 0. Let this be the target of our MCMC algorithm. Three non-adaptive simulation were completed with N(x,σ^2) proposals and one adaptive simulations were done. In addition, the starting point for each simulation was customized. The resulting Markov chain sample paths ($\{X_n\}$) are plotted. Refer to figure III.2.5. It is evident that a proposal with too narrow a spread would be stuck on only one side of the density even though that it is ergodic in the mathematical sense. The adaptive MCMC algorithm efficiently solved the problem.

III.2.7 Central Limit Theorem

Examples II.4.1 and III.1.2 seems to support that some form of CLT exists for adaptive MCMC. Below are the list of other related schemes that have shown a CLT like behaviour empirically. This table below states the *alterations* done to the original location-based exponentially adapting random-walk MH algorithm.

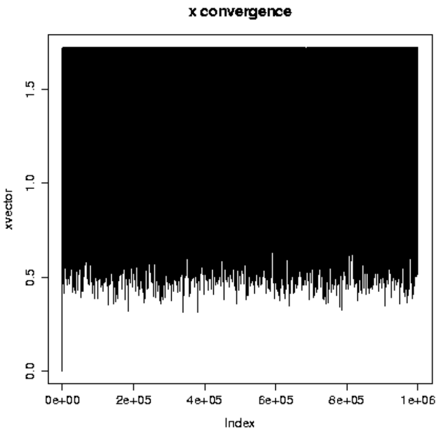
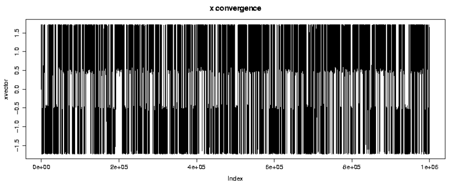
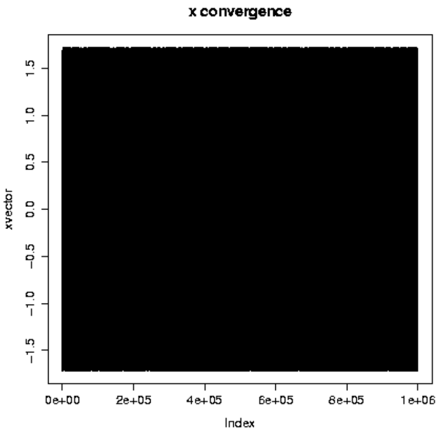
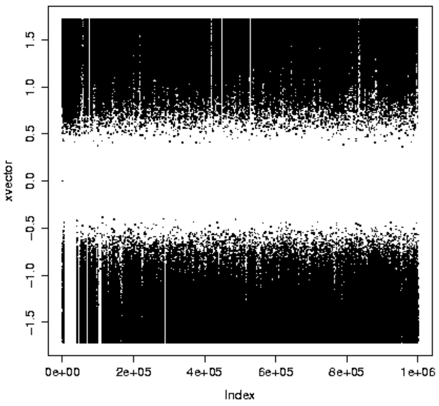
Scheme	Graph
$\sigma^2 = 0.1,\ x_0 = 1$	
$\sigma^2 = 0.5,\ x_0 = 1$	
$\sigma^2 = 1,\ x_0 = 1$	
Original Adaptation Scheme	

Figure III.2.6: Density with singularity

Target	Adaptation Method	Running Time
$\mathbb{E}(X)$ where X is N(0, 1)	$\sigma_x^2 = e^a$ where a changes by 1/100	batch size is 100, total iterations is 10000, 10000 simulations were repeated
$\mathbb{E}(X)$ where X is N(0, 1)	$\sigma_x^2 = e^1.7$ where 1.7 is the stable average from the adaptive case	50000 iterations per chain 5000 chains simulated
$\mathbb{E}(X)$ where X is N(0, 1)	$\sigma_x^2 = e^a$ where a changes by 1/100	batch size is 100, total iterations is 50000, 5000 chains simulated
$\mathbb{E}(X)$ where X is N(0, 1)	Original Scheme	batch size 100, total iterations 10000, 10000 chains simulated
$\mathbb{E}(X)$ where X is N(0, 1)	a and b adapts by $\{1/1, 1/2, \ldots, 1/100\}$	batch size is 100, total iterations is 10000, 10000 chains simulated
$\mathbb{E}(X)$ where X is Exp(1)	a and b adapts by 1/100	batch size is 100, total iterations is 10000, 10000 chains simulated
$\mathbb{E}(X)$ where X is Exp(3)	a and b adapts by 1/100	batch size is 100, total iterations is 10000, 1000 chains simulated

Appendix A

Notes on Simulation Codes

The simulation directory structure rooted at ....adaptlog/ has subdirectories organized in such a way that each subdirectory corresponds to one scheme and each sub-subdirectory will be some minor variation on one scheme. The scheme list is stored in the Finding Summary.doc file.

Below is the verbatim code from the original adaptation scheme:

```
/* Original Adaptation Scheme
==========================================================================

ADAPTLOG.C -- a program for doing one-dim "log(1+|x|)" adaptive MCMC

Copyright (c) 2004 by Jeffrey S. Rosenthal (jeff@math.toronto.edu).

Available from  http://probability.ca/jeff/comp/

Licensed for general copying, distribution and modification
according to the GNU General Public License
(http://www.gnu.org/copyleft/gpl.html).

----------------------------------------------------

Tab size is 3 spaces
----------------------------------------------------
Save as "adaptlog.c".
```

```
Compile with "cc adaptlog.c -lm", then run with "a.out".

Upon completion, can run 'source("adaptlogx")' in R to see a trace
plot.

Normal case: Can check pilogest in Mathematica (0.534822) with
command: NIntegrate[ (2*Pi)^(-0.5) * E^(-x*x/2) * Log[1+Abs[x]],
{x,-Infinity,Infinity}] Cauchy case: Can check pilogest in
Mathematica (0.929695) with command: NIntegrate[ 1/(1+x*x) *
Log[1+Abs[x]], {x,-Infinity,Infinity}] / NIntegrate[ 1/(1+x*x),
{x,-Infinity,Infinity}]

==========================================================================

*/


#include <stdio.h> #include <math.h> #include <sys/time.h>

#define PI 3.14159265

#define ADAPTLENGTH 100

#define PRINTLENGTH 1

#define NUMITS 10000

/* Target acceptance rate */ #define TARGACCEPT 0.45

#define XFILE "adaptlogx" #define AFILE "adaptloga" #define BFILE
"adaptlogb" #define PIFILE "adaptlogpi" #define ACCFILE
"adaptlogacc" #define MSQFILE "adaptlogmsq" /* #define CAUCHYTARG
true */

#define MAXABSB 10000.0

double drand48();
```

```
main()

{

   int i,j,k,t, adaptcount, printcount, xspacing;
    /* Global Iteration Count: numit */
   int numit, numaccept, accepted;
   int posaccept, negaccept;
    /* Simulation chain: x, Proposal chain: y*/
   double x, logsigma, sigma, a, b, y, A, logalpha;
   double possum, negsum, pilogest, pilogsum, logval;
   double absval(), normal(), targlogdens();
    /* int m1count; */
   FILE *fpx, *fpa, *fpb, *fppi, *fpacc, *fpmsq;

   /* INITIALISATIONS. */
   seedrand();
   numit = 0;
   x = 1.0;
   a = b = 0.0;
   logval = pilogest = pilogsum = 0.0;
   printf("\nBeginning \"pi(1+log|x|)\" adaption run.\n");
   printf("\nAdapting every %d iterations.\n", ADAPTLENGTH);
   printf("\nPrinting every %d iterations.\n", ADAPTLENGTH*PRINTLENGTH);
   printf("\n");
   if ((fpx = fopen(XFILE,"w")) == NULL) {
       fprintf(stderr, "Unable to write to file %s.\n", XFILE);
   }
   if ((fpa = fopen(AFILE,"w")) == NULL) {
       fprintf(stderr, "Unable to write to file %s.\n", AFILE);
   }
   if ((fpb = fopen(BFILE,"w")) == NULL) {
       fprintf(stderr, "Unable to write to file %s.\n", BFILE);
   }
    if ((fppi = fopen(PIFILE,"w")) == NULL) {
       fprintf(stderr, "Unable to write to file %s.\n", PIFILE);
   }
    if ((fpacc = fopen(ACCFILE, "w")) == NULL) {
```

```
    fprintf(stderr, "Unable to write to file %s.\n", ACCFILE);
}
 if ((fpmsq = fopen(MSQFILE,"w")) == NULL) {
    fprintf(stderr, "Unable to write to file %s.\n", MSQFILE);
}

xspacing = 1;

printf("\nxspacing = %d \n\n", xspacing);
fprintf(fpx, "\nxvector <- c(0.0");
fprintf(fpa, "\navector <- c(");
fprintf(fpb, "\nbvector <- c(");
 fprintf(fppi, "\npivector <- c(");
 fprintf(fpacc, "\naccvector <- c(");
 fprintf(fpmsq, "\nmsqvector <- c(0");

/* MAIN ITERATIVE LOOP. */
for (t=1; t<=NUMITS; t++)
 {
     /* THE PRINT CONTROL LOOP */
     for (printcount=0; printcount<PRINTLENGTH; printcount++)
     {
         /* Zero some counters. */
         numaccept = possum = negsum = posaccept = negaccept = 0;
         /* INDIVIDUAL ADAPTATION CYCLE LOOP */
         for (adaptcount=1; adaptcount<=ADAPTLENGTH; adaptcount++)
         {
             /* GENERATE PROPOSAL VALUE (stored in y). */
             sigma = exp(a/2)*pow((1+absval(x)),(b/2));
             y = x + sigma * normal();

             /* ACCEPT/REJECT. */
             logalpha = targlogdens(y) - targlogdens(x)
                     + b/2 * (log(1.0+absval(x)) - log(1.0+absval(y)) )
                     - (x-y)*(x-y)/2 *
                             ( exp( -a-b*(log(1.0+absval(y))) )
                             - exp( -a-b*(log(1.0+absval(x)) ) ) );
                     accepted = ( log(drand48()) < logalpha );
```

```
        if (accepted){
            fprintf(fpmsq, ", %f", pow(x-y, 2));
            x = y;
            numaccept++;
        }
        else
        {
            fprintf(fpmsq, ", 0");
        }
        fprintf(fpx, ", %f", x);

        /* Update various counts. */
        numit++;

        /* if (x<-1) m1count++; */
        /* printf("TESTER: numit=%d, m1count=%d\n", numit, m1count); */

        logval = log(1+absval(x));
        pilogsum = pilogsum + logval;

        // update certain counters
        if (logval > pilogest) {
            possum++;
            if (accepted) {
                posaccept++;
            }
        } else {
            negsum++;
            if (accepted) {
                negaccept++;
           }
        }

    } /* End of adaptcount for loop. */

    /* Update various estimates etc. */
    pilogest = pilogsum / numit;
```

```
/* DO THE ADAPTING. */

/* Adapt variable a. */
if (numaccept > ADAPTLENGTH * TARGACCEPT ) {
    a = a + 1.0/adaptcount;
} else {
    a = a - 1.0/adaptcount;
}

/*
> If average acceptance probability in the region where
>
> \log (1 + |x| ) - E_\pi (\log (1+|x|)) > 0
>
> is smaller than that in the region's compelment then decrease b by 1
> otherwise increase it by 1/i.
*/

/* Adapt b. */
/* if ( posaccept * negcount < negaccept * poscount ) */
if ( posaccept * negsum < negaccept * possum ) {
    b = b - 1.0/adaptcount;
} else if (posaccept * negsum > negaccept * possum){
    b = b + 1.0/adaptcount;
}

/* Prevent b from getting too extreme. */
if (b > MAXABSB)
    b = MAXABSB;
if (b < -MAXABSB)
    b = -MAXABSB;

if (t == xspacing * (t/xspacing)) {
    /* Write X[0] to file. */
    if (t > xspacing) { /* Not first one printed, so need commas. */
        fprintf(fpa, ",");
        fprintf(fpb, ",");
        fprintf(fppi, ",");
```

```
                    fprintf(fpacc, ",");
                }
                fprintf(fpa, " %f", a);
                fprintf(fpb, " %f", b);
                fprintf(fppi, " %f", pilogest);
                fprintf(fpacc, " %f", ((double)numaccept)/ADAPTLENGTH);
       }

        } /* End of printcount for loop. */

      /* OUTPUT SOME VALUES TO SCREEN. */
      printf("%9d: x=%.3f, acc=%.3f, pilogest=%f, a:=%.5f, b:=%.5f\n",
        numit, x, ((double)numaccept)/ADAPTLENGTH, pilogest, a, b);


    } /* End of main iteration loop. */

    fprintf(fpx, " )\n");
    fprintf(fpx, "plot(xvector, type=\"l\", main=\"x convergence\")\n");
    fclose(fpx);
    fprintf(fpa, " )\n");
    fprintf(fpa, "plot(avector, type=\"l\", main=\"a convergence\")\n");
    fclose(fpa);
    fprintf(fpb, " )\n");
    fprintf(fpb, "plot(bvector, type=\"l\", main=\"b convergence\")\n");
    fclose(fpb);
    fprintf(fppi, " )\n");
    fprintf(fppi, "plot(pivector, type=\"l\", main=\"pi convergence\")\n");
    fclose(fppi);
    fprintf(fpacc, " )\n");
    fprintf(fpacc, "plot(accvector, type=\"l\", main=\"acc convergence\")\n");
    fclose(fpacc);
    fprintf(fpmsq, " )\n");
    fprintf(fpmsq, "plot(msqvector, type=\"l\", main=\"msq convergence\")\n");
    fclose(fpmsq);
    return(0);

}
```

```
/* TARGLOGDENS Target Log Density
*/
double targlogdens( double w ) { #ifdef CAUCHYTARG
    return( -log(1.0+w*w) );
#endif
    return(-w*w/2);

}

/* ABSVAL */ double absval( double w ) {
  if (w<0)
    return(-w);
  else
    return(w);
}

/* SEEDRAND: SEED RANDOM NUMBER GENERATOR. */ seedrand() {
    int i, seed;
    struct timeval tmptv;
    gettimeofday (&tmptv, (struct timezone *)NULL);
    /* seed = (int) (tmptv.tv_usec - 1000000 *
                (int) ( ((double)tmptv.tv_usec) / 1000000.0 ) ); */
    seed = (int) tmptv.tv_usec;
    srand48(seed);
    (void)drand48();  /* Spin it once. */
    return(0);
}


/* NORMAL:  return a standard normal random number. */ double
normal() {
    double R, theta, drand48();

    R = - log(drand48());
    theta = 2 * PI * drand48();
    return( sqrt(2*R) * cos(theta));
}
```

Bibliography

[1] Andrieu C., and E. Moulines (2003), *On the ergodicity properties of some adaptive Markov Chain Monte Carlo algorithms* MCMC Preprint Service, Statistical Laboratory, Cambridge University. http://www.statslab.cam.ac.uk/ mcmc/

[2] Andrieu C., and Y.F. Atchadé (2005), *On the efficiency of adaptive MCMC algorithms* MCMC Preprint Service, Statistical Laboratory, Cambridge University. http://www.statslab.cam.ac.uk/ mcmc/

[3] Atchadé Y.F., and J.S. Rosenthal (2005), *On Adaptive Markov Chain Monte Carlo Algorithms* Bernoulli. Vol.11, pg. 815-828.

[4] Billingsley, Patrick (1995), *Probability and Measure.* Wiley-Interscience, 3rd edition.

[5] Cogburn, Robert (1972), *The central limit theorem for Markov processes.* Proceedings of the Sixth Berkeley Symposium in Mathematical Statistics and Probability, Vol.2, pg. 485-512, University of California Press, Berkeley.

[6] Dudley, R.M. (2002), *Real Analysis and Probability.* Cambridge University Press.

[7] Durrett, Richard (2005), *Probability: Theory and Examples.* Duxbury Advanced Series.

[8] Evans, M.J. and J.S. Rosenthal (2004), *Probability and Statistics: The Science of Uncertainty.* W.H. Freeman.

[9] Folland, Gerald B. (1999), *Real Analysis: Modern Techniques and Their Applications.* Wiley-Interscience Publication, 2nd edition.

[10] W.R. Gilks, G.O. Roberts, and S.K. Sahu (1998), *Adaptive Markov Chain Monte Carlo.* Journal of American Statistical Association. Vol.93, pg. 1045-1054.

[11] Grimmett, G.R. and D.R. Stirzaker (1992), *Probability and Random Processes.* Oxford Science Publications, 2nd edition.

[12] H.Haario, E.Saksman, and J.Tamminen (2001), *An adaptive Metropolis algorithm.* Bernoulli. Vol.7, pg. 223-242.

[13] J.P. Hobert, G.L. Jones, B. Presnell, and J.S. Rosenthal (2002), *On the Applicability of Regenerative Simulation in Markov Chain Monte Carlo.* Biometrika Vol.89, pg. 731-743.

[14] Hohendorff, J.M. and Rosenthal J.S. (2005), *An Introduction to Markov Chain Monte Carlo.* University of Toronto, Department of Statistics, supervised reading report (http://www.probability.ca/jeff/grad.html)

[15] Ibragimov, I.A. and Y.V. Linnik (1971), *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff. (English translation)

[16] Jones, Galin L. (2004), *On the Markov chain central limit theorem.* Probability Surveys, Vol.1 pg. 299-320.

[17] Lindvall, Torgny (1992), *Lectures on the Coupling Method.* Willey Series in Probability and Mathematical Statistics.

[18] Mengersen, K.L. and R.L. Tweedie (1996), *Rates of Cconvergence of the Hastings and Metropolis algorithms.* The Annals of Statistics, Vol.24, pg. 101-121.

[19] Meyn, S.P. and R.L. Tweedie (1993), *Markov chains and stochastic stability.* Springer-Verlag, Communications and Control Engineering Series.

[20] Neal, Radford M. (1993), *Probabilistic Inference Using Markov Chain Monte Carlo Methods.* University of Toronto, Department of Computer Science, Technical Report CRG-TR-93-1.

[21] Pasarica C., A. Gelman (2003), *Adaptively scaling the Metropolis algorithm using the average squared jumped distance.* MCMC Preprint Service, Statistical Laboratory, Cambridge University. http://www.statslab.cam.ac.uk/ mcmc/

[22] Revuz, D. (1975) *Markov Chains.* North-Holland.

[23] Roberts, Gareth O. (1999), *A note on acceptance rate criteria for CLTs for Metropolis-Hastings algorithms.* Journal of Applied Probability. V.36, pg 1210 - 1217.

[24] Roberts, G.O. and J.S. Rosenthal (2004) *Coupling and Ergodicity of Adaptive MCMC* MCMC Preprint Service, Statistical Laboratory, Cambridge University. http://www.statslab.cam.ac.uk/ mcmc/

[25] Roberts, G.O. and J.S. Rosenthal (2004), *General state space Markov chains and MCMC algorithms.* Probability Surveys. V.1, pg 20 - 71

[26] Roberts, G.O. and R.L. Tweedie (1996), *Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms.* Biometrika, Vol.3, pg. 95-110.

[27] Rosenthal, Jeffrey S. (2000), *A First Look at Rigorous Probability Theory.* World Scientific.

[28] Ross, Sheldon M. (2003), *Introduction to Probability Models.* Academic Press, Eigth Edition.

[29] Tierney, Luke (1994), *Markov chains for exploring posterior distributions.* Annals of Statistics, Vol.22, pg. 775-798.