# On The Weak Law Of Large Numbers For Unbounded Functionals For Adaptive MCMC

CHAO YANG

*Department of Mathematics*

*University of Toronto*

*M5S 2E4, Toronto ON, Canada*

chaoyang@math.toronto.edu

**Abstract**

We present counter-examples to demonstrate that when $g$ is unbounded the conditions of Simultaneous Uniform Ergodicity and Diminishing Adaptation are not enough to guarantee that the the weak law of large numbers (WLLN) holds, although from the results of Roberts and Rosenthal [4] we know that WLLN holds under these conditions when $g$ is bounded. Then we show various theoretical results of the WLLN for the adaptive MCMC and unbounded measurable function $g$. Finally we apply our results to the Adaptive Metropolis algorithm proposed by Haario et al. [7] (2001).

## 1 Introduction

Markov chain Monte Carlo (MCMC) is a popular method to simulate any distribution $\pi(\cdot)$ on the state space through the use of Markov chains. In practice, we can choose the transition probabiliy $P$ from the family where $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ is a collection of Markov chain kernels with stationary distribution $\pi(\cdot)$ on $\chi$. Then the question is how to optimize the choice of the Markov chain's kernel. The initial idea is to choose a "best" $P_\gamma$, but it has been proved by Gilks et al [5] (1996) that the optimal choice depends on the property of the target distribution $\pi$. Therefore another solution so-called adaptive MCMC has been proposed recently. The main idea of this method is to tune the parameter at every step using the information from the "history" so that the choice of the parameter is more reasonable than the fixed one. See Gilks et al [6] (1998), Haario et al. [7] (2001), Andrieu and Robert [3] (2001), Andrieu and Moulines [2] (2005), Roberts and Rosenthal

[4] (2005), Atchade and Rosenthal [8] (2005) , and Andieu and Achade [1] (2005) for example. Actually we can formalize the method as below (see Roberts and Rosenthal (2005) [4]):

Let $\{X_n\}$ be a discrete time series on a state space $\chi$ with $\sigma-$algebra $\mathcal{F}$, $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ be a collection of Markov chain kernels with stationary distribution $\pi(\cdot)$ on $\chi$, and $\Gamma_n$ be $\mathcal{Y}-$valued random variables which are updated according to specified rules. Thus we can define:

$$P[X_{n+1}\in A|X_n=x,\Gamma_n=\gamma,\mathcal{G}_n]=P_\gamma(x,A) \tag{1.1}$$

where $\mathcal{G}_n=\sigma(X_0,\cdots,X_n,\Gamma_0,\cdots,\Gamma_n)$. Then we call $\{X_n\}$ an adaptive MCMC with adaptive scheme $\Gamma_n$. Let

$$A^{(n)}((x,\gamma),B)=P[X_n\in B|X_0=x,\Gamma_0=\gamma],\qquad B\in\mathcal{F}$$

and

$$T((x,\gamma),n)=\|A^{(n)}((x,\gamma),\cdot)-\pi(\cdot)\|$$

In Roberts and Rosenthal [4] (2005), they proved the following theorems:

**Theorem 1.1.** *Consider an adaptive MCMC algorithm on a state space $\chi$, with adaptation index $\mathcal{Y}$ and the adaptive scheme is $\Gamma_n$. $\pi(\cdot)$ is stationary for each kernel $P_\gamma$ for $\gamma\in\mathcal{Y}$. Suppose also that:*

*(a)[Simultaneous Uniform Ergodicity] For all $\epsilon$, there is $N=N(\epsilon)\in N$ such that $\|P_\gamma^N(x,\cdot)-\pi(\cdot)\|\le\epsilon$ for all $x\in\mathcal{X}$ and $\gamma\in\mathcal{Y}$; and*

*(b)[Diminishing Adaption] $lim_{n\to\infty}D_n=0$ in probability, where $D_n=sup_{x\in\mathcal{X}}\|P_{\Gamma_{n+1}}-P_{\Gamma_n}\|$ is a $\mathcal{G}_{n+1}$-measurable random variable.*

*Then $lim_{n\to\infty}T(x,\gamma,n)=0$ for all $x\in\mathcal{X}$ and $\gamma\in\mathcal{Y}$.*

**Theorem 1.2.** *Consider an adaptive MCMC algorithm. Suppose that conditions (a) and (b) hold. Let $g:\mathcal{X}\to R$ be a bounded measurable function. Then for any starting values $x\in\mathcal{X}$ and $\gamma\in\Gamma$, conditional on $X_0=x$ and $\Gamma_0=\gamma$ we have:*

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

*in probability as $n\to\infty$.*

This leads to the following questions:

**Question 1.** We know that condition (a) is equivalent to the following statement: There exist $M > 0$ and $1 > \rho > 0$ such that for any $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, $\|P_\gamma^n(x,\cdot) - \pi(\cdot)\| < M\rho^n$. In particular, $\sup_{\gamma\in\mathcal{Y},\ x\in\mathcal{X}} \sum_{n=1}^{\infty} \|P_\gamma^n(x,\cdot) - \pi(\cdot)\| < \infty$. Then do we still have $\sum_{n=1}^{\infty} T(x,\gamma,n) < \infty$ for the Adaptive MCMC satisfying conditions (a) and (b)?

**Question 2.** Does the WLLN hold for all unbounded $g \in L(\pi)$ under the same conditions?

We will answer both of these questions in negative in section 2, prove the WLLN of adaptive MCMC under stronger conditions in section 3.

# 2 A Counterexample

Consider $\mathcal{X} = (0,1]$, $\mathcal{Y} = (0,1] \times \mathbf{N}$, $\pi(\cdot)$ is the Lebesgue measure on $\mathcal{X}$, and

$$g(x) = x^{-\frac{1}{2}}$$

therefore $\pi(g) = 2$. Furthermore, for $(\gamma, k) \in \mathcal{Y}$ define the kernel $P_{(\gamma,k)}$ by:

$$P_{(\gamma,k)}(x,A) = \begin{cases} \frac{2}{3}\pi(A) + \frac{1}{3}\delta_x(A) & \text{if } x \neq \gamma \\ \frac{2}{3}\pi(A) + \frac{1}{3}\delta_{\frac{1}{4^k}}(A) & \text{if } x = \gamma \end{cases}$$

and construct the adaptive scheme as below:

First we define $\{I_n\}_{n=1}^{\infty}$ to be an independent random variable sequence such that:

$$I_n = \begin{cases} 1 & \text{with probability} \frac{1}{n} \\ 0 & \text{with probability} \frac{n-1}{n} \end{cases}$$

Secondly we let $\Gamma_{n+1} = \Gamma_n \times (1 - I_n) + (X_{n+1}, n+1) \times I_n$.

## 2.1 The Answer To Question 2

**Theorem 2.1.** *The above adaptive MCMC algorithm satisfies conditions (a) and (b) and $\pi(|g|) < \infty$ , but the WLLN does NOT hold.*

**Lemma 2.1.** *The above adaptive MCMC algorithm satisfies conditions (a) and (b)*

*Proof.* Obviously each $P_{(\gamma,k)}$ is stationary with respect to $\pi$, and $\|P_{(\gamma,k)}(x,\cdot)-\pi(\cdot)\|_{var} \leq \frac{1}{3}$ for any $(\gamma,k)$, so such a family of kernels satisfy the condition (a) following the Proposition 7 in Roberts and Rosenthal [9] (2004);

And following the definition of $\Gamma_n$, we have:

$$\begin{aligned} D_n &= \sup_{x\in\mathcal{X}} \|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\| \\ &\leq P(\Gamma_{n+1} \neq \Gamma_n) \\ &= P(I_n = 1) \\ &= \frac{1}{n} \end{aligned}$$

Therefore we have the conditions (a) and (b) holds. □

To prove the theorem 2.1, we show the following lemmas first:

**Lemma 2.2.** *For any $\epsilon > 0$ and any sequence $\{x_i\}_{i=0}^{\infty}$, if $n$ and $k$ are two positive integers such that $n < k < \frac{2^n-(1+\epsilon)n-1}{1+\epsilon}$ and we also have $g(x_n) = 2^n$ then:*

$$\frac{\sum_{i=1}^{k} g(x_i)}{k} - 2 > \epsilon \tag{2.2}$$

*Proof.* Since $\frac{k+2^n-1}{k}$ strictly decreases with respect to $k$ and $g(x) \geq 1$, we have:

$$\begin{aligned} \frac{\sum_{i=1}^{k} g(x_i)}{k} - 2 - \epsilon &\geq \frac{k-1+2^n}{k} - 2 - \epsilon \\ &\geq \frac{\frac{2^n-(1+\epsilon)n-1}{1+\epsilon} - 1 + 2^n}{\frac{2^n-(1+\epsilon)n-1}{1+\epsilon}} - 2 - \epsilon \\ &\geq 1 + \frac{2^n-1}{\frac{2^n-(1+\epsilon)n-1}{1+\epsilon}} - 2 - \epsilon \\ &\geq 1 + \frac{2^n-1}{2^n-(1+\epsilon)n-1} \times [1+\epsilon] - 2 - \epsilon \\ &> 1 + 1 + \epsilon - 2 - \epsilon \\ &= 0 \end{aligned}$$

□

**Lemma 2.3.** *For any $\epsilon > 0$, there exists $M_\epsilon$ such that for any $m > M_\epsilon$ we have:*

$$\frac{2^{m+1} - (m+1)(1+\epsilon) - 1}{1+\epsilon} > m^2$$

*Proof.* Denote $h_m = \frac{2^{m+1}-(m+1)(1+\epsilon)-1}{1+\epsilon} - m^2$, then we have $\lim_{m\to\infty} h_m = \infty$. Therefore there exists $M_\epsilon$ such that for any $m > M_\epsilon$ we have $h_m > 0$, i.e. $\frac{2^{m+1}-(m+1)(1+\epsilon)-1}{1+\epsilon} > m^2$. □

For any $0 < \epsilon < \frac{1}{6}$, we define $N_\epsilon = \max\{M_\epsilon, \frac{1}{1-6\epsilon}\}$, then we can prove that:

**Lemma 2.4.** *For any $X_0 = x$, $\Gamma_0 = \gamma$ and $0 < \epsilon < \frac{1}{6}$, then we have:*

$$P(|\frac{\sum_{i=1}^n g(X_i)}{n} - \pi(g)| > \epsilon | X_0 = x, \Gamma_0 = \gamma) > 2\epsilon \quad \textit{for any} \quad n > N_\epsilon^2$$

*Proof.* For any $n > N_\epsilon^2$, we have:

$$\begin{aligned} & P(I_m = 0, \text{for any } m \text{ satisfies} \lfloor\sqrt{n}\rfloor + 1 \le m \le n) \\ = & \prod_{i=\lfloor\sqrt{n}\rfloor+1}^{n} \frac{i}{i+1} \\ = & \frac{\lfloor\sqrt{n}\rfloor}{n} \\ \le & \frac{1}{\sqrt{n}} \\ \le & \frac{1}{N_\epsilon} \end{aligned}$$

then:

$$P(\exists m, \lfloor\sqrt{n}\rfloor + 1 < m < n, I_m = 1) \ge \frac{N_\epsilon - 1}{N_\epsilon} \tag{2.3}$$

Whenever $\Gamma_{n+1} = (X_{n+1}, n+1)$, we have $g(X_{n+1}) = 2^{n+1}$ w.p. $\frac{1}{3}$. Since $N_\epsilon > \frac{1}{1-6\epsilon}$, we have $\frac{N_\epsilon - 1}{3N_\epsilon} > 2\epsilon$. Therefore:

$$P(\exists m,\ \lfloor\sqrt{n}\rfloor + 1 < m < n, g(X_m) = 2^m)) > \frac{N_\epsilon - 1}{3N_\epsilon} > 2\epsilon \tag{2.4}$$

Also since $m > N_\epsilon$, lemma 2.3 indicates that for any $\lfloor\sqrt{n}\rfloor + 1 < m < n$ we have:

$$\frac{2^m - (1+\epsilon)m - 1}{1+\epsilon} > m^2 + 1 > (\lfloor\sqrt{n}\rfloor + 1)^2 + 1 > n + 1$$

Following lemma 2.2 and $m < n < \frac{2^m-(1+\epsilon)m-1}{1+\epsilon}$, we know that

$$\frac{\sum_{i=1}^n g(x_i)}{n} - 2 \quad \ge \quad \epsilon$$

Therefore:

$$\begin{aligned} & P(|\frac{\sum_{i=1}^{n} g(X_i)}{n} - 2| > \epsilon) \\ \geq \; & P(\frac{\sum_{i=1}^{n} g(X_i)}{n} - 2 \geq \epsilon) \\ \geq \; & P(\exists m, \lfloor \sqrt{n} \rfloor < m < n, g(X_m) = 2^m)) \\ > \; & 2\epsilon \text{ following (2.4)} \end{aligned}$$

□

Now we can prove the theorem:

*Proof.* Consider the above example, following all the lemmas above we know that for any $\epsilon > 0$, we have:

$$\limsup_{n\to\infty} P(|\frac{\sum_{i=1}^{n} g(X_i)}{n} - \pi(g)| > \epsilon) > 2\epsilon$$

In other words, we do NOT have:

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

So the WLLN does NOT hold in this example. □

## 2.2 The Answer To Question 1

**Theorem 2.2.** *For the above adaptive MCMC which satisfies conditions (a) and (b), for any $x \in \mathcal{X}$, there exits a measurable set $B$ such that $\sum_{n=1}^{\infty} A^{(n)}(x, B) = \infty$.*

*Proof.* Consider the set $B = \{\frac{1}{4^k} | k = 1, 2, \cdots, \}$, suppose for any start valuve $X_0 = x$ and $\Gamma_0 = \gamma$, we have:

$$\sum_{i=1}^{\infty} A^i((x, \gamma), B) < \infty$$

then for any $0 < \epsilon < 1$, there exists $N_{x,\gamma} > 0$ such that

$$\sum_{i=N+1}^{\infty} A^i((x, \gamma), B) \leq \epsilon$$

Because

$$P(X_{n+1} = \frac{1}{4^n} | \Gamma_n = (X_n, n)) \geq \frac{1}{3}$$

and

$$P(\Gamma_n = (X_n, n)) \geq P(I_n = 1)$$

we can get

$$\begin{aligned} P(X_{n+1} = \frac{1}{4^n}) &\geq P(X_{n+1} = \frac{1}{4^n}, \Gamma_n = (X_n, n)) \\ &\geq \frac{1}{3} P(\Gamma_n = (X_n, n)) \\ &\geq \frac{1}{3} P(I_n = 1) \\ &= \frac{1}{3n} \end{aligned}$$

then following the Borel-Cantelli lemma see Jeffrey S. Rosenthal [11] (2000), we have:

$$\begin{aligned} 1 &= P(\exists\ m > N_{x,\gamma} \text{ s.t. } I_m = 1) \\ &\leq \sum_{i=N+1}^{\infty} P^i((X_i = \frac{1}{4^i} | X_0 = x, \Gamma_0 = \gamma) \\ &\leq \sum_{i=N+1}^{\infty} A^i((x,\gamma), B) \\ &\leq \epsilon \end{aligned}$$

Contradiction!! So we have $\sum_{i=1}^{\infty} A^i((x,\gamma), B) = \infty$. Since $\pi(B) = 0$, we can get:

$$\sum_{i=1}^{\infty} T^i((x,\gamma), B) < \infty$$

Therefore $A^i((x,\gamma), \cdot)$ is neither uniformly nor geometrically ergodic. □

## 3 Summable Adaptive Conditions

From the above counter-example, we know that conditions (a) and (b) are not sufficient conditions to the WLLN of unbounded functions, so we need to strengthen them. Intuitively if $n$ is large enough, for any $k$, $l > n$, $\Gamma_k$ and $\Gamma_l$ are "almost" the same, then the WLLN may hold for any $g \in L(\pi)$. Let us consider the following condition:
$(b^{'})$[Summable Adaption]$\sum_{i=1}^{\infty} sup_{x \in \mathcal{X}} \| P_{\Gamma_{i+1}}(x, \cdot) - P_{\Gamma_i}(x, \cdot) \| < \infty$ Actually we can prove the following theorem:

**Theorem 3.1.** *Consider an adaptive MCMC algorithm. Suppose that conditions (a) and* $(b^{'})$ *hold. Let* $g : \mathcal{X} \to R$ *be a measurable function such that* $\pi(|g|) < \infty$*. Then for any starting values* $x \in \mathcal{X}$ *and* $\gamma \in \Gamma$*, conditional on* $X_0 = x$ *and* $\Gamma_0 = \gamma$ *we have:*

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

*in probability as* $n \to \infty$*.*

*Proof.* Denote

$$S_n \quad = \quad \sum_{i=n}^{\infty} sup_{x \in \mathcal{X}} \| P_{\Gamma_{i+1}}(x, \cdot) - P_{\Gamma_i}(x, \cdot) \|$$

for any $\epsilon > 0$, following condition $(b^{'})$, there exists $N_1$ such that:

$$P(S_{N_1} > \epsilon) < \frac{\epsilon}{4}$$

We can denote $E = \{S_{N_1} < \epsilon\}$. Since $|g| < \infty$, there exists $N_2$, such that for any $n > N_2$

$$P(|\frac{\sum_{i=1}^{N_1} g(X_i)}{n}| > \frac{\epsilon}{2}) < \frac{\epsilon}{4}$$

Define $N = \max\{N_1, N_2\}$, and we can construct a second chain $\{X_n^{'}\}_{n=N}^{\infty}$ on $E$ such that $X_N^{'} = X_N$ and $X_n^{'} \sim P_{\Gamma_N}(X_{n-1}^{'}, \cdot)$ for $n > N$, and such that:

$$\sum_{n=N}^{\infty} P(X_n \neq X_n^{'}, E) < \frac{\epsilon}{4}$$

Define the events: $B^n(\epsilon) = \{|\frac{\sum_{i=N+1}^{n} g(X_i^{'})}{n}| > \frac{\epsilon}{2}, \text{given } X_N, \Gamma_N\}$, then we can get:

$$\lim_{n \to \infty} P(B^n(\epsilon)) = 0$$

That is when $n$ is large enough we have $P(B^n(\epsilon)) < \frac{\epsilon}{4}$. Then we have

$$\begin{aligned}
& P(|\frac{\sum_{i=1}^{n} g(X_i)}{n}| > \epsilon) \\
\leq \quad & P(|\frac{\sum_{i=1}^{N} g(X_i)}{n}| > \frac{\epsilon}{2}) + P(|\frac{\sum_{i=N+1}^{n} g(X_i)}{n}| > \frac{\epsilon}{2}) \\
\leq \quad & P(|\frac{\sum_{i=1}^{N} g(X_i)}{n}| > \frac{\epsilon}{2}) + P(|\frac{\sum_{i=N+1}^{n} g(X_i)}{n}| > \frac{\epsilon}{2}, E) + P(|\frac{\sum_{i=N+1}^{n} g(X_i)}{n}| > \frac{\epsilon}{2}, E^c) \\
\leq \quad & P(|\frac{\sum_{i=1}^{N} g(X_i)}{n}| > \frac{\epsilon}{2}) + P(|\frac{\sum_{i=N+1}^{n} g(X_i)}{n}| > \frac{\epsilon}{2}, E^c) \\
+ \quad & P(|\frac{\sum_{i=N+1}^{n} g(X_i^{'})}{n}| > \frac{\epsilon}{2}, E) + \sum_{i=N+1}^{n} P(X_i \neq X_i^{'}, E) \\
\leq \quad & \epsilon
\end{aligned}$$

□

**Remark:**According to the conditions in the above proposition, we know that when $N$ is large enough, the sequence $\{X_n\}_{n=N}^{\infty}$ is almost equal to $\{X_n^{'}\}_{n=N}^{\infty}$ which is a Markov chain with transition kernel $P_{\Gamma_n}$. At the first sight, adaptive algorithms that satisfy the conditions $(a)$ and $(b^{'})$ cannot show the adaptive MCMC's advantages sufficiently. But following Roberts and Rosenthal [10] (2005), we know that in lots of cases, the adaptive MCMC will tune the parameter to an "optimal" one after "learning" the information from the historical samples. So we can adjust the convergence speed of $S_n$ such that the adaptive chain can learn enough to find the optimal parameter, that is we can make $N$ very large, such that $\Gamma_N$ is almost a "good" parameter.

# 4 The WLLN For Adaptive Metropolis-Hastings Algorithm

Usually we construct the transition kernel using Metropolis-Hastings algorithms. Suppose that $\pi(\cdot)$ has a density $\pi$, and $Q(x,\cdot)$ is the proposal distribution with a density $q(x,y)$, i.e. $Q(x,dy) = q(x,y)dy$; then the Metropolis-Hasting algorithm proceeds as below:

First we need to choose the starting value $X_0$. Then given $X_n$, generate a proposal $Y_{n+1}$ from $Q(X_n,\cdot)$. Also flip an independent coin, whose probability of heads equals to $\alpha(X_n, Y_{n+1})$, where

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

Then if the coin is heads, "accept" the proposal by setting $X_{n+1} = Y_{n+1}$; otherwise set $X_{n+1} = X_n$. Replace $n$ by $n+1$ and repeat. We can observe that the Metropolis-Hastings algorithms do not have densities with respect to some finite measure. However, if the proposal kernels have uniformly bounded densities, Roberts and Rosenthal [4] (2005) have proved the following ergodic property:

**Corollary 4.1.** *Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaptation property, and also that each $P_\gamma$ is ergordic for $\pi(\cdot)$. Suppose further that for each $\gamma \in \mathcal{Y}$, $P_\gamma$ represents a Metropolis-Hastings algorithm with proposal kernel $Q_\gamma(x,dy) =$*

*$f_\gamma(x,y)\lambda(dy)$ having a density $f_\gamma(x,y)$ with respect to some finite reference measure $\lambda(\cdot)$ on $\mathcal{X}$, with corresponding density $w$ for $\pi(\cdot)$ so that $\pi(dy) = w(y)\lambda(dy)$. Finally, suppose $f_\gamma(x,y)$ are uniformly bounded, and that for each fixed $y \in \mathcal{X}$, the mapping $(x,\gamma) \mapsto f_\gamma(x,y)$ is continuous with respect to some product metric space topology, with respect to which $\mathcal{X}\times\mathcal{Y}$ is compact. Then $\lim_{n\to\infty} T(x,\gamma,n) = 0$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$.*

Then we can prove the following WLLN for unbounded measurable function $g$.

**Theorem 4.1.** *Consider an adaptive MCMC that satisfies the conditions in corrollary 4.1. Then for any measurable function $g$ such that $\lambda(|g|) < \infty$ and $\pi(|g|) < \infty$ we have:*

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

*in probability as $n \to \infty$, conditional on $X_0 = x_*$ and $\Gamma_0 = \gamma_*$.*

**Remark:** If there exist $M > m > 0$ such that $m < w(x) < M$, where $\pi(dy) = w(y)\lambda(dy)$, then we know that $\lambda(|g|) < \infty$ if and only if $\pi(|g|) < \infty$. A typical case is that the state space $\mathcal{X}$ is compact set in $R^d$, $w(y)$ is continuous function on $\mathcal{X}$ and $\lambda$ is Lebesgue measure. Then we have $M > w(x) > m > 0$, and the WLLN of the adaptive MCMC satisfying the conditions in corollary 4.1 will hold for any measurable function $g$ such that $\pi(|g|) < \infty$.

We will prove the theorem following the steps below:

**Step 1:** For all $M > 0$, denote $E_M = \{x \in \mathcal{X} || g(x)| \le M\}$ and for all $\varepsilon$ define:

$$\begin{aligned} M_\varepsilon &= inf\{M > 0 | \lambda(E_M) \ge 1-\varepsilon, \int_{E_M} |g(x)|\lambda(dx) \ge s - \varepsilon\} \\ &= inf\{M > 0 | \lambda(E_M^c) \le \varepsilon, \int_{E_M^c} |g(x)|\lambda(dx) \le \varepsilon\} \end{aligned}$$

If $\lambda(|g|) < \infty$, we will prove that $\varepsilon \cdot M_\varepsilon \to 0$ as $\varepsilon \to 0$;

**Step 2:** Suppose $P_\gamma(x,A) = \int_A \tilde{f}_\gamma(x,y)\lambda(dx) + r_\gamma(x)\delta_x(A)$ then Under the conditions of the theorem 4.1 we have $0 < r_\gamma(x) < \eta$;

**Step 3:** Suppose $A_\gamma^n(x,A) = P(X_n \in A | X_0 = 0, \Gamma_0 = \gamma)$, then there exist $L > 0$ and $0 < \eta < 1$, then under the conditions of the theorem, we have

$$A_\gamma^n(x,B) = \int_B h_\gamma^{(n)}(x,y)\lambda(dy) + w_\gamma^{(n)}(x)\delta_x(B)$$

such that $h_\gamma^{(n)}(x,y) < L$ and $w_\gamma^{(n)}(x) < \eta^n$;

**Step 4:** Prove the WLLN using coupling methods.

## 4.1 Some Technical Results

Suppose the probability of accepting a proposal $y$ generated from $x$ according to $Q_\gamma$ is given by $\alpha_\gamma(x,y) = \min\{1, \frac{g(y)f_\gamma(y,x)}{g(x)f_\gamma(x,y)}\}$, so we have:

$$P_\gamma(x,B) = \int_B f_\gamma(x,y)\alpha_\gamma(x,y)\lambda(dy) + (1 - \int_{\mathcal{X}} \alpha_\gamma(x,y)\lambda(dy))\delta_x(B)$$

We can denote $\tilde{f}_\gamma(x,y) = f_\gamma(x,y)\alpha_\gamma(x,y)$, $r_\gamma(x) = (1 - \int_{\mathcal{X}} \alpha_\gamma(x,y)\lambda(dy))$ and suppose $f_\gamma(x,y) < F$. Obviously we have $\tilde{f}_\gamma(x,y) < F$ since $\alpha_\gamma(x,y) \leq 1$. We also need to prove the following lemmas before we prove the theorem.

**Lemma 4.2.** *Suppose $(\chi, \mathfrak{F}, \lambda)$ is a probability space, and $g : \chi \to R$ is a measurable function such that $\lambda(|g|) = s < \infty$. Then for $\forall \varepsilon > 0$, there exists $M > 0$, such that: $\lambda(E_M) \geq 1 - \varepsilon$ and $\int_{E_M} |g(x)|\lambda(dx) \geq s - \varepsilon$*

*Proof.* Suppose there exits $\varepsilon_0 > 0$, for each $M$, we have

$$\lambda(E_M^c) \geq \varepsilon_0 \tag{4.5}$$

or

$$\int_{E_M} |g(x)|\lambda(dx) \leq s - \varepsilon_0 \tag{4.6}$$

If (4.5) holds, we have $\int_{E_M^c} |g(x)|\pi(dx) \geq M\varepsilon_0$ for all $M$, contradiction!
If (4.6) holds, we have $\int_\chi |g(x)|1_{E_n}(x)\pi(dx) \leq s - \varepsilon_0$ for all $n \in \mathbb{N}$, suppose

$$Y_n = |g(X)|1_{E_n}(x)$$

obviously $Y_n \uparrow |g(X)|$, then by the monotone convergence theorem

$$E_\lambda(|g(x)|) = lim_{n\to\infty} E(Y_n) \leq s - \varepsilon_0$$

which is contradicting with $E_\lambda(|g(x)|) = s$. □

**Lemma 4.3.** *Suppose $g : \chi \to R$ is a measurable function such that $\lambda(|g|) = s < \infty$. Then for each sequence $\{\varepsilon_n\} \to 0$, there exists a subsequence $\varepsilon_{n_k} \searrow 0$ such that $\varepsilon_{n_k} M_{\varepsilon_{n_k}} \to 0$ as $n \to 0$.*

*Proof.* Following lemma 3.2 we know that $0 \leq \frac{\lambda(E^c_{M_{\varepsilon_n}})}{\varepsilon_n} \leq 1$, there is a subsequence $\varepsilon_{n_k} \searrow 0$ such that $\{\frac{\lambda(E^c_{M_{\varepsilon_{n_k}}})}{\varepsilon_{n_k}}\}$ is convergent to some $a$. Then we can think about the problem in the following two cases:
(1) $0 < a \leq 1$; then there exists $N > 0$ such that for each $k > N$, $|\frac{\lambda(E^c_{M_{\varepsilon_{n_k}}})}{\varepsilon_{n_k}} - a| < \frac{a}{2}$, i.e. $\lambda(E^c_{M_{\varepsilon_{n_k}}}) > \frac{a}{2}\varepsilon_{n_k}$, so

$$\begin{aligned} 0 &= lim_{k\to\infty} \int_{E_{M^c_{\varepsilon_{n_k}}}} |g(x)|\pi(dx) \\ &\geq lim_{k\to\infty} \lambda(E^c_{M_{\varepsilon_n}}) M_{\varepsilon_{n_k}} \\ &\geq lim_{k\to\infty} \frac{a}{2}\varepsilon_{n_k} M_{\varepsilon_{n_k}} \\ &\geq 0 \end{aligned}$$

So $lim_{k\to\infty}\varepsilon_{n_k} M_{\varepsilon_{n_k}} = 0$
(2) $a = 0$; then there exists $N$, $k > N$, such that

$$\lambda(E^c_{M_{\varepsilon_{n_k}}}) < \frac{1}{2}\varepsilon_{n_k} \tag{4.7}$$

And following (4.5) for each $\delta > 0$,

$$\lambda(|g(x)| \geq M_{\varepsilon_{n_k}} - \delta) > \varepsilon_{n_k} \tag{4.8}$$

Following (4.7) and (4.8), let $\delta \to 0$, we can get:

$$\begin{aligned} \lambda(|g(x)| = M_{\varepsilon_{n_k}}) &\geq \varepsilon_{n_k} - \frac{1}{2}\varepsilon_{n_k} \\ &= \frac{1}{2}\varepsilon_{n_k} \end{aligned}$$

Since $\varepsilon_{n_{k+1}} < \varepsilon_{n_k}$, $M_{\varepsilon_{n_k}} \leq M_{\varepsilon_{n_{k+1}}}$,

$$\begin{aligned} 0 &= lim_{k\to\infty} \int_{E^c_{M_{\varepsilon_{n_k}}}} |g(x)|\lambda(dx) \\ &\geq lim_{k\to\infty} \int_{\{|g(x)|=M_{\varepsilon_{n+1}}\}} |g(x)|\lambda(dx) \\ &= lim_{k\to\infty} M_{\varepsilon_{n_{k+1}}} \cdot \lambda(|g(x)| = M_{\varepsilon_{n_{k+1}}}) \\ &\geq lim_{k\to\infty} \frac{1}{2} M_{\varepsilon_{k_{n+1}}} \cdot \varepsilon_{n_{k+1}} \\ &\geq 0 \end{aligned}$$

So $lim_{k\to\infty} M_{\varepsilon_{n_{k+1}}} \cdot \varepsilon_{n_{k+1}} = 0$ □

**Lemma 4.4.** *Suppose $g : \chi \to R$ is a measurable function such that $\lambda(|g|) = s < \infty$. Then $\varepsilon \cdot M_\varepsilon \to 0$ as $\varepsilon \to 0$.*

*Proof.* Suppose there exists $c > 0$ such that for each $n \in N$, there exists $\varepsilon_n < \frac{1}{n}$ and $\varepsilon_n \cdot M_{\varepsilon_n} \geq c$ for all $n$, then every subsequence $\{\varepsilon_{n_k}\}$ of $\{\varepsilon_n\}$ satisfies that $\varepsilon_{n_k} \cdot M_{\varepsilon_{n_k}} \geq c$, which is contradicting with the lemma 4.3. So $\varepsilon \cdot M_\varepsilon \to 0$ as $\varepsilon \to 0$. □

**Lemma 4.5.** *Under the conditions of corollary 4.1, we have that condition (a) holds.*

*Proof.* Following the proof of Corollary 12 in Roberts and Rosenthal [4](2005), we can get the lemma directly. □

**Lemma 4.6.** *Condition (a) is equivalent to: There exist $M > 0$ and $0 < \rho < 1$ such that for any $x$, $\gamma$ we have:*

$$\|P^n_\gamma(x, \cdot) - \pi(\cdot)\| \leq M\rho^n$$

*Proof.* Suppose $t_\gamma(n) = 2\sup_{x\in\mathcal{X}} \|P^n_\gamma(x, \cdot) - \pi(\cdot)\|$, following Roberts and Rosenthal [9] (2004) Proposition 3(c), we know that $t_\gamma(m + n) \leq t_\gamma(m)t_\gamma(n)$. Under condition (a), there exists $n$ which is independent of $\gamma$ such that $t_\gamma(n) \equiv \beta < 1$, so for all $j \in \mathbf{N}$, $t_\gamma(jn) \leq (t_\gamma(n))^j = \beta^j$. Therefore, we have:

$$\|P^m_\gamma(x, \cdot) - \pi(\cdot)\| \leq \|P^{\lfloor m/n \rfloor n}_\gamma(x, \cdot) - \pi(\cdot)\| \leq \frac{1}{2} t_\gamma(\lfloor m/n \rfloor n) \leq \beta^{\lfloor m/n \rfloor} \leq \beta^{-1}(\beta^{1/n})^m$$

so all the kernels are uniformly ergodic with $M = \beta^{-1}$ and $\rho = \beta^{1/n}$. □

**Lemma 4.7.** *Suppose* $P_\gamma(x,A) = \int_A \tilde{f}_\gamma(x,y)\lambda(dx) + r_\gamma(x)\delta_x(A)$*, then there exist measurable functions* $\tilde{f}_\gamma^{(n)}(x,y)$ *on* $\chi^2$ *such that* $P_\gamma^n(x,A) = \int_A \tilde{f}_\gamma^{(n)}(x,y)\lambda(dx) + r_\gamma^n(x)\delta_x(A)$

*Proof.* We will prove it by induction, and obviously the conclusion holds when $n = 1$. We suppose it also holds when $n = k$, then let's consider the case when $n = k+1$:

$$\begin{aligned}
P_\gamma^{k+1}(x,A) &= \int_\chi P_\gamma^k(y,A)P_\gamma(x,dy) \\
&= \int_\chi [\int_A \tilde{f}_\gamma^{(k)}(y,z)\lambda(dz) + r_\gamma^k(y)\delta_y(A)][\tilde{f}_\gamma(x,y)\pi(dy) + r_\gamma(x)\delta_x(dy)] \\
&= \int_\chi \int_A \tilde{f}_\gamma^{(k)}(y,z)\pi(dz)f_\gamma(x,y)\lambda(dy) + \tilde{f}_\gamma^{(k)}(y,z)\pi(dz)r_\gamma(x)\delta_x(dy) \\
&+ r_\gamma^k(y)\delta_y(A)\tilde{f}_\gamma(x,y)\lambda(dy) + r_\gamma^k(y)\delta_y(A)r_\gamma(x)\delta_x(dy) \\
&= \int_A [\int_\chi \tilde{f}_\gamma^{(k)}(y,z)f_\gamma(x,y)\lambda(dy)]\pi(dz) + \int_A r_\gamma(x)\tilde{f}_\gamma^k(x,z)\pi(dz) \\
&+ \int_A r_\gamma^k(y)\tilde{f}_\gamma(x,y)\lambda(dy) + r_\gamma^{k+1}(x)\delta_x(A) \\
&= \int_A \tilde{f}_\gamma^{(k+1)}(x,z)\lambda(dz) + r_\gamma^{k+1}(x)\delta_x(A)
\end{aligned}$$

where

$$\tilde{f}_\gamma^{(k+1)}(x,z) = \int_\chi \tilde{f}_\gamma^{(k)}(y,z)\tilde{f}_\gamma(x,y)\pi(dy) + r_\gamma(x)\tilde{f}_\gamma^k(x,z) + r_\gamma^k(x)\tilde{f}_\gamma(x,z) \quad (4.9)$$

□

**Lemma 4.8.** *Suppose* $P_\gamma(x,A) = \int_A \tilde{f}_\gamma(x,y)\lambda(dx) + r_\gamma(x)\delta_x(A)$ *where* $\lambda(\cdot)$ *is a finite reference measure on* $\mathcal{X}$ *such that* $\lambda(\{x\}) = 0$ *for any* $x$*, with corresponding density* $w$ *for* $\pi(\cdot)$ *so that* $\pi(dy) = w(y)\lambda(dy)$*. Then under condition (a),we have* $0 < r_\gamma(x) < \eta$*, where the* $\eta$ *is the same as in lemma 4.6.*

*Proof.* Because $P_\gamma(x,\{x\}^c) = \int_{\chi-x} \tilde{f}_\gamma(x,y)\pi(dx)$, $P_\gamma(x,x) = r_\gamma(x)$ and $\pi(x) = 0$, and following (4.9), we know that $|P_\gamma^n(x,\{x\}) - \pi(\{x\})| = r_\gamma^n(x)$ for each $x \in \chi$. Then following condition (a), we know for $\forall\epsilon > 0$, there exists $N$ such that $r_\gamma^N(x) < \epsilon$, that is $r_\gamma(x) < \epsilon^{\frac{1}{N}}$ for each $\gamma$ and $x$. Then we take $\epsilon < 1$, and we can get $\eta = \epsilon^{\frac{1}{N}} < 1$ □

**Lemma 4.9.** *Suppose* $A_\gamma^n(x,A) = P(X_n \in A | X_0 = 0, \Gamma_0 = \gamma)$*, then under the conditions of corollary 4.1, there exist* $L > 0$ *and* $0 < \eta < 1$

$$A_\gamma^n(x,B) = \int_B h_\gamma^{(n)}(x,y)\lambda(dy) + w_\gamma^{(n)}(x)\delta_x(B)$$

*such that* $h_\gamma^{(n)}(x,y) < L$ *and* $w_\gamma^{(n)}(x) < \eta^n$

*Proof.* Suppose the joint distribution of $(X_1, X_2, \cdots, X_n, \Gamma_1, \Gamma_2, \cdots, \Gamma_{n-1})$ given $X_0 = x$ and $\Gamma_0 = \gamma$ is $\mu_{(x,\gamma)}^{(n)}$, obviously the marginal distribution of $X_n$ is $A^{(n)}((x,\gamma),\cdot)$. Since $\gamma_n$ is a measurable function of $(x_1, x_2, \cdots, x_n, \gamma_1, \gamma_2, \cdots, \gamma_{n-1})$, we have:

$$\begin{aligned}
A^{(n+1)}((x,\gamma),B) &= \int_{\mathcal{X}^n\times\mathcal{Y}^{n-1}} P_{\Gamma_n}(x_n,B)\mu_{(x,\gamma)}^{(n)}(dx_1\cdots dx_n d\gamma_1\cdots d\gamma_{n-1}) \\
&= \int_{\mathcal{X}^n\times\mathcal{Y}^{n-1}} [\int_B \tilde{f}_{\gamma_n}(x_n,y)\lambda(dy) + r_{\gamma_n}(x_n)(\delta_{x_n}(B))]\mu_{(x,\gamma)}^{(n)}(dx_1\cdots dx_n d\gamma_1\cdots d\gamma_{n-1}) \\
&= \int_B \int_{\mathcal{X}^n\times\mathcal{Y}^{n-1}} \tilde{f}_{\gamma_n}(x_n,y)\mu_{(x,\gamma)}^{(n)}(dx_1\cdots dx_n d\gamma_1\cdots d\gamma_{n-1})\lambda(dy) \\
&+ \int_{\mathcal{X}^n\times\mathcal{Y}^{n-1}} r_{\gamma_n}(x_n)\delta_{x_n}(B)\mu_{(x,\gamma)}^{(n)}(dx_1\cdots dx_n d\gamma_1\cdots d\gamma_{n-1})
\end{aligned}$$

We can observe that the second term:

$$\begin{aligned}
&\int_{\mathcal{X}^n\times\mathcal{Y}^{n-1}} r_{\gamma_n}(x_n)\delta_{x_n}(B)\mu_{(x,\gamma)}^{(n)}(dx_1\cdots dx_n d\gamma_1\cdots d\gamma_{n-1}) \\
= &\int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-1}}\int_{\mathcal{X}} r_{\gamma_n}(x_n)\delta_{x_n}(B)P_{\gamma_{n-1}}(x_{n-1},dx_n)\mu_{(x,\gamma)}^{(n-1)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-1}) \\
= &\int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n}(x_n)P_{\gamma_{n-1}}(x_{n-1},dx_n)\mu_{(x,\gamma)}^{(n-1)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-2}) \\
= &\int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n}(x_n)\tilde{f}_{\gamma_{n-1}}(x_{n-1},x_n)\lambda(dx_n)\mu_{(x,\gamma)}^{(n-1)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-2}) \\
+ &\int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n}(x_n)r_{\gamma_{n-1}}(x_{n-1})\delta_{x_{n-1}}(dx_n)\mu_{(x,\gamma)}^{(n-1)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-2})
\end{aligned}$$

If $\gamma_n = \gamma_n(x, x_1, \cdots, x_n, \gamma, \gamma_1, \cdots, \gamma_{n-1})$, then we can define:

$$\gamma_n^i = \gamma_n(x, x_1, \cdots, x_{n-i-1}, x_{n-i}, x_{n-i}, \cdots, x_{n-i}, \gamma, \gamma_1, \cdots, \gamma_{n-i+1}^1, \cdots, \gamma_{n-1}^{i-1})$$

Similarly we can compute the second term of the above inequality:

$$\begin{aligned}
&\int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n}(x_n)r_{\gamma_{n-1}}(x_{n-1})\delta_{x_{n-1}}(dx_n)\mu_{(x,\gamma)}^{(n-1)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-2}) \\
= &\int_{\mathcal{X}^{n-2}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n^1}(x_{n-1})r_{\gamma_{n-1}}(x_{n-1})P_{\gamma_{n-2}}(x_{n-2},dx_{n-1})\mu_{(x,\gamma)}^{(n-2)}(dx_1\cdots dx_{n-2}d\gamma_1\cdots d\gamma_{n-3}) \\
= &\int_{\mathcal{X}^{n-2}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n^1}(x_{n-1})r_{\gamma_{n-1}}(x_{n-1})\tilde{f}_{\gamma_{n-2}}(x_{n-2},x_{n-1})\lambda(dx_{n-1})\mu_{(x,\gamma)}^{(n-2)}(dx_1\cdots dx_{n-2}d\gamma_1\cdots d\gamma_{n-3}) \\
+ &\int_{\mathcal{X}^{n-2}\times\mathcal{Y}^{n-1}}\int_{B} r_{\gamma_n^1}(x_{n-1})r_{\gamma_{n-1}}(x_{n-1})r_{\gamma_{n-2}}(x_{n-2})\delta_{x_{n-2}}(dx_{n-1})\mu_{(x,\gamma)}^{(n-2)}(dx_1\cdots dx_{n-2}d\gamma_1\cdots d\gamma_{n-3})
\end{aligned}$$

Inductively we have:

$$\begin{aligned}
h_{\gamma}^{(n+1)}(x,y) &= \int_{\mathcal{X}^n\times\mathcal{Y}^{n-1}} \tilde{f}_{\gamma_n}(x_n,y)\mu_{(x,\gamma)}^{(n)}(dx_1\cdots dx_n d\gamma_1\cdots d\gamma_{n-1}) \\
&+ \int_{\mathcal{X}^{n-1}\times\mathcal{Y}^{n-2}} r_{\gamma_n}(x_n)\tilde{f}_{\gamma_{n-1}}(x_{n-1},x_n)\mu_{(x,\gamma)}^{(n-1)}(dx_1\cdots dx_{n-1}d\gamma_1\cdots d\gamma_{n-2}) \\
&+ \int_{\mathcal{X}^{n-2}\times\mathcal{Y}^{n-3}}\int_B \prod_{i=0}^{1} r_{\gamma_{n-i}^i}(x_{n-i})\tilde{f}_{\gamma_{n-2}}(x_{n-2},x_{n-1})\mu_{(x,\gamma)}^{(n-2)}(dx\cdot\cdot dx_{n-2}d\gamma_1\cdot\cdot d\gamma_{n-3}) \\
&+ \cdots \\
&+ \int_B \prod_{i=0}^{n-1} r_{\gamma_{n-i}^i}(x_1)\tilde{f}_{\gamma}(x,x_1)\mu_{(x,\gamma)}^{1}(dx_1) \\
&\leq F\sum_{i=0}^{n-1}\eta^i \\
&\leq \frac{F}{1-\eta}
\end{aligned}$$

and

$$\begin{aligned}
w_{\gamma}^{(n+1)}(x) &= \prod_{i=0}^{n-1} r_{\gamma_{n-i}^{n-i}}(x) \\
&\leq \eta^n
\end{aligned}$$

□

## 4.2 The proof of Theorem 4.1

Now we state the proof using the above lemmas as below:

*Proof.* suppose $\pi(g) = 0$,$\lambda(|g|) = s$,$D_n = sup_{x\in\chi}\|P_{\Gamma_{n+1}}(x,\cdot) - P_{\Gamma_n}\|$ and $f_\gamma(x,y) < F$. Lemma 4.2 implies that given $\varepsilon > 0$, there exits $\eta_1 > 0$ such that $M_{\eta_1}\eta_1 < \epsilon$; denote $\eta_2 = \frac{\epsilon}{F}$, then we have:

$$\int_{E^c_{M_{\eta_2}}} |g(x)|\lambda(dx) \leq \varepsilon$$

Following lemma 4.4, we can find $\eta < \min\{\eta_1,\eta_2\}$ such that $M_\eta\eta < \epsilon$ and

$$\int_{E^c_{M_{\eta}}} |g(x)|\lambda(dx) \leq \varepsilon$$

Then we define $g_k(x) = g(x)\delta_{E_k}(x)$, Since $g_{M_\eta}(x)$ is a bounded measurable function, then we can find an integer $N$ such that:

$$E_{\gamma,x}[|\frac{\sum_{i=1}^N g_{M_\eta}(X_i)}{N}|] < \epsilon, \quad x\in\mathcal{X} \quad \gamma\in\mathcal{Y}$$

Denote $H_n = \{D_n \geq \frac{\eta}{N^2}\}$, then Diminishing Adaptive condition implies that we can find $N_1 \in N$ such that for each $n > N_1$, $P(H_n) \leq \frac{\eta}{N}$ and $\frac{|g(x_*)|\eta^{N_1}}{N(1-\eta)} < \epsilon$. Define the event $E = \bigcap_{i=n+1}^{n+N} H_i^c$. Then when $n > N_1$, we have $P(E^c) < \eta$.For all $n \geq N_1$, following the triangle inequality and induction, on event $E$ we have:

$$sup_{x\in\chi}\|P_{\Gamma_{n+k}}(x,\cdot) - P_{\Gamma_n}(x,\cdot)\| \leq \eta/N, k \leq N$$

In particular, for all $x \in \chi$ and $k - N \leq m \leq k$

$$\|P_{\Gamma_{k-N}}(x,\cdot) - P_{\Gamma_m}(x,\cdot)\| \leq \eta, on \;\; E$$

so $\|P^N_{\Gamma_{k-N}}(x,\cdot) - P(X_k \in \cdot | X_{k-N} = x, G_{k-N})\| \leq \eta \;\; on \;\; E$ for all $x \in \chi$. Then we can construct a second chain $\{X_n^{'}\}_{n=k-N}^{k}$ such that $X_{k-N}^{'} = X_{k-N}$ and $X_n^{'} \sim P_{\Gamma_{k-N}}(X_{n-1}^{'}, \cdot)$ for $k - N + 1 \leq n \leq k$ such that $P(X_k^{'} \neq X_k) \leq \eta$. So for any $n > N_1$, we have the following inequality (*):

$$\begin{aligned}
& E(\frac{1}{N}|\sum_{i=n+1}^{n+N} g(X_i)|X_0 = x_*, \Gamma_0 = \gamma_*) \\
\leq \;& E(E(|\frac{\sum_{i=n+1}^{n+N} g_{M_\eta}(X_i)}{N}||\mathcal{G}_n)|X_0 = x, \Gamma_0 = \gamma) + E(\frac{|\sum_{i=n+1}^{n+N}(g - g_{M_\eta})(X_i)}{N}||X_0, \Gamma_0) \\
\leq \;& E(E_{\Gamma_n, X_n}(|\frac{\sum_{i=1}^{N} g_{M_\eta}(X_i)}{N}|)|X_0, \Gamma_0) + M_\eta\eta + M_\eta P(E^c) + \frac{\sum_{i=n+1}^{n+N} E(|(g - g_\eta)(X_i)||X_0, \Gamma_0)}{N} \\
\leq \;& \epsilon + \epsilon + M_\eta\eta + \frac{\sum_{i=n+1}^{n+N}\int_{E^c_{M_\eta}} |g|(y)||A^{(i)}((x_*, \gamma_*), dy)}{N} \\
\leq \;& \epsilon + \epsilon + \epsilon + \frac{\sum_{i=n+1}^{n+N}\int_{E^c_{M_\eta}} |g|(y)|h^{(i)}_{\gamma_*}(x_*, y)\lambda(dy) + w^{(i)}_{\gamma_*}(x_*)|g(x_*)|}{N} \\
\leq \;& 3\epsilon + \frac{\sum_{i=n+1}^{n+N} L\int_{E^c_{M_\eta}} |g|(y)|\lambda(dy) + \eta^i|g(x_*)|}{N} \\
\leq \;& (3 + L)\epsilon + \frac{|g(x_*)|\eta^{n+1}}{N(1-\eta)} \\
\leq \;& (4 + L)\epsilon \qquad (*)
\end{aligned}$$

Now consider any integer $T$ sufficiently large such that:

$$max[\frac{N_1 Fs + \frac{|g(x_*)|}{1-\eta}}{T}, \frac{NFs + \frac{|g(x_*)|}{1-\eta}}{T}] \;\; \leq \;\; \varepsilon \qquad (4.10)$$

Then we have

$$\begin{aligned} & E(|\frac{\sum_{i=1}^{T} g(X_i)}{T}||X_0 = x_*, \Gamma_0 = \gamma_*) \\ \leq \; & E(|\frac{\sum_{i=1}^{N_1} g(X_i)}{T}||X_0 = x_*, \Gamma_0 = \gamma_*) \\ + \; & E(\frac{1}{\lfloor \frac{T-N_1}{N} \rfloor} \sum_{j=1}^{\lfloor \frac{T-N_1}{N} \rfloor} \frac{1}{N} \sum_{k=1}^{N} g(X_{N_1+(j-1)N+k}|X_0 = x_*, \Gamma_0 = \gamma_*)) \\ + \; & E(|\frac{\sum_{N_1+\lfloor \frac{T-N_1}{N} \rfloor N+1}^{T} g(X_i)}{T}||X_0 = x_*, \Gamma_0 = \gamma_*) \end{aligned}$$

For the first term we have:

$$\begin{aligned} & E(|\frac{\sum_{i=1}^{N_1} g(X_i)}{T}||X_0 = x_*, \Gamma_0 = \gamma_*) \\ \leq \; & \frac{\sum_{i=1}^{N_1} E(|g(X_i)||X_0 = x_*, \Gamma_0 = \gamma_*)}{T} \\ \leq \; & \frac{\sum_{i=1}^{N_1} \int_{\mathcal{X}} |g(y)| A^{(n)}((x_*, \gamma_*), dy)}{T} \\ \leq \; & \frac{\sum_{i=1}^{N_1} \int_{\mathcal{X}} |g(y)| h_{\gamma}^{(n)}(x_*, y)\lambda(dy) + |g(x_*)|\eta^i}{T} \\ \leq \; & \frac{N_1 Fs + \frac{|g(x_*)|}{1-\eta}}{T} \\ \leq \; & \epsilon \end{aligned}$$

and for the third one we know that:

$$\begin{aligned} & E(|\frac{\sum_{N_1+\lfloor \frac{T-N^*}{N} \rfloor N+1}^{T} g(X_i)}{T}|) \\ \leq \; & \frac{\sum_{N^*+\lfloor \frac{T-N^*}{N} \rfloor N+1}^{T} E(|g(X_i)|}{T} \\ \leq \; & \frac{NFs + \frac{|g(x_*)|}{1-\eta}}{T} \\ \leq \; & \epsilon \end{aligned}$$

Finally following the inequality (*), the second term $\leq (4 + L)\epsilon$, so we have

$$E(|\frac{\sum_{i=1}^{T} g(X_i)}{T}|) \quad \leq \quad (6 + L)\epsilon$$

Markov's inequality then gives that

$$P(|T^{-1} \sum_{i=1}^{T} g(X_i)| \geq \varepsilon^{\frac{1}{2}}) \leq (6 + L)\epsilon^{\frac{1}{2}}$$

Since this holds for all sufficiently large $T$, and since $\varepsilon > 0$ is arbitrary, the result follows. □

**Remark:** Here we actually get the conclusion: for any $\epsilon > 0$, $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, there exists $N$ such that for any $n > N$ we have:

$$P(|\frac{\sum_{i=1}^{n} g(X_i)}{n}| > \epsilon) < \epsilon$$

**But** here the "$N$" is **dependent** on the choice of the starting value $x$, but **independent** of the starting value $\gamma$. In fact, this kind of dependence of the starting value is reasonable when $g$ is unbounded. Let us consider the following example which is a general Markov chain with the kernel being uniformly ergodic:

Consider $\mathcal{X} = (0, 1]$, and

$$P(x, A) = \frac{2}{3}\mu(A) + \frac{1}{3}\delta_x(A)$$

where $\mu$ is Lebesgue measure on $(0, 1]$. Since

$$\begin{aligned} \int_{\mathcal{X}} P(x, A)\mu(dx) &= \int_{\mathcal{X}} [\frac{2}{3}\mu(A) + \frac{1}{3}\delta_x(A)]\mu(dx) \\ &= \frac{2}{3}\mu(A) + \frac{1}{3}\mu(A) \\ &= \mu(A) \end{aligned}$$

$\pi$ is stationary with respect to $P(x, \cdot)$. And following that:

$$\|P(x, \cdot) - \pi(\cdot)\|_{var} = \| - \frac{1}{3}\mu(A) + \frac{1}{3}\delta_x(A)\|_{var} \leq \frac{1}{3}$$

Therefore, $P$ is uniformly ergodic with respect to $\mu$. Now suppose $g(x) = x^{-\frac{1}{2}}$, then $\mu(g) = 2$, and then $P(X_1 \in (0, \frac{1}{m^2}]|X_0 = \frac{1}{m^2}) = \frac{2}{3m^2} + \frac{1}{3}$ for each $m \in \mathbf{N}$. Suppose for some $0 < \epsilon < \frac{1}{3}$, there exists $N$ such that $P(|\frac{\sum_{i=1}^{N} g(X_i)}{N}| > \epsilon|X_0 = x_0) < \epsilon$ for all $x_0 \in \mathcal{X}$. If we take $x_0 = (3N)^{-2}$, since $g(X_i) > 0$, we have:

$$\begin{aligned} P(|\frac{\sum_{i=1}^{N} g(X_i)}{N} - \pi(g)| > \epsilon|X_0 = \frac{1}{(3N)^2}) &\geq P(\frac{g(X_1)}{N} - 2 > \epsilon|X_0 = \frac{1}{(3N)^2}) \\ &\geq P(g(X_1) \geq 3N|X_0 = \frac{1}{(3N)^2}) \\ &\geq P(X_1 \leq \frac{1}{(3N)^2}|X_0 = \frac{1}{(3N)^2}) \\ &> \frac{1}{3} \end{aligned}$$

Contradiction!

## 4.3 A Corollary

In Roberts and Rosenthal [4] (2005), they also studied the adaptive MCMC with bounded densities and proved the following corollary:

**Corollary 4.10.** *Suppose an adaptive MCMC algorithm satisfies the Diminishing Adaptation property, and also that each $P_\gamma$ is ergordic for $\pi(\cdot)$. Suppose further that for each $\gamma \in \mathcal{Y}$, $P_\gamma(x, dy) = f_\gamma(x, y)\lambda(dy)$ has a density $f_\gamma(x, y)$ with respect to some finite reference measure $\lambda(\cdot)$ on $\mathcal{X}$. Finally, suppose $f_\gamma(x, y)$ are uniformly bounded, and that for each fixed $y \in \mathcal{X}$, the mapping $(x, \gamma) \mapsto f_\gamma(x, y)$ is continuous with respect to some product metric space topology, with respect to which $\mathcal{X} \times \mathcal{Y}$ is compact. Then $\lim_{n\to\infty} T(x, \gamma, n) = 0$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$.*

We also have the WLLN for the unbounded measurable function $g$ under the same conditions in the corollary 4.10. Actually $P_\gamma(x, A) = \int_A f_\gamma(x, y)\lambda(dy)$ is a special case of $P_\gamma(x, A) = \int_A f_\gamma(x, y)\lambda(dy) + r_\gamma(x)\delta_x(A)$ when $r_\gamma(x) \equiv 0$. We just plug in $\eta = 0$ to the proof of theorem 4.1, then we can prove the following corollary:

**Corollary 4.11.** *Consider an adaptive MCMC that satisfies the conditions in Corollary 4.10, then for any measurable function $g$ such that $\lambda(|g|) < \infty$ and $\pi(g) < \infty$ we have:*

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

*in probability as $n \to \infty$, conditional on $X_0 = x$ and $\Gamma_0 = \gamma$.*

**Remark:**The corollary 4.11 indicates that: for any $\epsilon > 0$, $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$, there exists $N$ such that for any $n > N$ we have:

$$P(|\frac{\sum_{i=1}^{n} g(X_i)}{n}| > \epsilon) < \epsilon$$

However it is not hard to find that such an "$N$" is **independent** of the choice of the initial values $x$ and $\gamma$.

## 4.4 Applications

As an application of theorem 4.1, we will think about the Adaptive Metropolis algorithm of Haario et al. [7] (2001) , in which the target distribution $\pi$ is supported on the subset

$S \subseteq \mathbb{R}^d$ and it has the density $\pi$ with a slight abuse of notation with respect to the Lebesgue measure on $S$.

Now let us state the adaptive algorithm, at $n-$step, we will use the Gaussian distribution $q_n$ with mean at the current point $X_{n-1}$ and covariance $C_n = C_n(X_0, X_1, \cdots, X_{n-1})$ as the proposal distribution, where $C_n$ is defined as following:

$$C_n = \begin{cases} C_0, & n \leq n_0 \\ s_d cov(X_0, \cdots, X_{n-1}) + s_d \epsilon I_d, & n > n_0 \end{cases}$$

Here $s_d$ is a parameter that depends only on dimension $d$ , $\epsilon > 0$ is a constant that we may choose very small compared to the size of $S$, $I_d$ denotes the d-dimensional identity matrix and the initial covariance $C_0$ is an arbitrary strictly positive definite matrix according to our best prior knowledge. Haario et al. [7] (2001) have prove the following Strong Laws of Large Number(SLLN):

**Theorem 4.2.** *Let $\pi$ be the density of a target distribution supported on a bounded measurable subset $S \subseteq \mathbb{R}^d$, and assume that $\pi$ is bounded from above. Let $\epsilon > 0$ and let $\mu_0$ be any initial distribution on $S$. Define the adaptive MCMC as above. Then the AMCMC simulates properly the target distribution $\pi$: for any bounded and measurable function $f : S \to R$, the equality:*

$$\lim_{n\to\infty} \frac{1}{n+1}(f(X_0) + f(X_1) + \cdots + f(X_n)) = \int_S f(x)\pi(dx)$$

*holds almost surely.*

However following our theorem 4.1, we actually can prove that the WLLN holds for any unbounded measurable function $g$ with $\lambda(|g|) < \infty$ where $\lambda$ is Lesbesgue measure.

**Corollary 4.12.** *The WLLN holds for the above adaptive MCMC and any measurable function $g$ satisfying $\lambda(|g|) < \infty$ and $\pi(|g|) < \infty$.*

*Proof.* In this adaptive algorithm, according to the formula (14) in Haario et al. [7] (2001), the parameter space $\mathcal{Y}$ consists of all the $d \times d$ matrix $\gamma$ satisfying that $c_1 I_d \leq \gamma \leq c_2 I_d$ for some $c_1 > 0$ and $c_2 > 0$. If we consider $\mathcal{Y}$ as a $d^2$ vector space and define

the metric on it as $d(\gamma_1, \gamma_2) = \sqrt{\sum_{1 \le i \le j \le d} \left( (\gamma_1)_{ij} - (\gamma_2)_{ij} \right)^2}$. Obviously $\mathcal{Y}$ is compact with respect this metric topology, hence $\mathcal{X} \times \mathcal{Y}$ is also compact. Furthermore since the proposal distribution $Q_\gamma(x, \cdot) = MVN(x, \gamma)$, $P_\gamma$ is ergodic for $\pi(\cdot)$ and the density mapping $(x, \gamma) \to f_\gamma(x, y)$ are continuous and bounded. Therefore following the theorem 4.1 we have the conclusion. □

**Acknowledgements.** We would like to thank Prof. Jeffrey Rosenthal for his assistance in writing this paper.

REFERENCES

[1] Adrieu, C. & Atchade, Y. F. (2005). On the efficiency of adaptive MCMC algorithms

[2] Andrieu, C. & Moulines, E. (2005). On the ergodicity properties of some Adaptive MCMC Algorithms. *To appear Ann. Appl. Probab.*.

[3] Andrieu, C. & Roberts, C. P. (2001). Conrolled MCMC for optimal samling. Technique report. University Paris Dauphine, Ceremade 0125.

[4] Roberts, G. O. & Rosenthal, J. S. (2005). Coupling and ergodicity of adaptive MCMC *Technical Report*, MCMC preprints.

[5] Gilks, W.R. Richardson, S. & Spiegelhalter, D. J. (1996). Markov chain Monte Carlo in practice. *Interdisciplinary Statistics*, Chapman & Hall, London.

[6] Gilks, W.R.Roberts, G. O. & Sahu,S.K. (1998). Adaptive Markov chain Monte Carlo through regeneration.*J. Amer. Statist. Assoc.*, **93**, 1045-1054.

[7] Haario, H. , Saksman, E. & Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli* **7**, 223-242

[8]Atchade, Y. F. & Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithm *Bernoulli*, **11** 815-828.

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

[10] Roberts, G. O. & Rosenthal, J. S. (2005). Example of adaptive MCMC.

[11] Rosenthal, J. S. (2000). First Look at Rigorous Probability Theory.