PROPP-WILSON ALGORITHM AND BEYOND

by

Ümit Işlak

B.S., in Mathematics, Boğaziçi University, 2007

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Mathematics
Boğaziçi University
2010

ACKNOWLEDGEMENTS

I especially want to thank my supervisor, Alp Eden, for his guidance during my master studies. His perpetual energy and enthusiasm in research had motivated me throughout this period. Research life was very smooth and rewarding for me.

Serdar Altok and Taylan Cemgil deserve special thanks for participating in my thesis committee and for their endless support on all sides of my master studies.

I am deeply thankful to Alper Kanar and Mehmet Demir for their true friendship and motivation. I also thank to all of my officemates for such a friendly atmosphere.

I owe my loving thanks to Sevgi.

Most importantly, I am indebted to my family for their self-sacrifice and patience through my education.

ABSTRACT

PROPP-WILSON ALGORITHM AND BEYOND

Propp-Wilson algorithm is a Markov chain Monte Carlo method that produces samples that are drawn exactly from the stationary distribution of a given Markov chain. The aim of this master thesis is to unify the underlying ideas of this algorithm and to embed it into a more general framework. For this purpose, we use coupling theory as the primary tool. We also introduce Letac's principle which states that if the backward process corresponding to a Markov chain converges independent of the initial position, then its limit is distributed according to the stationary distribution of the Markov chain. With Letac's principle, sufficient conditions for the convergence of backward processes become very important and this convergence is usually satisfied with the choice of contractive maps. We detail this with examples and work on a case in which we have contractivity on the average.

ÖZET

PROPP-WILSON ALGORİTMASI VE ÖTESİ

Propp-Wilson algoritması, verilen bir Markov zincirinin durağan dağılımına tam olarak uyan örnekler almamızı sağlayan bir Markov zinciri Monte Carlo metodudur. Bu tezin amacı, bu algoritmanın altında yatan fikirleri bir araya getirmek ve algoritmayı daha genel bir çerçevenin içine oturtmaktır. Eşleşim teorisi temel araç olarak kullanılmaktadır. Ayrıca, bir Markov zincirine denk düşen geri süreç başlangıç pozisyonundan bağımsız olarak yakınsıyorsa, geri sürecin limitinin Markov zincirinin durağan dağılımına göre dağıldığını söyleyen Letac'ın prensibine odaklanılmıştır. Letac'ın prensibi ile beraber, geri sürecin yakınsaması için yeter koşullar önem kazanmaktadır. Bu koşullar genelde çökücü fonksiyonlar üzerinden verilirler. Tezin son bölümünde, bu nokta örneklerle detaylandırılmış ve ortalamada çöken bir durum incelenmiştir.

TABLE OF CONTENTS

LIST OF SYMBOLS/ABBREVIATIONS

$\square$	End of proof
$:=$	Equality that includes a definition
$\subset$	Subset
a^+	Max{a,0}
$\mathfrak{B}(\mathcal{X})$	Borel σ-algebra on $\mathcal{X}$
$BC(\mathcal{X})$	Bounded continuous function on $\mathcal{X}$
d_W	Wasserstein distance
d_K	Kantorovich distance
d_P	Prokhorov distance
δ_x	Kronecker's delta function
$\mathcal{F} \otimes \mathcal{G}$	The product σ-algebra formed from the σ-algebras $\mathcal{F}$ and $\mathcal{G}$
L_X	Law of the random variable X
$\mathcal{L}(\mathcal{X})$	Lipschitz functions on $\mathcal{X}$
$\mathbb{N}$	The set of natural numbers
$(\Theta, \mathfrak{F}, Q)$	A probability space
$\mathrm{P}(\cdot, \cdot)$	Transition probability
$\mathbb{R}^n$	The $n-$dimensional Euclidean space
$\| . \|_\infty$	The sup-norm
$\| . \|_{TV}$	Total variation distance
$U(0,1)$	Uniform distribution over $(0,1)$
$X \sim$	Random variable X is distributed according to
$\mathbb{Z}$	The set of integers
$\mathbb{Z}^+$	The set of non-negative integers
a.s.	Almost surely
CFTP	Coupling from the past
GCFTP	Generalized coupling from the past

GM	Green-Murdoch
i.i.d.	Independent identically distributed
IFS	Iterative Function System
MRF	Markov random field
MCMC	Markov Chain Monte Carlo
PW	Propp-Wilson
ROCFTP	Read once coupling from the past
WFP	Weak Feller property

1. INTRODUCTION

Markov Chain Monte Carlo (MCMC) methods date back to the same time as the development of ordinary Monte Carlo (MC) methods which were introduced by Stanislaw Ulam and John von Neumann in late 1940's ([1]). The rise of these methods were in a close relation with the emergence of computers. In fact, Neumann gave the first such method in 1947 after the appearance of the first computer, ENIAC (Electronic Numerical Integrator And Computer), in 1946. At the same time, Ulam and Neumann invented inversion and accept-reject algorithms to generate random numbers. See [2], [3] and [4] for history and analysis of MC methods.

The first MCMC algorithm is associated with another computer, MANIAC (Mathematical Analyzer, Numerical Integrator and Computer), built in Los Alamos in 1952. This algorithm was published in 1953 in the Journal of Chemical Physics by Metropolis et al ([5]). Their primary focus in this paper is the computation of integrals of the form:

$$I = \frac{\int F(x,y)\exp(-E(x,y)/kT)dxdy}{\int \exp(-E(x,y)/kT)dxdy}$$

where the energy E is defined by

$$E(x,y) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} V(d_{ij})$$

with N being the number of particles, V the potential function and d_{ij} the distance between particles i and j. Since numerical integration was not efficient in computing these integrals due to high dimensionality, they invented a genuine method that resembles the Gibbs Sampler algorithm of modern days.

Nowadays, MCMC methods are used in a variety of problems in diverse fields of applied sciences such as physics, biology and astronomy. Solutions to linear systems of equations, calculations of integrals and many other problems can be expressed in terms of Markov chains and their stationary distributions. See [4] and [6] for introductory books, [7] and [8] for survey papers on MCMC methods.

When solving problems as in the previous paragraph, one eventually faces the hard problem of getting samples from complicated distributions living in high dimensions. Inversion Method and Rejection Sampling are classical methods for getting exact samples from a given distribution which become useless as the dimension of the underlying space increases ([4]). Besides these, there are other methods such as Metropolis-Hastings' algorithm and Gibbs sampling which are widely used to get approximate samples. These latter algorithms are in fact examples of MCMC methods.

The idea behind Metropolis-Hastings' algorithm is running a Markov chain whose stationary distribution is same as the one from which we are trying to sample from. So, once the chain has *evolved enough*, the distribution of the chain will be *close enough* to the stationary distribution of the Markov chain and the problem will be approximately solved. In this methodology, although there are many results on the convergence rates to the stationary distribution in some particular cases, it is in general not possible to be sure that the chain has evolved enough or equivalently that its distribution is close enough. This is called the *'How long is long enough?'* phenomenon in the literature. See, for example, [2] and [9] for results on Metropolis algorithm.

In 1996, James Propp and David Wilson published a sampling algorithm which is called as Coupling from the Past (CFTP) or Propp-Wilson (PW) algorithm that can be used to get exact samples from stationary distributions of finite state space ergodic Markov chains ([10]). Their idea was a combination of a well-known fact from Markov chain theory and a genuine observation on coupling of copies of a Markov chain. Using monotonicity arguments, PW algorithm can be applied to chains with huge state spaces such as Ising model, Hardcore model and more general Markov random fields. Such application areas brought a significant reputation to PW algorithm.

Many modifications and generalizations of PW algorithm were accomplished in a very short time. The natural extension to the Markov chains with continuous state spaces was carried out by Peter Green and Duncan Murdoch in 1998 ([11]). These adaptations are quite restrictive as it is usually impossible to deal with infinitely many copies of a chain. In the same year, James Allen Fill devised an alternative algorithm for exact sampling whose primary aim was preventing the User Impatiance Bias in CFTP ([12]). In short, PW algorithm excited many people from various disciplines and created a new field of study which is now called *Perfect Sampling*.

The underlying idea of Perfect Sampling was already available in 1984 in a paper of Gerard Letac: If the backward process corresponding to a Markov chain (with some properties that will be discussed) converges a.s. independent of the initial position, then its limit is distributed according to the stationary distribution of the Markov chain ([13]). Once the importance of this idea was understood, finding sufficient conditions for the convergence of the backward processes became extremely important. In fact, PW algorithm relies heavily on the observation that backward processes corresponding to finite state space Markov chains converge with probability one independent of the initial position.

In 1999, Persi Diaconis and David Freedman published a survey paper in which they used algebraic tail and global contractivity on the average conditions for the convergence of the backward processes ([14]). This paper convinced many that the theory behind was applicable to areas such as Queueing theory, Image Processing and speeded up the research on both theoretical and applied sides of this subject.

During the first years of 2000's, necessary conditions on the continuity of the update functions of the underlying Markov chains were relaxed with contributions of Örjan Stenflo ([15]). Under quite general conditions, Ö. Stenflo not only shows the convergence of the backward process but also gives results on the convergence rate in terms of the Kantorovich distance between probabilities. See [16] and [17] as well for Stenflo's further contributions to perfect sampling on general state spaces and random iterated function systems.

The fundamental aim of this thesis is to understand the ideas given in [14] and analyze them in a detailed way. The rest of the thesis is organized as follows. In *Chapter 2*, we give some notations, definitions and preliminary results.

In *Chapter 3*, we discuss CFTP algorithm and present a generalization of it which we call *Generalized Coupling From the Past* (GCFTP). We then consider Green-Murdoch (GM) algorithms as special cases of GCFTP. Application to Markov random fields and further comments on CFTP algorithms are also given in this chapter.

In *Chapter 4*, after a short review of forward coupling times, the necessary background on general backward coupling times is given. In particular, we focus on vertical backward coupling times and using this concept, we analyze CFTP algorithms in a wider sense.

Chapter 5 is devoted to a discussion on the theory behind Letac's principle which is closely related to CFTP algorithms. We present theorems on stationary distributions of Markov chains that contract on the average as well.

Appendix A consists of the necessary background on probability metrics. *Appendix B* is devoted to Gibbs Sampler.

2. PRELIMINARIES

In this chapter, we discuss representations of Markov chains on Polish spaces with Iterative Function Systems and explain introductory concepts on Markov chains. Probability metrics are recalled in Appendix A.

2.1. Probability on Polish spaces

We start with a short review of basic facts about Polish spaces and probability measures on Polish spaces to make the thesis self-contained. For the proofs and details, see [18] and [19].

Definition 2.1.1. *A metric space is said to be a* Polish space *if it is separable and if there is an equivalent metric for which it is complete.*

$\mathbb{R}^n, \mathbb{N}$, finite subsets of $\mathbb{R}$, Cantor set and $C([0,1], \|.\|_\infty)$ are examples of Polish spaces. Two very useful properties of probability measures on Polish spaces are given in the following theorems.

Theorem 2.1. *[18] Let $\mathcal{X}$ be a Polish space and μ be a probability measure on $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ where $\mathfrak{B}(\mathcal{X})$ is the Borel σ-algebra on $\mathcal{X}$. Then for every $A \in \mathfrak{B}(\mathcal{X})$ and $\epsilon > 0$, there exist an open set U and a closed set F that satisfy*

$$F \subset A \subset U \subset \mathcal{X} \quad \textit{with} \quad \mu(U - F) < \epsilon.$$

Denote by $BC(\mathcal{X})$ the set of bounded continuous functions on a metric space $\mathcal{X}$.

Theorem 2.2. *[18] If $\mathcal{X}$ is a Polish space and μ_1, μ_2 are probability measures on $\mathfrak{B}(\mathcal{X})$ with*

$$\int_{\mathcal{X}} f(x)\mu_1(dx) = \int_{\mathcal{X}} f(x)\mu_2(dx)$$

for every $f \in BC(\mathcal{X})$, then $\mu_1 = \mu_2$.

2.2. Representation of Markov Chains with Iterative Function Systems

This section is devoted to advancing a way of representing Markov chains that will give us the chance of managing these processes in a more flexible way.

Definition 2.2.1. *Let $(\mathcal{X}, d)$ be a Polish space with its Borel σ-algebra $\mathfrak{B}(\mathcal{X})$ and $(\Theta, \mathfrak{F})$ be a measurable space. Let $f : \mathcal{X} \times \Theta \to \mathcal{X}$ be a jointly measurable function. Writing $f_\theta(x) = f(x, \theta)$ for $\theta \in \Theta$, the set $\{\mathcal{X}; f_\theta, \theta \in \Theta\}$ is called an* Iterative Function System (IFS) *and the mapping f is called an* update function.

Now we wish to discuss processes that are generated using IFSs. For this purpose, let $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ be a Polish space and $(\Theta, \mathfrak{F})$ be a measurable space. Letting $(\theta_n)_{n=0}^{\infty}$ be a stochastic sequence taking values in Θ and fixing an initial point $x \in \mathcal{X}$, we may define a stochastic sequence $(X_n(x))$ by

$$X_0(x) = x \quad \text{and} \quad X_n(x) = (f_{\theta_{n-1}} o ... o f_{\theta_0})(x), \quad n \geq 1. \tag{2.1}$$

The process $(X_n(x))$ which is generated via the IFS $\{\mathcal{X}; f_\theta, \theta \in \Theta\}$ will be called a *stochastically recursive sequence with randomness source* (θ_n) ([20]).

Stochastically recursive sequences are in a very close relation with Markov chains. We begin by recalling the definition of transition probabilities to get into this relation.

Definition 2.2.2. *Let $(\mathcal{X}, \mathfrak{B})$ be a measurable space. A mapping $P : \mathcal{X} \times \mathfrak{B} \to [0, 1]$ is said to be a* transition probability *if for each $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure and for each $A \in \mathfrak{B}$, $P(\cdot, A)$ is $\mathfrak{B}$ measurable.*

It is worth noting that, for a given stochastically recursive sequence $(X_n(x))$ on a Polish space $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ and for any $m \in \mathbb{N}$, the map $P_m : \mathcal{X} \times \mathfrak{B}(\mathcal{X}) \to [0, 1]$ defined by

$$P_m(x, A) = \mathbb{P}(X_m(x) \in A)$$

is a transition probability ([17]).

Throughout this thesis, we will always work on stochastically recursive sequences with independent and identically distributed (i.i.d.) randomness source sequence (θ_n). Such constructed stochastically recursive sequences are in fact Markov chains as we now state and prove for the case where θ_n's are uniformly distributed over $(0, 1)$.

Theorem 2.3. *Let $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ be a Polish space and $(\theta_n)_{n=0}^{\infty}$ be a sequence of independent random variables that are uniformly distributed over $(0, 1)$. Also let the stochastic sequence $(X_n(x))_{n=0}^{\infty}$ be defined as in (2.1). Then $(X_n(x))$ is a Markov chain starting from x with transition probability*

$$P(x, A) = m(\theta : f_\theta(x) \in A), \quad x \in \mathcal{X}, \quad A \in \mathfrak{B}(\mathcal{X})$$

where m is the Lebesgue measure on the Borel $\sigma-$algebra of $(0, 1)$.

Proof. Let $x \in \mathcal{X}$. The process $(X_n(x))$ starts from x by definition of $(X_n(x))$. Let $\{\mathcal{F}_n\}$ be the filtration induced by the process $(X_n(x))$. Then for any $B \in \mathfrak{B}(\mathcal{X})$, we have

$$\begin{aligned}\mathbb{P}(X_n \in B|\mathcal{F}_{n-1}) = \mathbb{P}(f(X_{n-1}, \theta_{n-1}) \in B|\mathcal{F}_{n-1}) &= m(\theta : f(X_{n-1}, \theta) \in B) \\ &= P(X_{n-1}, B)\end{aligned}$$

where the second equality follows from the disintegration theorem. Thus $X_n(x)$ depends on $\mathcal{F}_n$ only through $X_{n-1}(x)$ and so $(X_n(x))$ is a Markov process with transition probability P as asserted. □

Now we face the following important converse problem: Which Markov chains can be represented by stochastically recursive sequences with i.i.d. randomness source? Theorem 2.4 below gives a sufficient criterion for the existence of such a representation. See [22] for further representation theorems.

Recall that a metric space $\mathcal{X}$ is said to be *Borel measurably isomorphic* to a Borel subset of $\mathbb{R}$ when there exists a one-to-one Borel map $\phi : \mathcal{X} \to \mathbb{R}$ for which $M := \phi(\mathcal{X})$ is a Borel subset of $\mathbb{R}$ with the property that $\phi^{-1} : M \to \mathcal{X}$ is also Borel measurable.

Theorem 2.4. *[17] Let P be a transition probability on a metric space $(\mathcal{X}, d)$ which is Borel measurably isomorphic to a Borel subset of $\mathbb{R}$. Then there exists a jointly measurable function $f : \mathcal{X} \times (0,1) \to \mathcal{X}$ such that for any $x \in \mathcal{X}$ and for any $A \in \mathfrak{B}(\mathcal{X})$,*

$$P(x, A) = m(\theta \in (0,1) : f_\theta(x) \in A)$$

where m is the Lebesgue measure on the Borel σ-algebra of $(0,1)$.

Proof. General case will follow once we prove that Markov chains on $\mathcal{X} = \mathbb{R}$ can be represented in the form stated in the theorem. So let $\mathcal{X} = \mathbb{R}$ and define $f : \mathbb{R} \times (0,1) \to \mathbb{R}$ by

$$f(x, \theta) = \inf\{y : P(x, (-\infty, y]) \geq \theta\}.$$

We have,

$$f(x, \theta) > a \Leftrightarrow P(x, (-\infty, a]) < \theta$$

for $x \in \mathcal{X}, \theta \in (0,1)$ and $a \in \mathbb{R}$. Thus, for each fixed $x \in \mathcal{X}$, $f(x, \theta)$ is Borel measurable in θ by the measurability of the transition probability P. Also since

$$\begin{aligned} m(\theta \in (0,1) : f_\theta(x) > a) &= m(\theta \in (0,1) : P(x, (-\infty, a]) < \theta) \\ &= 1 - P(x, (-\infty, a]) \\ &= P(x, (a, \infty)) \end{aligned}$$

and sets of the form (a, ∞) generate the Borel σ-algebra on $\mathbb{R}$ ([23]), it follows that

$$P(x, A) = m(\theta \in (0,1) : f_\theta(x) \in A)$$

for any $x \in \mathcal{X}$ and any $A \in \mathfrak{B}(\mathbb{R})$.

Next observe that for fixed $\theta \in (0,1)$,

$$\{x \in \mathbb{R} : f_\theta(x) > a\} = \{x \in \mathbb{R} : P(x, (-\infty, a]) < \theta\},$$

is a Borel set since P is measurable in its first coordinate once the second coordinate is fixed. So $f_\theta : \mathbb{R} \to \mathbb{R}$ is a Borel map.

For the case $\mathcal{X} = \mathbb{R}$, it only remains to prove that f is jointly measurable. Firstly we note that for fixed $x \in \mathcal{X}$, $f(x,\theta) = f_\theta(x) : (0,1) \to \mathbb{R}$ is nondecreasing and left continuous. Now set $f(x,0) = -\infty$ and define for $n \geq 1$,

$$f_n(x,\theta) = f(x, \frac{j}{n}), \quad \text{when} \quad \frac{j}{n} \leq \theta < \frac{j+1}{n}, \quad j = 0, 1, ..., n-1.$$

Now, for any $A \in \mathfrak{B}(\mathbb{R})$,

$$\begin{aligned} \{(x,\theta) : f_n(x,\theta) \in A\} &= \bigcup_{j=0}^{n-1} \{(x,\theta) : f(x, \frac{j}{n}) \in A, \frac{j}{n} \leq \theta < \frac{j+1}{n}\} \\ &= \bigcup_{j=0}^{n-1} \left(\{(x,\theta) : f(x, \frac{j}{n}) \in A\} \cap \{(x,\theta) : \frac{j}{n} \leq \theta < \frac{j+1}{n}\} \right) \end{aligned}$$

is a Borel subset of $\mathbb{R}^2$. Since $f(x,\theta)$ is left continuous in θ, it follows that

$$f(x,\theta) = \lim_{n \to \infty} f_n(x,\theta)$$

for all (x,θ) and thus

$$\{(x,\theta) : f(x,\theta) \in (a,\infty)\} = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \{(x,\theta) : f_n(x,\theta) \in (a,\infty)\},$$

for any $a \in \mathbb{R}$. Therefore f is jointly measurable. This completes the proof when $\mathcal{X} = \mathbb{R}$.

Now assume that the state space $\mathcal{X}$ is Borel measurably isomorphic to a Borel subset of $\mathbb{R}$ and let $\phi : \mathcal{X} \to \mathbb{R}$ be a one-to-one Borel map for which $M := \phi(\mathcal{X})$ is a Borel subset of $\mathbb{R}$ with the property that $\phi^{-1} : M \to \mathcal{X}$ is also Borel measurable.

Suppose that $\psi : \mathbb{R} \to \mathcal{X}$ equals ϕ^{-1} on M and maps $\mathbb{R} - M$ on some fixed point $x' \in \mathcal{X}$. For each $x \in \mathbb{R}$ and $B \in \mathfrak{B}(\mathbb{R})$, define $\widetilde{P}(x, B) = P(\psi(x), \phi^{-1}(B \cap M))$. Then $\widetilde{P}$ is a transition probability on $\mathbb{R}$.

We define $g : \mathbb{R} \times (0, 1) \to \mathbb{R}$ by $g(x, \theta) = \inf\{y : \widetilde{P}(x, (-\infty, y]) \geq \theta\}$. First part of the proof shows that g is jointly measurable. Next letting $f(x, \theta) = \psi(g(\phi(x), \theta))$, f turns out to be jointly measurable and for any measurable subset A of $\mathcal{X}$ we have

$$\begin{aligned} m(\theta : f(x,\theta) \in A) = m(\theta : \psi(g(\phi(x),\theta)) \in A) &= m(\theta : g(\phi(x),\theta) \in \psi^{-1}(A)) \\ &= \widetilde{P}(\phi(x), \psi^{-1}(A)) \\ &= P(\psi(\phi(x)), \phi^{-1}(\psi^{-1}(A) \cap M)) \\ &= P(x, A). \end{aligned}$$

This completes the proof. □

Corollary 2.5. *Any Markov chain on a Polish space $\mathcal{X}$ can be represented by an IFS $\{\mathcal{X}, f_\theta, \theta \in \Theta\}$ with the randomness source sequence, (θ_n), being independent and uniformly distributed over* $(0, 1)$.

Proof. Any Polish space is Borel measurably isomorphic to a Borel subset of $\mathbb{R}$ ([24], [25]). Result follows from Theorem 2.4. □

Remark 2.6. *Note that there is no uniqueness claim in Theorem 2.4. In fact, choice of an appropriate IFS plays a crucial role in many problems.*

2.3. Definitions and Basic Results Related to Markov Chains

For this section, $(\mathcal{X}, d)$ is a Polish space with its Borel σ-algebra $\mathfrak{B}(\mathcal{X})$ and (X_n) denotes a Markov chain on $\mathcal{X}$ with transition probability P and corresponding update function $f : \mathcal{X} \times \Theta \to \mathcal{X}$ where $(\Theta, \mathfrak{F}, Q)$ is a probability space.

Definition 2.3.1. *A* stationary distribution π *for the transition probability P with corresponding update function f is the distribution of an $\mathcal{X}$-valued random variable X whose distribution satisfies $L_X = L_{f(X,\theta)}$.*

By a stationary distribution of a Markov chain (X_n), we of course mean a stationary distribution for the transition probability P of (X_n).

Remark 2.7. *If the Markov chain (X_n) is started according to the stationary distribution π, that is, if $X_0 \sim \pi$, then $L_{X_1} = L_{f(X_0,\theta)} = \pi$. Induction reveals $L_{X_n} = \pi$ for every $n \in \mathbb{N}$.*

Definition 2.3.2. *For a Markov chain on a Polish space $\mathcal{X}$ whose IFS representation is given by $\{\mathcal{X}; f_\theta, \theta \in \Theta\}$, we define the* forward process *starting from $x \in \mathcal{X}$ by*

$$F_0(x) = x, \quad F_n(x) = (f_{\theta_{n-1}} o ... o f_{\theta_0})(x), \quad n = 1, 2, 3...$$

and the backward process *starting from $x \in \mathcal{X}$ by*

$$B_0(x) = x, \quad B_n(x) = (f_{\theta_0} o ... o f_{\theta_{n-1}})(x), \quad n = 1, 2, 3...$$

Remark 2.8. *Note that the forward process adds the newly generated randomness at the last step whereas the backward process adds it through the first step. Although these two processes are quite different in nature, we have*

$$L_{B_n(x)} = L_{F_n(x)}, \quad \textit{for} \quad n \in \mathbb{N} \quad \textit{and} \quad x \in \mathcal{X},$$

since (θ_n) is an i.i.d. sequence.

Theorem 2.9. *[13] Consider a Markov chain (X_n) on a Polish space $\mathcal{X}$ with transition probability P whose IFS representation is given by $\{\mathcal{X}; f_\theta, \theta \in \Theta\}$ where f_θ is continuous for each $\theta \in \Theta$. If the corresponding forward process $F_n(x)$ converges weakly to a probability distribution π independent of $x \in \mathcal{X}$, then π is the unique stationary distribution for P.*

Proof. For $g \in BC(\mathcal{X})$ and for every $n \geq 1$ we have,

$$\begin{aligned}\int_{\mathcal{X}} g(y) L_{F_n(x)}(dy) = \mathbb{E}(g(F_n(x))) &= \mathbb{E}(\mathbb{E}(g(F_n(x))|F_{n-1}(x))) \\ &= \int_{\mathcal{X}} \int_{\Theta} g(f(y,\theta))Q(d\theta)L_{F_{n-1}(x)}(dy).\end{aligned}$$

Define $\varphi : \mathcal{X} \to \mathcal{X}$ by $\varphi(y) = \int_{\Theta} g(f(y,\theta))Q(d\theta)$. Then φ is bounded since g is bounded and Q is a probability measure. By the assumption on the continuity of f, we may also conclude that φ is continuous using dominated convergence theorem. Thus $\varphi \in BC(\mathcal{X})$.

Now we may use the weak convergence assumption of the forward process to get

$$\int_{\mathcal{X}} g(y)\pi(dy) = \int_{\mathcal{X}} \int_{\Theta} g(f(y,\theta))Q(d\theta)\pi(dy)$$

by letting $n \to \infty$.

So if X_0 is a random variable distributed according to π and $g \in BC(\mathcal{X})$, we have

$$\int_{\mathcal{X}} g(z)L_{X_0}(dz) = \int_{\mathcal{X}} g(z)L_{f(X_0,\theta)}(dz).$$

Since this is true for all $g \in BC(\mathcal{X})$, we conclude that $L_{X_0} = L_{f(X_0,\theta)}$ holds by Theorem 2.2 and this reveals the required result that π is a stationary distribution for P.

For the uniqueness, suppose that π' is any stationary distribution and $L_{X_0} = \pi'$. Then $L_{F_n(X_0)} = L_{X_0}$ for every $n \geq 1$ and $d_W(L_{F_n(X_0)}, \pi) \to 0$ as $n \to \infty$ by the first part of the proof where d_W is the Wasserstein distance between probability measures (*Appendix I*). Since $L_{F_n(X_0)} = \pi'$ for all n, we conclude that $\pi = \pi'$. □

When the forward process converges weakly to some random variable B, the distribution of B is called the *limiting distribution* of the Markov chain.

Definition 2.3.3. *Let (X_n) be a Markov chain with an IFS representation $\{\mathcal{X}; f_\theta, \theta \in \Theta\}$. (X_n) is said to have* weak Feller property *(WFP) if for any $g \in BC(\mathcal{X})$, the mapping*

$$x \to \mathbb{E}g(f_\theta(x))$$

is continuous.

For more on WFP, see [21]. We now give a corollary of Theorem 2.9 that will be useful in the sequel.

Corollary 2.10. *Theorem 2.9 remains valid when the assumption on the continuity of the update functions is replaced with the WFP assumption for the Markov Chain (X_n).*

Next we briefly discuss uniform ergodicity of Markov chains.

Definition 2.3.4. *Let (X_n) be a Markov chain on a Polish space $\mathcal{X}$ with transition probability P and unique stationary distribution π. The chain is said to be* uniformly ergodic *when*

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \|P^n(x,\cdot) - \pi\|_{TV} = 0.$$

Theorem 2.11. *[26] The following conditions are equivalent for a Markov chain (X_n) on a Polish space $\mathcal{X}$ with transition probability P and unique stationary distribution π:*

(i) (X_n) is uniformly ergodic.

(ii) There exists $c \in (0,\infty)$ and $\lambda \in (0,1)$ such that

$$\|P^n(x,\cdot) - \pi\|_{TV} < c\lambda^n, \quad \forall x \in \mathcal{X}, \quad \forall n \in \mathbb{Z}^+.$$

Theorem 2.12. *[26] Let (X_n) be a Markov chain on a Polish space $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ with transition probability P. If (X_n) is uniformly ergodic, then there exists a probability measure Φ on $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$, $m \in \mathbb{Z}^+$ and $\beta \in (0,1]$ that satisfy $P^m(x,\cdot) \geq \beta\Phi(\cdot)$ for all $x \in \mathcal{X}$.*

Remark 2.13. *Markov chains having the necessary criterion for uniform ergodicity given in Theorem 2.12 are called* Doeblin chains. *These special processes will have great importance throughout this thesis.*

3. PROPP-WILSON ALGORITHM

3.1. Introduction and Illustration of the Basic Idea

Consider an ergodic Markov chain on a finite state space $\mathcal{X}$. Since $|\mathcal{X}|$ is finite, ergodicity assures the existence of a unique stationary (and limiting) distribution which we call π ([27]). Let $f : \mathcal{X} \times (0,1) \to \mathcal{X}$ be one of the possible update functions for this Markov chain.

Our aim is getting exact and independent samples from π. As a starting point, one may think of starting $|\mathcal{X}|$ copies of the underlying chain from each possible initial state and update these chains according to f with the same random numbers over (0,1). It is worth knowing whether the state at which all of these chains coalesce is distributed according to π as the effect of initial position is swept. Let's see that this is not the case with the following example:

Example 3.1. Consider a Markov chain with 3 states that has the following transition matrix,

$$\mathbf{P} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

This Markov chain is ergodic and its stationary distribution is given by $\pi = (0.5, 0.25, 0.25)$. But the three chains initiated from three different states can never coalesce at state 3. Hence we would get biased samples if we had used the described strategy directly.

What PW algorithm does to get exact samples is reversing the above procedure ([10]). This can be formulated as follows. At time 1, generate $\theta_{-1} \sim U(0,1)$ and start Markov chains from all possible initial states and let them evolve using θ_{-1}. If they have coalesced at time 0, take the common value at time 0 as a stationary pick. If not, generate

$\theta_{-2} \sim U(0,1)$, start Markov chains from all possible initial states and let them evolve using θ_{-2} and θ_{-1}. If they have coalesced up to time 0, take the common value at time 0 as a sample. If not, do the same until you get an output. Main result of next section assures that the returned value is in fact a *perfect sample*; that is, the returned value is exactly distributed according to π. Generalizations of this scenario will also be considered in following sections. Figure 3.1 below illustrates PW algorithm for a Markov chain with 4 states.

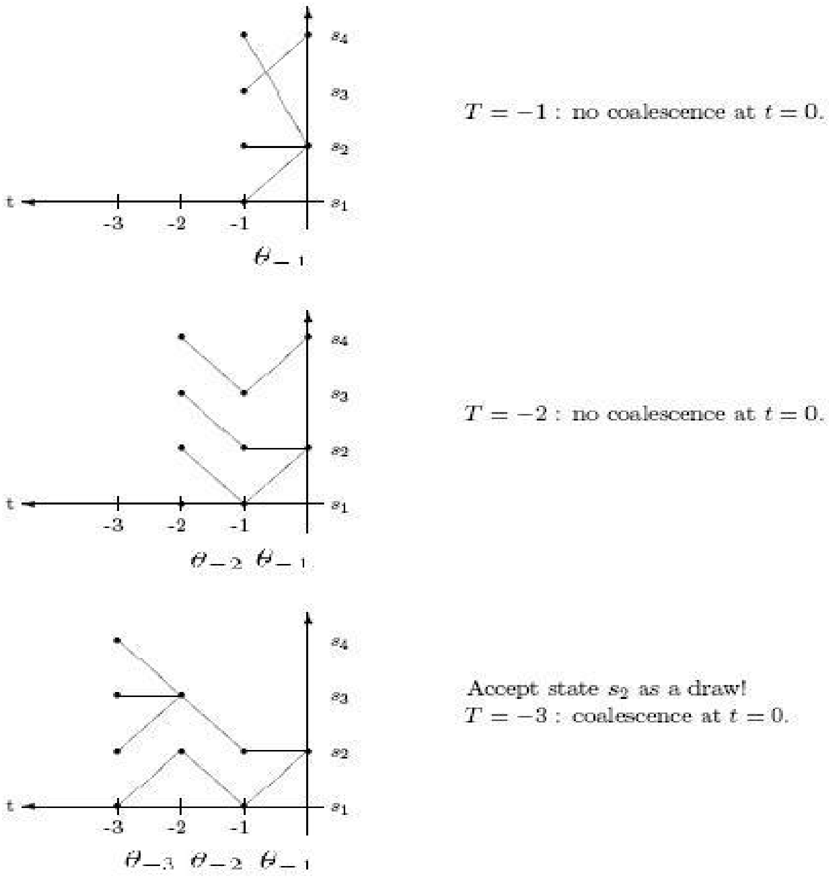
s_4
s_3
s_2
s_1
t
-3
-2
-1
θ_{-1}
$T = -1$: no coalescence at $t = 0$.
θ_{-2} θ_{-1}
$T = -2$: no coalescence at $t = 0$.
θ_{-3} θ_{-2} θ_{-1}
Accept state s_2 as a draw!
$T = -3$: coalescence at $t = 0$.

Figure 3.1. Illustration of how perfect sampling works for a Markov chain with 4 states. Coalescence occurs at $T = -3$.

3.2. Coupling from the Past Algorithm

This section is devoted to CFTP algorithm of James Propp and David Wilson ([10]). Advantages of monotonicity arguments is discussed and a toy example is presented. Note that CFTP algorithm is also known as Propp-Wilson (PW) algorithm in the literature and both names will be used throughout this thesis.

For this section, (X_n) is a Markov chain on a finite state space $\mathcal{X}$ which may always thought to be $\{1, 2, ...m\}$ for some $m \in \mathbb{Z}^+$. Let $f : \mathcal{X} \times (0, 1) \to \mathcal{X}$ be one of the update functions for (X_n). So once the chain is at $x \in \mathcal{X}$, a random number θ is generated uniformly over $(0, 1)$ and the chain moves to $f(x, \theta)$. Now, the pseudocode of the PW algorithm is as follows.

```
PW(M)
    $t = -M$
    $\mathcal{X}_t = \mathcal{X}$
    while $t < 0$
        $t = t + 1$
        $\mathcal{X}_t = f(\mathcal{X}_{t-1}, \theta_{t-1})$
    if $|\mathcal{X}_0| = 1$ then
        return $x_0 \in \mathcal{X}_0$
    else
        PW(M+1)
```

Figure 3.2. Pseudocode for Propp-Wilson Algorithm

For a rigorous analysis of this algorithm, we need to introduce some notation. For $x \in \mathcal{X}$, consider the backward process given by $B_0(x) = x$ and $B_n(x) = (f_{\theta_{-1}} o ... o f_{\theta_{-n}})(x)$

for $n \in \mathbb{Z}^+$. Also define $\tau = \min\{n \geq 1 : B_n(x) = B_n(y), \forall x, y \in \mathcal{X}\} \leq \infty$. This notation needs an explanation. When there exists an $n \in \mathbb{N}$ with $B_n(x) = B_n(y)$, $\forall x, y \in \mathcal{X}$, we set τ to be the minimum of such n's. If such an n does not exist, we set $\tau = \infty$.

In this setting, when τ is finite, τ is the coalescence time of the Markov chains initiated from all possible states and the common value $B_\tau(x)$ with $x \in \mathcal{X}$ is the returned value of PW algorithm.

Theorem 3.2. *[10] (PW algorithm) Let (X_n) be an ergodic Markov chain on a finite state space $\mathcal{X}$ with stationary distribution π and let f be an update function for (X_n). Define τ and $B_n(x)$ as above and suppose that $\tau < \infty$ a.s.. Then, $B_\tau(x)$ is distributed according to π for any $x \in \mathcal{X}$.*

Proof. As τ is finite a.s., $B_n(x) = (f_{\theta_{-1}} o ... o f_{\theta_{-n}})(x)$ becomes constant at a finite time a.s.. So $B_n(x)$ converges a.s. to some random limit independent of x which we call B. This gives $L_{F_n(x)} = L_{B_n(x)} \rightarrow L_B$. That is, $F_n(x)$ converges weakly to B independent of x. By Theorem 2.9, we know that L_B is the stationary distribution of (X_n). Since an ergodic Markov chain on a finite state space has a unique stationary distribution, we conclude $L_B = \pi$.

We also have $B = \lim_{n\to\infty} B_n(x) = B_\tau(x)$ a.s.. Hence $B_\tau(x)$ is distributed according to π independent of x as asserted. □

Note that the idea behind PW algorithm can be seen as a special case of backward coupling times or Letac's principle. More on these topics will emerge in the next two chapters.

Remark 3.3. *Note that the arguments carried out in the proof of 3.2 remains valid when the state space of the underlying Markov chain is replaced by a Polish space and a unique stationary distribution exists for the chain.*

Remark 3.4. *The sequence $1, 2, ..., M, M + 1, ...$ in PW algorithm can be replaced by $1, 2, 4, ..., M, 2M$ or in fact by any increasing sequence. This follows from the proof PW algorithm directly. This observation is very useful in B. Wilson's* Read Once Coupling from the Past Algorithm (ROCFTP) *which is a modified version of the CFTP algorithm ([31]).*

Now we present a toy example for PW algorithm.

Example 3.5. Consider a Markov chain (X_n) with state space $\{0, 1, 2\}$ whose transition matrix is given by

$$\boldsymbol{P} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \end{pmatrix}.$$

An update function for (X_n) can be given by,

$$X_n = f(X_{n-1}, \theta_{n-1}) = \begin{cases} \max\{X_{n-1} - 1, 0\} & \textit{if } \theta_{n-1} \in (0, 1/2] \\ \min\{X_{n-1} + 1, 2\} & \textit{if } \theta_{n-1} \in (1/2, 1) \end{cases}$$

Now PW algorithm works for this chain as follows: Firstly, start 3 chains from 3 different states of (X_n) at $T = -1$ and check for coalescence at $t = 0$. If coalescence occurs, the common state at $t = 0$ is accepted as a sample from the stationary distribution. Otherwise, the starting time is moved to $T = -2$, and the chains are evolved and again checked for coalescence at $t = 0$. If coalescence occurs, the state of the chain at $t = 0$ is accepted as a sample from the desired distribution. The whole process is repeated until the 3 chains coalesce (which happens a.s. with the chosen update function). When the coalescence occurs, the state of the chain at $t = 0$ is taken as a sample from the stationary distribution $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

A crucial part of PW algorithm is the use of stochastic monotonicity. It becomes impossible to deal with Markov chains with huge state spaces without monotonicity.

Definition 3.2.1. *An update function f of a Markov chain (X_n) on a partially ordered state space $(\mathcal{X}, \leq)$ is said to be a* monotone update function *if for any $\theta \in (0, 1)$, the* *inequality*

$$f(x,\theta) \leq f(y,\theta)$$

holds whenever $x \leq y$.

For a detailed discussion of stochastic monotonicity and monotone update functions, see [20].

Theorem 3.6. *Consider an ergodic Markov chain (X_n) on a finite linearly ordered state space $(\mathcal{X}, \leq)$ with a monotone update function f and let $\underline{x}, \overline{x}$ be the minimal and maximal elements in $\mathcal{X}$ respectively. Define $\tau_e = \min\{n : B_n(\underline{x}) = B_n(\overline{x})\} \leq \infty$ where $B_n(x) = (f_{\theta_{-1}} o ... o f_{\theta_{-n}})(x)$. If $\tau_e < \infty$ a.s., then $B_{\tau_e}(x)$ is distributed according to the stationary distribution π of (X_n) for any $x \in \mathcal{X}$.*

Proof. Firstly observe that

$$B_n(\underline{x}) \leq B_n(x) \leq B_n(\overline{x}) \tag{3.1}$$

for any $x \in \mathcal{X}$ since f is given to be monotone.

Next, defining $\tau = \min\{n : B_n(x) = B_n(y), \forall x, y \in \mathcal{X}\} \leq \infty$, we clearly have $\tau_e \leq \tau$. Also when $B_n(\overline{x}) = B_n(\underline{x})$, (3.1) reveals that

$$B_n(x) = B_n(\overline{x}) = B_n(\underline{x})$$

for any $x \in \mathcal{X}$. This gives $\tau \leq \tau_e$ from which we conclude that $\tau = \tau_e$.

So, $B_{\tau_e}(x) = B_\tau(x)$ for any $x \in \mathcal{X}$. Since $B_\tau(x)$ is distributed according to π by Theorem 3.2, result follows. □

Now going back to Example 3.5, with the trivial ordering $0 \leq 1 \leq 2$ on $\mathcal{X}$, the given update function is monotone. That is $f(0,\theta) \leq f(1,\theta) \leq f(2,\theta)$ for every $\theta \in (0,1)$. Hence when applying the PW algorithm to this case, it will be enough to check the coalescence of the Markov chains that are initiated from the states 0 and 2. Although monotonicity does not help too much for this toy example, it is certainly inevitable for Markov chains on huge state spaces. In section 3.4, we shall show the use of monotonicity on Markov random fields.

Remark 3.7. *An immediate corollary for Theorem 3.6 can be given by replacing the linear order by a partial order. In this case, the coalescence of the chains starting from various minimal and maximal elements should be checked.*

Remark 3.8. *Choice of an inconvenient IFS may turn PW into a useless algorithm. This is best illustrated with an example. Consider a 2-state Markov chain* (X_n) *with state space* $\mathcal{X} = \{1, 2\}$ *and transition matrix* $\boldsymbol{P} = \begin{pmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{pmatrix}$ *An update function for* (X_n) *can be given by:* $f(1, \theta) = 1$, $f(2, \theta) = 2$ *if* $\theta \in (0, 1/3)$ *and* $f(1, \theta) = 2$, $f(2, \theta) = 1$ *otherwise. We can not use PW algorithm with this update function since chains starting from different states can never coalesce.*

3.3. Generalized Coupling from the Past

In general, CFTP algorithm can not be applied to Markov chains with general state spaces in its original form as it is impossible to deal with the coalescence of infinitely many chains. For these cases, we present a variation of PW algorithm which we call *Generalized Coupling from the Past (GCFTP).*

Let $\mathcal{X}$ be a Polish space and (X_n) be a Markov chain on $\mathcal{X}$ with an update function $f : \mathcal{X} \times (0, 1) \to \mathcal{X}$ and a unique stationary distribution π. Following the PW algorithm given in the previous section, we define $\tau = \min\{n : B_n(x) = B_n(y), \forall x, y \in \mathcal{X}\} \leq \infty$ where $B_n(x) = (f_{\theta_{-1}} o ... o f_{\theta_{-n}})(x)$ for $n \in \mathbb{Z}^+$ as before. We already know that $B_\tau(x) \sim \pi$ when τ is finite a.s. by Theorem 3.2. For the cases of more general state spaces, we need to define another random time besides τ. For this purpose, firstly consider the following pseudocode.

```
GCFTP(M)
    $t = -M$
    $\mathcal{Y}_t = \mathcal{X}$
    while $t < 0$
        $t = t + 1$
        $\mathcal{Y}_t$ = a set containing $f(\mathcal{Y}_{t-1}, \theta_{t-1})$
        if $|\mathcal{Y}_0| = 1$ then
            return $x_0 \in \mathcal{Y}_0$
        else
            GCFTP(M+1)
```

Figure 3.3. Pseudocode for Generalized Coupling from the Past

Now, let f' be the update mechanism described in the pseudocode so that $\mathcal{Y}_t = f'(\mathcal{Y}_{t-1}, \theta_{t-1})$. We write $f'_{\theta_{t-1}}(\mathcal{Y}_{t-1})$ for $f'(\mathcal{Y}_{t-1}, \theta_{t-1})$ as usual. Next define

$$T = \min\{m \geq 1 : |(f'_{\theta_{-1}} o ... o f'_{\theta_{-m}})(\mathcal{X})| = 1\} \leq \infty.$$

Note that $f_\theta(A) \subset f'_\theta(A)$ for any $A \subset \mathcal{X}$. So, $\tau \leq T$.

Theorem 3.9. *(GCFTP) Let (X_n) be a Markov chain on a Polish space $\mathcal{X}$ with a unique stationary distribution π. Let T and $B_n(x)$ be defined as above for $x \in \mathcal{X}$ and suppose that $T < \infty$ a.s.. Then, $B_T(x)$ is distributed according to π.*

Proof. Firstly observe that $\tau \leq T$. Now since for any $m \geq \tau$, $B_m(x) = B_\tau(x)$, we get $B_T(x) = B_\tau(x)$ for any $x \in \mathcal{X}$. This implies that $B_T(x) \sim \pi$ by Theorem 3.2 (See Remark 3.3). □

Appropriate choice of the update mechanism f' helps us to have control over the uncoalesced chains and reveal PW-type algorithms on Polish state spaces. We now show the use of GCFTP with the aid of Multigamma Coupler. See [11] and [28] for similar examples.

Example 3.10. [11] Let (X_n) be a Markov chain with state space $\mathcal{X} = \mathbb{R}$ and transition probability density $p(.|x)$. Suppose that

$$p(y|x) \geq g(y), \qquad \forall x, y \in \mathcal{X},$$

is satisfied for some positive continuous function g with $\rho := \int_{\mathbb{R}} g(y)dy \in (0,1)$ so that (X_n) is a very special case of Doeblin chains. Also define

$$G(y) = \rho^{-1} \int_{-\infty}^{y} g(v)dv \quad \text{and} \quad Q(y|x) = (1-\rho)^{-1} \int_{-\infty}^{y} (p(v|x) - g(v))dv.$$

An update function for (X_n) can be given by

$$f(x, (\theta^1, \theta^2)) = \begin{cases} G^{-1}(\theta^2) & , \text{ if } \theta^1 < \rho \\ Q^{-1}(\theta^2|x) & , \text{ otherwise} \end{cases}$$

where θ^1, θ^2 are independent random numbers uniformly distributed over (0,1). To see that this is the case, we observe that

$$\begin{aligned} \mathbb{P}(f(x,\theta) \leq y) &= \rho\mathbb{P}(G^{-1}(\theta^2) \leq y) + (1-\rho)\mathbb{P}(Q^{-1}(\theta^2|x) \leq y) \\ &= \rho\mathbb{P}(\theta^2 \leq G(y)) + (1-\rho)\mathbb{P}(\theta^2 \leq Q(y|x)) \\ &= \rho G(y) + (1-\rho)Q(y|x) \\ &= \int_{-\infty}^{y} g(v)dv + \int_{-\infty}^{y} (p(v|x) - g(v))dv \\ &= \mathbb{P}(X_{n+1} \leq y | X_n = x) \end{aligned}$$

as required. Using this update function, we shall now give the pseudocode for the multigamma coupler as stated in [11].

```
Multigamma($M$):
    $t = -M$
    $\mathcal{Y}_t = \mathcal{X}$
    while $\mathcal{Y}_t$ infinite and $t < 0$
        $t = t + 1$
        if $\theta_t^1 < \rho$ then
            $\mathcal{Y}_t = \{G^{-1}(\theta_t^2)\}$
        else
            $\mathcal{Y}_t = \mathcal{X}$
    while $t < 0$
        $t = t + 1$
        if $\theta_t^1 < \rho$ then
            $\mathcal{Y}_t = \{G^{-1}(\theta_t^2)\}$
        else
            $\mathcal{Y}_t = \{Q^{-1}(\theta_t^2|x)\}$ where $\mathcal{Y}_{t-1} = \{x\}$
    if $|\mathcal{Y}_0| = 1$ then
        return $x_0 \in \mathcal{Y}_0$
    else
        Multigamma($M + 1$)
```

Figure 3.4. Pseudocode for Multigamma Coupler

To show that the Multigamma Coupler returns a value that is distributed according to the stationary distribution of the underlying Markov chain, we shall make use of GCFTP. So we need to show that: (i) $f(\mathcal{Y}_{t-1}, \theta) \subset \mathcal{Y}_t$ and (ii) T is finite a.s.

First of these assertions follows directly from the choice of the sets $\mathcal{Y}_t$. For the second one we observe that we will get an outcome whenever we have $\theta_k^1 < \rho$ for some $k \in \mathbb{N}$ and the probability that we do not have any $\theta_k^1 < \rho$ for $k = 1, ...n$ equals $(1-\rho)^n$. So the probability that we will not have an outcome at a finite time is $\lim_{n\to\infty}(1-\rho)^n = 0$ from which (ii) follows.

Remark 3.11. In practice it is in general impossible to use multigamma coupler for perfect sampling. The main reasons for this are:
(i) The normalized update densities are not usually known by the user.
(ii) It is not possible to find a suitable lower bound function g in many cases.

3.4. More on Coupling from the Past Algorithm

3.4.1. An Application on Markov Random Fields

In this section, we describe getting perfect samples from Markov random fields(MRF). See [29] and [30] for a detailed discussion of MRFs. We start by giving basic definitions on graph theory and random fields. For the following, let $G = (V, E)$ be a finite graph and $\mathcal{X}$ be a finite set.

For $v, w \in V$, write $v \sim w$ if there exists $e \in E$ connecting v and w. Also write $< v, w >$ for an edge in E connecting v and w. For $W \subset V$, we define the *boundary of* W, ∂W, to be $\partial W = \{v \in V - W : \exists w \in W \, \text{such that} \, v \sim w\}$.

A *random field* on V with values in $\mathcal{X}$ is a collection $X = \{X(v)\}_{v\in V}$ of random variables with each of $X(v)$ taking values in $\mathcal{X}$. Note that a random field can be regarded as a random variable taking its values in the *configuration space* $\mathcal{X}^V$.

A *configuration* $\xi \in \mathcal{X}^V$ is of the form $(\xi(v) : v \in V)$ where $\xi(v) \in \mathcal{X}$ for $v \in V$. For a given configuration $\xi \in \mathcal{X}^V$ and a given subset $W \subset V$, define $\xi(W) = (\xi(v) : v \in W)$ to be the *restriction of* ξ *to* W.

Definition 3.4.1. *A random field X on a finite graph $G = (V, E)$ is said to be a* Markov random field (MRF) *with distribution π if for any configuration ξ, we have $\pi(X = \xi) > 0$ and*

$$\pi(X(W) = \xi(W)|X(V - W) = \xi(V - W)) = \mathbb{P}(X(W) = \xi(W)|X(\partial W) = \xi(\partial W))$$

where W is any subset of V.

Here we will focus on a particular example of MRFs, namely, Ising model which was introduced by Ernst Ising in 1925. In Ising's finite model, $\mathcal{X} = \{-1, 1\}$, $V = \mathbb{Z}_m^2$ and the underlying neighborhood system is the nearest neighborhood. The *energy of the system* at some configuration $\xi \in \{-1, 1\}^V$ is given by

$$E(\xi) = \frac{-J}{k} \sum_{<x,y>\in E} \xi(x)\xi(y) - \frac{H}{k} \sum_{x \in V} \xi(x)$$

where k is the Boltzman constant, J is the internal energy of an elementary magnetic dipole and H is the external magnetic field. The *(Gibbs) distribution* π on $\mathcal{X}^V$ in terms of the energy function is now given by $\pi(\xi) = \frac{1}{Z}e^{-E(\xi)}$ where Z is the normalizing constant. For the following, we assume that there is no external magnetic field, that is $H = 0$. Also we set $\beta = \frac{-J}{k}$ so that $\pi = \pi_\beta$ is given by

$$\pi_\beta(\xi) = \frac{1}{Z_\beta} e^{\beta \sum_{<x,y>\in E} \xi(x)\xi(y)}, \quad \xi \in \{-1, 1\}^V \tag{3.2}$$

where Z_β is the corresponding normalizing constant. Now to see that π_β defines a MRF, we observe that the right hand side of equation (3.2) can be factorized into factors involving only the states of neighbor vertices ([29]).

For a perfect sampling algorithm for the Ising model, we follow the steps sketched in [30] and use random Gibbs sampler (Appendix B) to form a Markov chain on $\mathcal{X}^V$ having π as its stationary distribution. Such a construction requires the knowledge of full conditionals of π.

We denote the number of positive and negative neighbors of some $v \in V$ when the configuration is ξ by $P(v,\xi)$ and $N(v,\xi)$ respectively.

Lemma 3.12. *Full conditional distributions of π_β are given by*

$$\pi_\beta(X(v)=1|X(V-\{v\})=\xi(V-\{v\}))=\frac{\exp(2\beta(P(v,\xi)-N(v,\xi)))}{1+\exp(2\beta(P(v,\xi)-N(v,\xi)))}$$

where ξ is a given configuration and v is any element in V.

Proof. Fix $v \in V$. Define $\xi^+ \in \{-1,1\}^V$ to be the configuration that agrees with ξ on $V-\{v\}$ and takes the value 1 at v. Similarly define ξ^- to be the configuration that agrees with ξ on $V-\{v\}$ and takes the value -1 at v. We have

$$\begin{aligned}
\frac{\pi_\beta(\xi^+)}{\pi_\beta(\xi^-)} &= \frac{\exp\left(\beta\sum_{<x,y>\in E}\xi^+(x)\xi^+(y)\right)}{\exp\left(\beta\sum_{<x,y>\in E}\xi^-(x)\xi^-(y)\right)}\\
&= \exp\left(\beta\left(\sum_{<x,y>\in E}\xi^+(x)\xi^+(y)-\sum_{<x,y>\in E}\xi^-(x)\xi^-(y)\right)\right)\\
&= \exp\left(\beta\left(\sum_{<v,y>\in E}\xi^+(v)\xi^+(y)-\xi^-(v)\xi^-(y)\right)\right)\\
&= \exp\left(\beta\sum_{<v,y>\in E}(\xi(y)+\xi(y))\right)\\
&= \exp(2\beta(P(v,\xi)-N(v,\xi))).
\end{aligned}$$

Using this we get

$$\begin{aligned}
\pi_\beta(X(v)=1|X(V\backslash\{v\})=\xi(V\backslash\{v\})) &= \frac{\pi_\beta(X(v)=1,X(V\backslash\{v\})=\xi(V\backslash\{v\}))}{\pi_\beta(X(V\backslash\{v\})=\xi(V\backslash\{v\}))}\\
&= \frac{\pi_\beta(\xi^+)}{\pi_\beta(\xi^+)+\pi_\beta(\xi^-)}\\
&= \frac{\exp(2\beta(P(v,\xi)-N(v,\xi)))}{1+\exp(2\beta(P(v,\xi)-N(v,\xi)))}
\end{aligned}$$

from which the required result follows. □

Using these full conditionals, the Gibbs sampler can be constructed with the following update rule: Given X_n, obtain X_{n+1} by picking a vertex $v \in V$ at random, picking $X_{n+1}(v)$ according to the full conditionals and leaving all other vertices unchanged (Appendix B). The updating mechanism can be realized by choosing a random number $\theta_n \sim U(0,1)$ and setting

$$X_{n+1}(v) = 1 \quad \text{if} \quad \theta_n < \frac{\exp(2\beta(P(v,\xi) - N(v,\xi)))}{1 + \exp(2\beta(P(v,\xi) - N(v,\xi)))}$$

and setting $X_{n+1}(v) = -1$ otherwise.

We can now construct a CFTP algorithm based on the Gibbs sampler by using 2^{m^2} chains starting from all possible initial configurations and checking their coalescence. But of course using so many chains bring computational burden when m is large. To avoid this, we establish an order on the configuration space.

For $\xi_1, \xi_2 \in \{-1,1\}^V$, write $\xi_1 \leq_m \xi_2$ if $\xi_1(v) \leq \xi_2(v)$ for all $v \in V$. In this partial ordering, the configuration $\overline{\xi}$ with $\overline{\xi}(v) = 1$ for all $v \in V$ is the maximal configuration and the configuration $\underline{\xi}$ with $\underline{\xi}(v) = -1$ for all $v \in V$ is the minimal configuration.

Now we claim that the update mechanism given above is monotone with respect to $\leq_m$. Let $\xi, \eta \in \{-1,1\}^V$ such that $\xi \leq_m \eta$. Suppose that the randomly chosen vertex is v so that only the value of $X(v)$ changes during the update. We have

$$\frac{\exp(2\beta(P(v,\xi) - N(v,\xi)))}{1 + \exp(2\beta(P(v,\xi) - N(v,\xi)))} \leq \frac{\exp(2\beta(P(v,\eta) - N(v,\eta)))}{1 + \exp(2\beta(P(v,\eta) - N(v,\eta)))}$$

since $\frac{e^a}{e^a+1} \leq \frac{e^b}{e^b+1}$ whenever a and b are real numbers with $a \leq b$. This gives us the required result and so the chains starting from the intermediate states will be sandwiched between the ones starting from the two extreme states. In fact this is exactly why CFTP algorithm can be applied to so many computationally difficult problems in diverse areas.

3.4.2. Drawbacks of Propp-Wilson Algorithm

Although the coalescence times in PW algorithm are finite a.s., they can be arbitrarily large. So the user should take an action and put some bound for the working time of the algorithm. This may cause biased samples especially for Markov chains with large mixing times. This kind of bias is known as the *user-impatience bias*. We explain this with Example 3.5. Recall that (X_n) was a Markov chain on $\mathcal{X} = \{0, 1, 2\}$ with transition matrix

$$\mathbf{P} = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \end{pmatrix}.$$

The stationary distribution of (X_n) is $\pi = (1/3, 1/3, 1/3)$ and a monotone update function for it is given by

$$X_n = f(X_{n-1}, \theta_{n-1}) = \begin{cases} \max\{X_{n-1} - 1, 0\} & \text{if } \theta_{n-1} \in (0, 1/2] \\ \min\{X_{n-1} + 1, 2\} & \text{if } \theta_{n-1} \in (1/2, 1) \end{cases}$$

Now if the user aborts runs not completed in 2 steps, then the output of the algorithm in the 4 possible cases will be as in the following: if $\theta_1 \leq 1/2$ and $\theta_2 \leq 1/2$ then the output is 0, if $\theta_1 \geq 1/2$ and $\theta_2 \geq 1/2$ then the output is 2 and in all other cases coalescence does not occur. Hence in this case we see that PW algorithm returns a biased output with distribution $(1/2, 0, 1/2)$. Note that this example is also the motivation for Fill's algorithm which is another perfect sampling algorithm that takes care of user-impatience bias ([12]).

Another drawback of PW algorithm is the memory problem. Namely, at each iteration of the process, the random numbers generated at previous stages are used again and again until the coalescence occurs. This causes memory problems especially for Markov chains with huge state spaces. This problem is solved by the *Read Once Coupling From The Past Algorithm* of Bruce Wilson ([31]).

4. COUPLING THEORY

In this chapter, we firstly consider some elementary notions from forward coupling theory. Then we introduce the concept of backward coupling times and eventually specialize on vertical backward coupling times which are closely related to the algorithms described in Chapter 3. Our discussion will be mainly based on [32]. A standard reference for coupling theory is [33].

4.1. Forward Coupling

Let $\mathbf{X} = (X_n)$ and $\mathbf{X}' = (X'_n)$ be Markov chains on a Polish space $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ with different initial values x_0 and x'_0, but with the same IFS representation $\{\mathcal{X} : f_\theta, \theta \in (0,1)\}$. They evolve in time via the recursions $X_{n+1} = f(X_n, \theta_n)$ and $X'_{n+1} = f(X'_n, \theta_n)$ for $n \geq 0$ using the same randomness source sequence (θ_n). So if $X_n = X'_n$ for some $n \in \mathbb{N}$, then we necessarily have $X_{n+m} = X'_{n+m}$ for every $m \geq 0$.

Definition 4.1.1. *The* minimal forward coupling time τ *of the Markov chains* $\boldsymbol{X} = (X_n)$ *and* $\boldsymbol{X'} = (X'_n)$ *is defined by*

$$\tau(\boldsymbol{X}, \boldsymbol{X'}) = \min\{n \geq 0 : X_n = X'_n\} \leq \infty.$$

τ *is said to be* successful *if* $\tau(\boldsymbol{X}, \boldsymbol{X'}) < \infty$ *a.s..*

Since it is not possible to detect minimal forward coupling times in most cases, we need to introduce a more general definition for forward coupling times.

Definition 4.1.2. *A random variable* τ *taking values in* $\mathbb{N} \cup \{\infty\}$ *is said to be a* forward coupling time *for the Markov chains* $\boldsymbol{X}$ *and* $\boldsymbol{X'}$ *if*

$$\tau \leq n \Longrightarrow X_{n+m} = X'_{n+m}, \ \forall m \geq 0.$$

Note that, if τ is a forward coupling time and τ' is another random time that satisfies $\tau' \geq \tau$ a.s., then τ' is also a forward coupling time.

Definition 4.1.3. *If $\{\boldsymbol{X}^{(j)}\}_{j\in J}$ is any family of Markov chains with different initial states x_0^j, then*

$$\tau(\{\boldsymbol{X}^{(j)}\}_{j\in J}) := \sup_{i,k\in J} \tau(\boldsymbol{X}^{(i)}, \boldsymbol{X}^{(k)})$$

is said to be the minimal forward coupling time for the family $\{\boldsymbol{X}^{(j)}\}_{j\in J}$.

We now prove the important forward coupling inequality that will play a key role in the sequel.

Theorem 4.1. *(Coupling inequality) Let $\boldsymbol{X} = (X_n)$ be a Markov chain starting from $x_0 \in \mathcal{X}$ with transition kernel P for which a unique stationary distribution π exists. Further let $\boldsymbol{X}' = (X_n')$ be a stationary version of $\boldsymbol{X}$ and τ be the minimal forward coupling time of $\boldsymbol{X}$ and $\boldsymbol{X}'$. Then we have*

$$\|P^n(x_0, \cdot) - \pi\|_{TV} \leq \mathbb{P}(\tau > n).$$

Proof. We define a new stochastic process $\overline{\mathbf{X}}$ via $\mathbf{X}$ and $\mathbf{X}'$ as

$$\overline{X}_n = \begin{cases} X_n' & \text{if } n < \tau \\ X_n & \text{if } n \geq \tau. \end{cases}$$

Then $(\overline{X}_n)$ is a stationary Markov chain with transition kernel P. Also for any $A \in \mathfrak{B}(\mathcal{X})$ we have,

$$\begin{aligned} \mathbb{P}(\overline{X}_n \in A) &= \mathbb{P}(\overline{X}_n \in A | \tau \leq n)\mathbb{P}(\tau \leq n) + \mathbb{P}(\overline{X}_n \in A | \tau > n)\mathbb{P}(\tau > n) \\ &= \mathbb{P}(X_n \in A | \tau \leq n)\mathbb{P}(\tau \leq n) + \mathbb{P}(\overline{X}_n \in A, \tau > n) \end{aligned}$$

and

$$P^n(x_0, A) = \mathbb{P}(X_n \in A) = \mathbb{P}(X_n \in A | \tau \leq n)\mathbb{P}(\tau \leq n) + \mathbb{P}(X_n \in A, \tau > n).$$

The last two equations immediately give

$$|P^n(x_0, A) - \pi(A)| = |\mathbb{P}(X_n \in A, \tau > n) - \mathbb{P}(\overline{X}_n \in A, \tau > n)| \leq \mathbb{P}(\tau > n).$$

Since this is true for each Borel subset A, we get

$$\|P^n(x_0, \cdot) - \pi\|_{TV} \leq \mathbb{P}(\tau > n)$$

from which the required result follows. □

With the aid of coupling inequality, we can get sharp bounds for the rate of convergence to the stationary distribution in the cases where we have information on the tail behavior of the minimal forward coupling time. One such instance can be found below for finite state space Markov chains. See [20] and [32] for further notes on the tail behavior of coupling times.

The following corollary is immediate from the definition of successful minimal forward coupling times.

Corollary 4.2. *If the minimal forward coupling time is successful, then*

$$\|P^n(x_0, .) - \pi\|_{TV} \to 0 \quad as \quad n \to \infty.$$

Next we prove that the $n-$step probabilities of a finite state space ergodic Markov chain converges to the stationary distribution geometrically fast. The idea is finding out a suitable bound for the tail behavior of the minimal forward coupling time and using coupling inequality. Precise statement is as follows.

Theorem 4.3. *Suppose that (M_n) is an ergodic Markov chain on a finite state space $\mathcal{X}$ with $|\mathcal{X}| = M$ whose transition probability is given by P. Then (M_n) is uniformly ergodic.*

Proof. Suppose X_n and X'_n are two independent Markov chains having P as their transition probability. Suppose further that the initial distributions of these chains are $\delta_{\{x\}}$ and π

respectively where x is any fixed element in $\mathcal{X}$ and π is the unique stationary distribution corresponding to P. Also let τ be the minimal forward coupling time of these two chains.

Since the chain is ergodic, there exist $N \in \mathbb{N}$ and $\epsilon > 0$ such that

$$P^N(a,b) > \epsilon, \quad \forall a,b \in \mathcal{X}.$$

Using this observation, we get

$$\begin{aligned}\mathbb{P}(X_N = X'_N) = \sum_{y\in\mathcal{X}} \mathbb{P}(X_N = y, X'_N = y) &= \sum_{y\in\mathcal{X}} P^N(x,y)\mathbb{P}(X'_N = y) \\ &> \epsilon \sum_{y\in\mathcal{X}}\sum_{z\in\mathcal{X}} \mathbb{P}(X'_N = y | X'_0 = z)P(X'_0 = z) \\ &> \epsilon^2 \sum_{y\in\mathcal{X}}\sum_{z\in\mathcal{X}} P(X'_0 = z) \\ &= \epsilon^2 M\end{aligned}$$

and so $\mathbb{P}(X_N \neq X'_N) \leq 1 - \epsilon^2 M$. For any $k \geq 1$, we get

$$\begin{aligned}\mathbb{P}(\tau > kN) &\leq \mathbb{P}(X_N \neq X'_N, ..., X_{kN} \neq X'_{kN}) \\ &= \mathbb{P}(X_N \neq X'_N)\mathbb{P}(X_{2N} \neq X'_{2N} | X_N \neq X'_N) \\ &\qquad ...\mathbb{P}(X_{kN} \neq X'_{kN} | X_{(k-1)N} \neq X'_{(k-1)N}, ..., X_N \neq X'_N) \\ &\leq (1 - \epsilon^2 M)^k\end{aligned}$$

where the last inequality holds since the upper bound for $\mathbb{P}(X_N \neq X'_N)$ is independent of the initial state of (X_n). Now for $n \in \mathbb{N}$, we set $p_n = \max\{m \in \mathbb{N} : mN \leq n\}$. We have

$$\begin{aligned}\mathbb{P}(\tau > n) \leq \mathbb{P}(\tau > p_n N) \leq (1-\epsilon^2 M)^{p_n} &= \frac{(1-\epsilon^2 M)^{\frac{1}{N}((p_n+1)N)}}{1-\epsilon^2 M} \\ &\leq \frac{((1-\epsilon^2 M)^{\frac{1}{N}})^n}{1-\epsilon^2 M}\end{aligned}$$

Now we set $c = \frac{1}{1-\epsilon^2 M}$ and $\lambda = (1-\epsilon^2 M)^{\frac{1}{N}}$ (with ϵ small enough) and use Theorem 2.11 with Theorem 4.1 to get the required result. □

4.2. Backward Coupling

This section is devoted to a study of backward coupling times which will be an intermediate step to treat CFTP algorithms in a more general setting. Consider the probability space $(\Omega, \mathcal{F}, m)$, where $\Omega = (0,1)^{\mathbb{Z}}$, $\mathcal{F}$ is the cylinder σ-algebra and m is the Lebesgue measure. Define the coordinate maps θ_n by $\theta_n(\omega) = \omega_n$ where $\omega = \{\omega_n\}_{n=-\infty}^{\infty} \in \Omega$.

Definition 4.2.1. *The* m-shift transformation $\mathcal{T}^m$ *on* Ω *is defined by* $\mathcal{T}^m(\omega) = \{\omega_{n+m}\}_{n=-\infty}^{\infty}$ *for any* $\{\omega_n\}_{n=-\infty}^{\infty} \in \Omega,$ *i.e.* $(\mathcal{T}^m(\omega))_n = \omega_{n+m}$. The m-shift transformation of a set $B \in \mathcal{F}$ *is defined by* $\mathcal{T}^m(B) = \{\mathcal{T}^m(\omega) : \omega \in B\}$.

Note that $\mathcal{T}^{m+k}(\omega) = \mathcal{T}^m(\mathcal{T}^k(\omega))$ for any $m, k \in \mathbb{Z}$ and any $\omega \in \Omega$.

Definition 4.2.2. *For any random variable* $\Psi : \Omega \to \mathcal{X}$ *and any* $m \in \mathbb{Z}$*, the* m-shifted random variable Ψ_m *is defined by* $\Psi_m(\omega) = \Psi(\mathcal{T}^m(w))$ *where* $\omega \in \Omega$.

Remark 4.4. *The shift transformation* $\mathcal{T} : \Omega \to \Omega$ *is measure preserving. See [34] and [35] for more on shift transformations.*

The definition of shifted random variables suggests that we may also define shifted Markov chains once we represent them with stochastically recursive sequences . For this purpose, consider a Markov chain $\mathbf{X} = (X_n)_{n=0}^{\infty}$ on a Polish space $\mathcal{X}$ that is represented with an IFS $\{\mathcal{X}; f_\theta, \theta \in (0,1)\}$. We define the *m-shifted Markov chain* by $\mathcal{T}^m\mathbf{X} = (\mathcal{T}^m X_n)$.

So when $\mathbf{X}$ initiates from x_0, the m-shifted Markov chain starts at time m from $\mathcal{T}^m x_0$ and takes the value $\mathcal{T}^m X_n$ at time $m+n$ with $m \in \mathbb{Z}$ and $n \in \mathbb{N}$. Here x_0 is a random variable which is assumed to be adapted to the past. In most cases, x_0 will be a constant.

It is important to keep in mind that the definition of a shifted Markov chain makes use of the IFS representation of the original Markov chain $\mathbf{X}$. Shifted chains become more tractable once we express this relation more explicitly. Firstly, since the state of (X_n) at time n is given by

$$X_n = (f_{\theta_{n-1}} o ... o f_{\theta_0})(x_0)$$

we have,

$$\mathcal{T}X_n = (f_{\theta_n}o...of_{\theta_1})(\mathcal{T}x_0).$$

More generally we have,

$$\mathcal{T}^m X_n = (f_{\theta_{m+n-1}}o...of_{\theta_m})(\mathcal{T}^m x_0).$$

When x_0 is constant, $\mathcal{T}^m X_n = (f_{\theta_{m+n-1}}o...of_{\theta_m})(x_0)$.

We are now ready to give the definition of backward coupling times corresponding to a given Markov chain. These will be the key tools in embedding CFTP type algorithms into a more general framework.

Definition 4.2.3. *For a Markov chain $\boldsymbol{X} = (X_n)_{n=0}^{\infty}$ with a given IFS representation $\{\mathcal{X}; f_\theta, \theta \in (0,1)\}$, the* minimal backward coupling time *$\nu(\boldsymbol{X})$ is defined by*

$$\nu(\boldsymbol{X}) = \min\{m \geq 0 : \mathcal{T}^{-n_1}X_{n_1} = \mathcal{T}^{-n_2}X_{n_2}, \quad \forall n_1, n_2 \geq m\} \leq \infty.$$

Definition 4.2.4. *A random variable ν taking values in $\mathbb{N} \cup \{\infty\}$ is said to be a* backward coupling time *for a Markov chain $\boldsymbol{X} = (X_n)_{n=0}^{\infty}$ if*

$$\nu \leq m,\ m \in \mathbb{N} \Longrightarrow \mathcal{T}^{-n_1}X_{n_1} = \mathcal{T}^{-n_2}X_{n_2},\ \forall n_1, n_2 \geq m.$$

ν is said to be successful *when $\nu < \infty$ a.s..*

Note that the definition of a backward coupling time depends on just one chain starting from x_0 which was not the case for forward coupling times.

Now we shall give the fundamental theorem of this section.

Theorem 4.5. *[32] Let $\boldsymbol{X} = (X_n)$ be a Markov chain on a Polish space $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ with transition probability P and IFS representation $\{\mathcal{X}; f_\theta, \theta \in (0,1)\}$. Suppose that ν is a successful backward coupling time for $\boldsymbol{X}$. Define another stochastic sequence $\widetilde{\boldsymbol{X}} = (\widetilde{X}^n)$ by* *setting* $\widetilde{X}^0 = \mathcal{T}^{-\nu} X_\nu$ *and* $\widetilde{X}^n = \mathcal{T}^n \widetilde{X}^0$ *for* $n \geq 1$. *Then the sequence* $\widetilde{\boldsymbol{X}} = (\widetilde{X}^n)$ *forms a stationary Markov chain with transition probability* P, *and satisfies the recursion*

$$\widetilde{X}^{n+1} = f(\widetilde{X}^n, \theta_n)$$

a.s. for each $n \in \mathbb{N}$.

Proof. We start by showing that the stochastic sequence $(\widetilde{X}_n)$ satisfies the given recursion using induction. For $n = 0$, we have

$$\begin{aligned} \widetilde{X}^1 = \mathcal{T}\widetilde{X}^0 = \mathcal{T}\mathcal{T}^{-\nu} X_\nu = \lim_{m\to\infty} \mathcal{T}\mathcal{T}^{-m} X_m = \lim_{m\to\infty} \mathcal{T}^{-m+1} X_m &= \lim_{n\to\infty} \mathcal{T}^{-n} X_{n+1} \\ &= \lim_{n\to\infty} \mathcal{T}^{-n} f(X_n, \theta_n) \\ &= \lim_{n\to\infty} f(\mathcal{T}^{-n} X_n, \theta_0) \\ &= f(\mathcal{T}^{-\nu} X_\nu, \theta_0) \\ &= f(\widetilde{X}^0, \theta_0) \end{aligned}$$

where we just used the fact that $\mathcal{T}^{-\nu} X_\nu = \mathcal{T}^{-n} X_n$ for sufficiently large n a.s.. Next suppose that the result is true for some $n \in \mathbb{N}$. We have

$$\widetilde{X}^{n+1} = \mathcal{T}\widetilde{X}^n = \mathcal{T} f(\widetilde{X}^{n-1}, \theta_{n-1}) = f(\mathcal{T}\widetilde{X}^{n-1}, \mathcal{T}\theta_{n-1}) = f(\widetilde{X}^n, \theta_n)$$

as required. Now, since θ_n's are i.i.d., it follows from Theorem 2.3 that the stochastically recursive sequence $(\widetilde{X}_n)$ is a Markov chain with update function f and corresponding transition probability P.

Lastly, using the fact that $\mathcal{T}$ is measure preserving, we get

$$\mathbb{P}(\widetilde{X}^1 \in A) = \mathbb{P}(\mathcal{T}\widetilde{X}^0 \in A) = \mathbb{P}(\widetilde{X}^0 \in A)$$

for any $A \in \mathfrak{B}(\mathcal{X})$. So $(\widetilde{X}_n)$ is stationary as asserted. □

Remark 4.6. *Note that the random variable $\mathcal{T}^{-\nu}X_\nu$ in the proof of Theorem 4.5 is distributed according to a stationary distribution of the Markov chain (X_n). Henceforth, existence of a successful backward coupling time assures the existence of a stationary version of the Markov chain.*

4.3. Vertical Backward Coupling Times and Perfect Sampling

In this section, we embed CFTP type algorithms into a more general framework using the techniques developed in the previous section. For a Markov chain $\mathbf{X} = (X_n)$ with state space $\mathcal{X}$, denote by $X_n^{(z)}$ the chain that starts from $z \in \mathcal{X}$. That is, $X_n^{(z)}$ has the same transition probability with (X_n) but starts from a possibly different initial state $z \in \mathcal{X}$.

Theorem 4.7. *[32] Let $\boldsymbol{X} = (X_n)$ be a Markov chain on a Polish space $\mathcal{X}$ that is represented with an IFS $\{\mathcal{X}; f_\theta, \theta \in (0,1)\}$ and suppose that*

$$T = \min\{n \geq 0 : \mathcal{T}^{-n}X_n^{(z)} = \mathcal{T}^{-n}X_n^{(y)}, \quad \forall z, y \in \mathcal{X}\} \leq \infty$$

is well-defined and measurable. Then we have:

(i) For any $x_0 \in \mathcal{X}$, T is a backward coupling time for the Markov chain $\boldsymbol{X} = (X_n)$ starting from x_0.

(ii) If T is successful, then for any $x_0 \in \mathcal{X}$, $\mathcal{T}^{-T}X_T^{(x_0)}$ is distributed according to π where π is the unique stationary distribution of the Markov chain $(X_n^{(x_0)})$.

Proof. (i) Fix $x_0 \in \mathcal{X}$ and let $\mathbf{X} = (X_n)$ be the Markov chain starting at time zero from x_0. We wish to prove that $\mathcal{T}^{-m}X_m$ equals to a constant value for $m \geq T$ so that T is a backward coupling time for (X_n).

Let y be any element in $\mathcal{X}$ and set $\widetilde{X}^0 = \mathcal{T}^{-T}X_T^{(y)}$. Note that $\widetilde{X}^0$ is independent of y by definition of T. Now for any $m \geq T$, if $\mathcal{T}^{-m}X_{m-T} = z$, or equivalently if

$(f_{\theta_{-T-1}} o ... o f_{\theta_{-m}})(x_0) = z$ for some $z \in \mathcal{X}$, then we have

$$\mathcal{T}^{-m}X_m = (f_{\theta_{-1}} o ... o f_{\theta_{-T}} o f_{\theta_{-T-1}} o ... o f_{\theta_{-m}})(x_0) = (f_{\theta_{-1}} o ... o f_{\theta_{-T}})(z) = \mathcal{T}^{-T}X_T^{(z)}.$$

Since $\mathcal{T}^{-T}X_T^{(z)} = \mathcal{T}^{-T}X_T^{(y)} = \widetilde{X}^0$, we conclude that $\mathcal{T}^{-m}X_m = \widetilde{X}^0$ for $m \geq T$; that is, $\mathcal{T}^{-m}X_m$ becomes constant. Thus T is a backward coupling time for the Markov chain starting from x_0 as asserted.

(ii) By Theorem 4.5 we already know $\mathcal{T}^{-T}X_T^{(x_0)}$ is distributed according to π where π is a stationary distribution. In this case we claim that π is necessarily unique. Let π' be any stationary distribution. Suppose $x_0 \sim \pi'$. By the stationarity of π' we have, $\mathcal{T}^{-T}X_T^{(x_0)} \sim \pi'$. Following the above proof we also get, $\mathcal{T}^{-T}X_T^{(x_0)} = \mathcal{T}^{-T}X_T^{(y)} = \widetilde{X}^0 \sim \pi$. Thus $\pi = \pi'$. □

Definition 4.3.1. *For a Markov chain $\boldsymbol{X} = (X_n)$ on a Polish space $\mathcal{X}$, the random time $T = \min\{n \geq 0 : \mathcal{T}^{-n}X_n^{(z)} = \mathcal{T}^{-n}X_n^{(y)}, \quad \forall z, y \in \mathcal{X}\} \leq \infty$ is said to be the* minimal vertical backward coupling time *for the Markov chain $\boldsymbol{X}$.*

A random variable τ taking values in $\mathbb{N} \cup \{\infty\}$ is said to be a *vertical coupling time* for the Markov chain $\mathbf{X}$ when we have $\tau \geq T$ a.s. where T is the minimal vertical coupling time of $\mathbf{X}$. Also τ is said to be *successful* when τ is finite a.s.. Now the following corollary is immediate since the random time in the mentioned algorithm was just a particular example of a vertical backward coupling time.

Corollary 4.8. *Propp-Wilson algorithm is a valid algorithm to get perfect samples from a given distribution π.*

Remark 4.9. *The monotonicity arguments in PW algorithm can be extended to a vertical backward coupling time setting. For instance, suppose that (X_n) is a Markov chain with a monotone update function f on a Polish space $\mathcal{X}$. Suppose further that $\mathcal{X}$ is linearly ordered with a minimal element x and a maximal element y. Then it is true that the random time defined by $T = \min\{n \geq 0 : \mathcal{T}^{-n}X_n^x = \mathcal{T}^{-n}X_n^y\} \leq \infty$ is a vertical backward* *coupling time. To see this, observe that for any $m \in \mathbb{N}$ and any $z \in \mathcal{X}$, we have*

$$\mathcal{T}^{-m} X_m^x \leq \mathcal{T}^{-m} X_m^z \leq \mathcal{T}^{-m} X_m^y. \tag{4.1}$$

Thus, when T is finite, for any $m \geq T$, we have $\mathcal{T}^{-m} X_m^z = \mathcal{T}^{-m} X_m^x = \mathcal{T}^{-m} X_m^y$ where we use the fact that $\mathcal{T}^{-m} X_m^x = \mathcal{T}^{-m} X_m^y$ for $m \geq T$ by the definition of T. This reveals that T is in fact the minimal vertical backward coupling time. Generalizations for partially ordered state spaces are also possible [20].

We devote the rest of this section to proving the following fundamental theorem which gives a necessary and sufficient condition for the existence of successful minimal vertical backward coupling times for Markov chains.

Theorem 4.10. *[32] Let $\boldsymbol{X} = (X_n)$ be a Markov chain on a Polish space $\mathcal{X}$ with a unique stationary distribution π. Then the minimal vertical backward coupling time T of $\boldsymbol{X}$ is successful if and only if $\boldsymbol{X}$ is uniformly ergodic.*

We need to give two lemmas before working on the proof this theorem that are very important themselves.

Lemma 4.11. *If T is the minimal vertical backward coupling time for a Markov chain $\boldsymbol{X} = (X_n)$, then we have*

$$\mathbb{P}(T > m + n) \leq \mathbb{P}(T > m)\mathbb{P}(T > n)$$

for any $m, n \in \mathbb{Z}^+$.

Proof. For $m, n \in \mathbb{Z}^+$, we define the following events regarding the coalescence of Markov chains starting from different states,

$$\begin{aligned}
C_m &= \{\mathcal{T}^{-m} X_m^{(x)} = \mathcal{T}^{-m} X_m^{(y)} : \quad \forall x, y \in \mathcal{X}\}, \\
C_{n,m} &= \{\mathcal{T}^{-(m+n)} X_n^{(x)} = \mathcal{T}^{-(m+n)} X_n^{(y)} : \quad \forall x, y \in \mathcal{X}\}, \\
C_{m+n} &= \{\mathcal{T}^{-(m+n)} X_{m+n}^{(x)} = \mathcal{T}^{-(m+n)} X_{m+n}^{(y)} : \quad \forall x, y \in \mathcal{X}\}.
\end{aligned}$$

Since $C_m \cup C_{n,m} \subset C_{m+n}$, we have $\mathbb{P}(C_m \cup C_{n,m}) \leq \mathbb{P}(C_{m+n})$. This gives

$$\mathbb{P}(C^c_{m+n}) \leq \mathbb{P}(C^c_m \cap C^c_{n,m}) = \mathbb{P}(C^c_m)\mathbb{P}(C^c_{n,m})$$

as the events C_m and $C_{n,m}$ are independent. But we also have $\mathbb{P}(C^c_{m+n}) = \mathbb{P}(T > m+n)$, $\mathbb{P}(C^c_m) = \mathbb{P}(T > m)$ and $\mathbb{P}(C^c_{n,m}) = \mathbb{P}(C^c_n) = \mathbb{P}(T > n)$. Hence

$$\mathbb{P}(T > m+n) = \mathbb{P}(C^c_{m+n}) \leq \mathbb{P}(C^c_m)\mathbb{P}(C^c_{n,m}) = \mathbb{P}(T > m)\mathbb{P}(T > n)$$

as required. □

Lemma 4.12. *If the minimal vertical backward coupling time T of a Markov chain $\boldsymbol{X} = (X_n)$ is successful, then there exist $c \in (0, \infty)$ and $\lambda \in (0,1)$ such that*

$$\mathbb{P}(T > n) \leq c\lambda^n,$$

for every $n \in \mathbb{N}$.

Proof. Since T is successful, for a given $\beta \in (0,1)$, there exists $N \in \mathbb{N}$ for which we have $\mathbb{P}(T > n) < \beta$ for every $n \geq N$. Using Lemma 4.11 we get,

$$\mathbb{P}(T > mN) \leq (\mathbb{P}(T > N))^m < \beta^m = (\beta^{\frac{1}{N}})^{mN}, \quad \text{for} \quad m \in \mathbb{N}.$$

Set $\lambda = \beta^{\frac{1}{N}}$. For a given $k = pN + s \in \mathbb{N}$ with $s \in \{0, 1, ..., N-1\}$ we have

$$\mathbb{P}(T > k) \leq \mathbb{P}(T > pN)\mathbb{P}(T > s) < (\lambda)^{pN}\mathbb{P}(T > s).$$

Now if we choose $c > 0$ large enough so that

$$\mathbb{P}(T > s) < c(\lambda)^s \quad \text{for} \quad s = 0, 1, ..., N-1,$$

then we have for $k = pN + s \in \mathbb{N}$ with $s \in \{0, 1, ..., N-1\}$,

$$\mathbb{P}(T > k) < (\lambda)^{pN} \mathbb{P}(T > s) < (\lambda)^{pN} c\lambda^s = c\lambda^k$$

from which the result follows. □

Proof of Theorem 4.10. Suppose firstly that the minimal vertical backward coupling time T of $\mathbf{X}$ is successful. Then there exist $c \in (0, \infty)$ and $\lambda \in (0, 1)$ that satisfy $\mathbb{P}(T > n) \leq c\lambda^n$ for every $n \in \mathbb{N}$ by Lemma 4.12. Denote by $\tau_{\mathbf{X}}$ the minimal forward coupling time of the family $\{\mathbf{X}^{(x)}\}_{x \in \mathcal{X}} = \{(X_n^{(x)})\}$. Since $\mathcal{T}$ is measure preserving we have

$$\begin{aligned} \mathbb{P}(T > n) &= \mathbb{P}(\exists x, y : \mathcal{T}^{-n} X_n^{(x)} \neq \mathcal{T}^{-n} X_n^{(y)}) \\ &= \mathbb{P}(\exists x, y : X_n^{(x)} \neq X_n^{(y)}) \\ &= \mathbb{P}(\tau_{\mathbf{X}} > n). \end{aligned}$$

So $\mathbb{P}(\tau_{\mathbf{X}} > n) \leq c\lambda^n$.

We know by Theorem 4.1 that the coupling inequality $\|P^n(x_0, .) - \pi\| \leq \mathbb{P}(\widetilde{\tau} > n)$ holds where $\widetilde{\tau}$ is the forward coupling time for two Markov chains one of which is stationary and the other one is initiated from a state $x_0 \in \mathcal{X}$. Now since $\widetilde{\tau} \leq \tau_{\mathbf{X}}$ a.s., we also have $\mathbb{P}(\widetilde{\tau} > n) \leq \mathbb{P}(\tau_{\mathbf{X}} > n)$. Combining all of these we get

$$\|P^n(x_0, .) - \pi\| \leq \mathbb{P}(\widetilde{\tau} > n) \leq \mathbb{P}(\tau_{\mathbf{X}} > n) \leq c\lambda^n$$

which in particular says that $\mathbf{X}$ is uniformly ergodic by Theorem 2.11.

For the converse, assume that (X_n) is uniformly ergodic. Then by Theorem 2.12 there exists a probability measure Φ on $\mathcal{X}$, $m \geq 1$ and $\beta \in (0, 1]$ such that

$$P^m(x, .) \geq \beta \Phi(.), \quad \forall x \in \mathcal{X}.$$

Now, for $n \in \mathbb{Z}$, sample and fix i.i.d. sequences U_n uniform from $(0, 1)$ and V_n from Φ.

We construct the sample path of X_{mn} by setting $X_{mn} = V_{mn}$ if $U_{mn} \leq \beta$ and drawing $X_{mn} \sim \frac{1}{1-\beta}(P(X_{(n-1)m}, .) - \beta\Phi(.))$ if $U_{mn} > \beta$. Note that with this update mechanism we have,

$$\begin{aligned}\mathbb{P}(X_{mn} \leq y | X_{m(n-1)} = x) &= \beta\Phi((-\infty, y]) + (1-\beta)\frac{1}{1-\beta}(P(x, (-\infty, y]) - \beta\Phi((-\infty, y])) \\ &= P(x, (-\infty, y])\end{aligned}$$

and so the transitions are done according to P. Hence this mechanism gives an update function f for (X_n). Now if we set $T_U = min\{n \geq 1 : U_{mn} \leq \beta\}$, then for any $k \geq T_U$,

$$\begin{aligned}\mathcal{T}^{-k} X_k^{x_0} &= (f_{-1} o ... o f_{-T_U+1} o f_{-T_U} o ... o f_{-k})(x_0) \\ &= (f_{-1} o ... o f_{-T_U+1})(V_{-T_U})\end{aligned}$$

which is independent of x_0. Thus, T_U is a vertical backward coupling time for $\mathbf{X}$. Since $\mathbb{P}(T_U > k) \leq (1-\beta)^k$, T_U is finite a.s.. As the minimal vertical backward coupling time T satisfies $T \leq T_U$ a.s., we conclude that T is successful as asserted. □

Thus, one can use CFTP-type algorithms for perfect sampling only if the chain is uniformly ergodic.

5. LETAC'S PRINCIPLE

5.1. Letac's Principle

Consider a Markov chain (X_n) on a Polish space $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ with an IFS representation $\{\mathcal{X}; f_\theta, \theta \in \Theta\}$ where $(\Theta, \mathfrak{F}, Q)$ is a probability space. Recall that for such a Markov chain, we defined the *forward process* starting from $x \in \mathcal{X}$ by

$$F_0(x) = x, \quad F_n(x) = (f_{\theta_{n-1}} o...o f_{\theta_0})(x), \quad n = 1, 2, 3...$$

and the *backward process* starting from $x \in \mathcal{X}$ by

$$B_0(x) = x, \quad B_n(x) = (f_{\theta_0} o...o f_{\theta_{n-1}})(x), \quad n = 1, 2, 3...$$

Theorem 5.1. *[13] (Letac's Principle) Consider a Markov chain (X_n) on a Polish space $\mathcal{X}$ whose IFS representation is given by $\{\mathcal{X}; f_\theta, \theta \in \Theta\}$ where f_θ is continuous for each $\theta \in \Theta$. If $B = \lim_{n\to\infty} B_n(x)$ exists a.s. independent of x, then $\pi := L_B$ is the unique stationary distribution for (X_n).*

Proof. Since $B_n(x) \to B$ almost surely, $B_n(x) \to B$ weakly. Using $L_{B_n(x)} = L_{F_n(x)}$ for every n, we see that $F_n(x)$ converges to B weakly independent of x. Using Theorem 2.9, we conclude that L_B is the unique stationary distribution for the Markov chain. □

The following corollary relaxes the continuity condition in Theorem 5.1 using Corollary 2.10.

Corollary 5.2. *Theorem 5.1 remains valid when the assumption on the continuity of the update functions is replaced with the WFP assumption for the Markov Chain (X_n).*

Remark 5.3. *CFTP algorithms are closely related to Letac's principle. Indeed PW algorithm mainly relies on the following observation: Backward processes corresponding to* *finite state Markov chains with an appropriate IFS become constant at a finite time a.s. and so converge to some random limit. Convergence of the backward process allows us to have a random variable distributed according to the stationary distribution and get perfect samples from this distribution.*

Now we present an application of Letac's principle on processes evolving according to affine transition maps.

5.1.1. Affine Maps

Throughout this subsection, we consider a Markov chain (X_n) on $\mathcal{X} = \mathbb{R}$ whose update function $f : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}$ is given by $f(x, \theta) = f(x, (a, b)) = ax + b$ with $\theta = (a, b)$. Now, for $\theta_n = (a_n, b_n)$, the forward and backward processes turn out to be:

$$F_n(x) = (\Pi_{j=0}^{n-1} a_j)x + \sum_{k=1}^{n} b_{k-1}(\Pi_{j=k}^{n-1} a_j) \tag{5.1}$$

and

$$B_n(x) = (\Pi_{j=0}^{n-1} a_j)x + \sum_{k=0}^{n-1} b_k(\Pi_{j=0}^{k-1} a_j). \tag{5.2}$$

where we set $\Pi_{j=k}^{k-1} a_j = 1$ by convention.

To provide some motivation, we follow [14] and suppose for a moment that $a \in (0, 1)$ is constant. Then the first few terms of these two processes are;

$$F_0(x) = x, \quad F_1(x) = ax + b_0, \quad F_2(x) = a^2x + ab_0 + b_1, \quad F_3(x) = a^3x + a^2b_0 + ab_1 + b_2,$$

and

$$B_0(x) = x, \quad B_1(x) = ax + b_0, \quad B_2(x) = a^2x + ab_1 + b_0, \quad B_3(x) = a^3x + a^2b_2 + ab_1 + b_0.$$

What differs between these processes is that the new randomness in the backward process is damped by a power of a whereas the randomness in the forward process is preserved all the time. This explains why the backward process $B_n(x)$ converges independent of x and Letac's principle is applicable to various problems.

Next we detail an argument on affine maps given in [13] that provides sufficient conditions for the convergence of $B_n(x)$ in the case of affine update functions.

Theorem 5.4. *The following two conditions are sufficient for the convergence of backward process $B_n(x)$ in (5.2) independent of the initial position $x \in \mathbb{R}$:*

$$\gamma := \mathbb{E}(\log |a_0|) \in (-\infty, 0) \quad \textit{and} \quad \mathbb{E}(\log^+ |b_0|) < \infty \tag{5.3}$$

where a^+ =max$\{a, 0\}$.

Proof. We firstly show that

$$\lim_{k\to\infty} (\Pi_{j=0}^{k-1}|a_j|)^{1/k} < 1 \quad \text{a.s} \quad \text{and} \quad \lim_{n\to\infty} (\Pi_{j=0}^{n-1} a_j)x = 0 \quad \text{a.s.}$$

hold. Set $A_n = (\Pi_{j=0}^{n-1}|a_j|)^{1/n}$. Then the first one follows since we have,

$$\log A_n = \frac{1}{n}\sum_{j=0}^{n-1} \log |a_j| \to \gamma < 0$$

by Strong Law of Large Numbers. This also gives $\Sigma_{j=0}^{n-1} \log |a_j| \to -\infty$ as $n \to \infty$ which reveals the second one. Note that the second one in particular says that the first term in (5.2) drops as $n \to \infty$ for any $x \in \mathbb{R}$.

Next as an intermediate step, for $k \in \mathbb{Z}^+$ and $Y_k = \frac{-2}{\gamma} log^+|b_k|$, we observe

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{P}(Y_k > k) = \sum_{k=1}^{\infty}\sum_{j=k}^{\infty} \mathbb{P}(j < Y_1 \leq j+1) &= \sum_{j=1}^{\infty}\sum_{k=1}^{j} \mathbb{P}(j < Y_1 \leq j+1) \\
&= \sum_{j=0}^{\infty} j\mathbb{P}(j < Y_1 \leq j+1) \\
&= \sum_{j=0}^{\infty} \mathbb{E}(jI_{j<Y_1\leq j+1}) \\
&\leq \sum_{j=0}^{\infty} \mathbb{E}(Y_1 I_{j<Y_1\leq j+1}) \\
&= \mathbb{E}(\sum_{j=0}^{\infty} Y_1 I_{j<Y_1\leq j+1}) \\
&= \mathbb{E}(Y_1) \\
&= \frac{-2}{\gamma}\mathbb{E}(\log^+|b_1|) < \infty
\end{aligned}$$

where in the last step we use our assumption in (5.3). Now it follows that $\limsup_{k\to\infty} \frac{1}{k}\log^+|b_k| \leq \frac{-\gamma}{2}$ a.s. by Borel-Cantelli's lemma. So

$$\begin{aligned}
\limsup_{k\to\infty} \frac{1}{k}\log|a_0a_1...a_{k-1}b_k| &= \limsup_{k\to\infty}(\frac{1}{k}\log|a_0a_1...a_{k-1}| + \frac{1}{k}\log|b_k|) \\
&\leq \limsup_{k\to\infty}(\frac{1}{k}\log|a_0a_1...a_{k-1}|) + \limsup_{k\to\infty}(\frac{1}{k}\log|b_k|) \\
&\leq \gamma - \frac{\gamma}{2} \\
&= \frac{\gamma}{2} < 0
\end{aligned}$$

and this gives

$$\limsup_{k\to\infty} |a_1...a_{k-1}b_k|^{1/k} < 1, \quad a.s..$$

Using Cauchy's root test we see that the random series in (5.2) converges a.s.. As the first term in (5.2) also drops for any $x \in \mathcal{X}$, we conclude that the backward process $B_n(x)$ converges a.s. independent of x. □

The following corollary is an immediate consequence of Letac's principle.

Corollary 5.5. *Under the assumption (5.3), the random variable $\sum_{k=0}^{\infty} b_k(\Pi_{j=0}^{k-1} a_j)$ has the stationary distribution of the Markov chain $F_n(x)$ given by (5.1) for any $x \in \mathbb{R}$.*

A complete treatment of the convergence of the backward process in the case of affine transition maps relies on an analysis of the Lyapounov exponent. See [36] for results involving a necessary and sufficient condition for the convergence of backward process in the case of affine maps. Also see [13] and [37] for specific examples.

5.1.2. Systems Contracting on the Average

Our aim in this subsection is getting conclusions on stationary distributions of Markov chains that contract on the average. Analysis of the backward processes will be our primary tool in proving the following fundamental result as in Letac's principle.

Theorem 5.6. *[15] Consider a Markov chain (X_n) on a Polish space $\mathcal{X}$ with an IFS representation $\{\mathcal{X}; f_\theta, \theta \in \Theta\}$. Suppose that*

$$\mathbb{E}(d(f_{\theta_0}(x), f_{\theta_0}(y))) \leq cd(x,y), \quad \forall x, y \in \mathcal{X}$$

holds for some $c \in (0,1)$ and

$$\mathbb{E}(d(x_0, f_{\theta_0}(x_0))) < \infty, \quad \textit{for some} \quad x_0 \in \mathcal{X}.$$

Then there exists a unique stationary distribution μ for the Markov chain (X_n). Furthermore, for a given bounded subset S of $\mathcal{X}$ there exists a positive constant α_S satisfying

$$\sup_{x \in S} d_K(\mu_n^x, \mu) \leq \alpha_S c^n, \quad n \geq 0 \tag{5.4}$$

where d_K is the Kantorovich distance between probability measures and μ_n^x is the distribution of the chain starting from $x \in \mathcal{X}$ at time n.

Remark 5.7. *Note that the update functions are not assumed to be continuous.*

Proof of Theorem (5.6). As in the proof of Letac's principle, the backward process defined by $B_0(x) = x$ and $B_n(x) = (f_{\theta_0} o ... o f_{\theta_{n-1}})(x)$, $n \geq 1$ will play a key role in the proof. Firstly we start by showing that $B_n(x)$ converges a.s.. Proving that it is Cauchy will suffice as $\mathcal{X}$ is complete.

For $N \leq n \leq m$ we have, $d(B_n(x), B_m(x)) \leq \sum_{k=N}^{\infty} d(B_k(x), B_{k+1}(x))$. Now, by an application of Fatou's lemma, observe that if $\mathbb{E}\left(\sum_{k=N}^{\infty} d(B_k(x), B_{k+1}(x))\right) \to 0$, then $\sum_{k=N}^{\infty} d(B_k(x), B_{k+1}(x)) \to 0$ and so the Cauchyness of $B_n(x)$ follows. We have,

$$\begin{aligned}
\mathbb{E}\left(\sum_{k=N}^{\infty} d(B_k(x), B_{k+1}(x))\right) &= \sum_{k=N}^{\infty} \mathbb{E}(d(B_k(x), B_{k+1}(x))) \\
&= \sum_{k=N}^{\infty} \mathbb{E}(\mathbb{E}(d(B_k(x), B_{k+1}(x))|f_{\theta_1}, ..., f_{\theta_k})) \\
&= \sum_{k=N}^{\infty} \mathbb{E}(\mathbb{E}(d(f_{\theta_0}(f_{\theta_1} o ... o f_{\theta_{k-1}}(x)), f_{\theta_0}(f_{\theta_1} o ... o f_{\theta_k}(x)))|f_{\theta_1}, ..., f_{\theta_k})) \\
&\leq \sum_{k=N}^{\infty} c\mathbb{E}(d(f_{\theta_1} o ... o f_{\theta_{k-1}}(x)), (f_{\theta_1} o ... o f_{\theta_k}(x)))
\end{aligned}$$

Inductively we get

$$\mathbb{E}\left(\sum_{k=N}^{\infty} d(B_k(x), B_{k+1}(x))\right) \leq \sum_{k=N}^{\infty} c^k \mathbb{E}(d(x, f_{\theta_k}(x))) = \frac{c^N}{1-c}\mathbb{E}(d(x, f_{\theta_0}(x))). \tag{5.5}$$

Next we observe

$$\mathbb{E}(d(x, f_{\theta_0}(x))) \leq \mathbb{E}(d(x, x_0)) + \mathbb{E}(d(x_0, f_{\theta_0}(x_0))) + \mathbb{E}(d(f_{\theta_0}(x_0), f_{\theta_0}(x))) < \infty \tag{5.6}$$

and so $\sum_{k=N}^{\infty} d(B_k(x), B_{k+1}(x)) \to 0$. Thus $B_k(x)$ is Cauchy and therefore converges.

We shall now prove that this convergence is independent of x. Let $y \in \mathcal{X}$ and $\epsilon > 0$.

We have

$$\begin{aligned}
\mathbb{P}(d(B_n(x), B_n(y)) > \epsilon) &\leq \frac{\mathbb{E}(d(B_n(x), B_n(y)))}{\epsilon} \\
&= \frac{1}{\epsilon}\mathbb{E}(\mathbb{E}(d(B_n(x), B_n(y)|f_{\theta_1}, ..., f_{\theta_{n-1}})) \\
&\leq \frac{c}{\epsilon}\mathbb{E}(d(f_{\theta_1}o...of_{\theta_{n-1}}(x), f_{\theta_1}o...of_{\theta_{n-1}}(y)))
\end{aligned}$$

Iterating n times, we get

$$\mathbb{P}(d(B_n(x), B_n(y)) > \epsilon) \leq \frac{c^n}{\epsilon}\mathbb{E}(d(x, y)).$$

Setting $A_n = \{d(B_n(x), B_n(y)) > \epsilon\}$, this gives $\sum_{n=0}^{\infty} \mathbb{P}(A_n) < \infty$. Using Borel-Cantelli lemma, we conclude that for a given $\epsilon > 0$ and for $x, y \in \mathcal{X}$, there exists $N = N_{x,y,\epsilon} \in \mathbb{N}$ such that for all $n \geq N$, we have $d(B_n(x), B_n(y)) \leq \epsilon$. But this says that $d(B_n(x), B_n(y)) \to 0$ as $n \to \infty$.

Next let $X = \lim_{n\to\infty} B_n(x_0)$ where x_0 is the special point given in the assumptions of the theorem. For any $x \in \mathcal{X}$, we have $d(B_n(x), X) \leq d(B_n(x), B_n(x_0)) + d(B_n(x_0), X) \to 0$ so that $d(B_n(x), X) \to 0$ a.s.. This gives $B_n(x) \to X$ a.s. independent of x. Now letting $\mu = L_X$, we observe that for $n \geq 0$,

$$\begin{aligned}
d_K(\mu_n^x, \mu) &= \sup\{|\int_{\mathcal{X}} f d(\mu_n^x - \mu)| : \|f\|_L \leq 1\} \\
&= \sup\{|\mathbb{E}(f(B_n(x))) - \mathbb{E}(f(X))| : \|f\|_L \leq 1\} \\
&\leq \sup\{\mathbb{E}(|f(B_n(x)) - f(X)|) : \|f\|_L \leq 1\} \\
&\leq \mathbb{E}(d(B_n(x), X)) \\
&= \mathbb{E}(\lim_{m\to\infty} d(B_n(x), B_m(x))) \\
&\leq \mathbb{E}(\lim_{m\to\infty} \sum_{k=n}^{m-1} d(B_k(x), B_{k+1}(x))).
\end{aligned}$$

So we get

$$d_K(\mu_n^x, \mu) \leq \mathbb{E}(\sum_{k=n}^{\infty} d(B_k(x), B_{k+1}(x))) \leq \frac{c^n}{1-c}\mathbb{E}(d(x, f_{\theta_0}(x)))$$

where in the last step we used our calculation in (5.5). For a given bounded subset S of $\mathcal{X}$, we get

$$\sup_{x \in S} d_K(\mu_n^x, \mu) \leq \frac{c^n}{1-c}\gamma_S, \quad n \geq 0,$$

with $\gamma_S := \sup_{x \in S} \mathbb{E}(d(x, f_{\theta_0}(x))) < \infty$ since $\gamma_S \leq \sup_{x \in S}(\mathbb{E}(d(x_0, f_{\theta_0}(x_0)) + (c+1)d(x, x_0))) < \infty$ by (5.6). Setting $\alpha_S = \frac{\gamma_S}{1-c}$, we get

$$\sup_{x \in S} d_K(\mu_n^x, \mu) \leq \alpha_S c^n, \quad n \geq 0$$

which gives us the convergence rate result given in (5.4). Now we only need to prove that μ is a stationary probability and actually the unique one having this property. We recall from Corollary 5.2 that if the Markov chain (X_n) has WFP, then the limiting distribution is the unique stationary distribution. So proving that our chain has WFP will suffice.

For this purpose, let (x_n) be a sequence in $\mathcal{X}$ with $x_n \to x$. We claim that $\mathbb{E}(g(f_{\theta_0}(x_n))) \to \mathbb{E}(g(f_{\theta_0}(x)))$ for any $g \in BC(\mathcal{X})$. Using Markov's inequality, we get

$$\begin{aligned} \mathbb{P}(d(f_{\theta_0}(x_n), f_{\theta_0}(x)) > \epsilon) &\leq \frac{\mathbb{E}(d(f_{\theta_0}(x_n), f_{\theta_0}(x)))}{\epsilon} \\ &\leq c\frac{d(x_n, x)}{\epsilon} \to 0 \end{aligned}$$

as $n \to \infty$. So $f_{\theta_0}(x_n) \to f_{\theta_0}(x)$ in probability and from this we get $f_{\theta_0}(x_n) \to f_{\theta_0}(x)$ in distribution. From this we get, $\lim_{n \to \infty} \mathbb{E}(g(f_{\theta_0}(x_n))) = \mathbb{E}(g(f_{\theta_0}(x)))$ for any $g \in BC(\mathcal{X})$. Thus WFP is satisfied and μ is the unique stationary distribution. □

Some authors study algebraic tail conditions for the tail behavior of appropriate random variables to catch sufficient conditions for the convergence of the backward process corresponding to Markov chains. By definition, a random variable X has *algebraic tail* if there exist $\alpha, \beta \in (0, \infty)$ such that $\mathbb{P}(X > x) < \alpha/x^{\beta}$ for all $x > 0$. Algebraic tail conditions not only bring insight to IFS but also help getting important theorems on stationary distributions of Markov chains. See [14] and [38] for such instances. Here we quote one such theorem whose proof is very similar to the proof of Theorem 5.6.

Theorem 5.8. *[14] Let $\mathcal{L}(\mathcal{X})$ be the set of Lipschitz functions on a Polish space $(\mathcal{X}, d)$ and μ be a probability measure on $\mathcal{L}(\mathcal{X})$. Suppose that*

$$f \mapsto K_f \quad \textit{has an algebraic tail with respect to} \quad \mu$$

and for some $x_0 \in \mathcal{X}$

$$f \mapsto d(f(x_0), x_0) \quad \textit{has an algebraic tail with respect to} \quad \mu.$$

Now, consider a Markov chain on $\mathcal{X}$ that moves according to the following rule: starting from x, the chain chooses $f \in \mathcal{L}(\mathcal{X})$ according to μ and goes to $f(x)$. Furthermore we assume that

$$\int_{\mathcal{L}(X)} \log K_f \mu(df) < 0$$

where the integral can be $-\infty$. Letting μ_n^x be the law of chain after n moves starting from x, we have:

(i) There is a unique stationary distribution π for the Markov chain.

(ii) There exist $A_x \in (0, \infty)$ and an $r \in (0, 1)$ such that $d_P(\mu_n^x, \pi) \leq A_x r^n$ for $n \geq 1$ and $x \in X$. The constant r does not depend on n or x; the constant A_x does not depend on n, and $A_x < a + bd(x, x_0)$ where $0 < a, b < \infty$.

The essence of the problem is as before: Backward process converges at a geometric rate to the stationary distribution independent of the initial state. See [14] for the proof and interesting examples that study what happens without the regularity conditions given in the theorem. [14] also contains ideas from applications to queueing theory and image processing.

Note that there is also a considerable interest in the case where we have a Markov chain with an IFS of finitely many (affine) strict contractions. The motivation for this case is obtaining a unified method for generating and classifying a broad class of fractals. See [39] for an analysis of this case.

APPENDIX A: PROBABILITY METRICS

In this section, $(\mathcal{X}, d)$ is a Polish space with its Borel $\sigma-$algebra $\mathfrak{B}(\mathcal{X})$. Denote by $BL(\mathcal{X})$ the set of bounded continuous functions $f : \mathcal{X} \to \mathbb{R}$ that also satisfy the Lipschitz condition

$$\|f\|_L := \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)} < \infty.$$

Define $\|f\|_{BL} = \|f\|_\infty + \|f\|_L$. Then, $(BL(\mathcal{X}), \| \, \|_{BL})$ is a normed vector space and for any $f, g \in BL(\mathcal{X})$, we have $\|fg\|_{BL} \leq \|f\|_{BL}\|g\|_{BL}$. See [19] for the proofs.

Definition A.1. *For two probability measures μ_1, μ_2 on $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$, we define the* Wasserstein distance *by*

$$d_W(\mu_1, \mu_2) = \sup\{\left|\int_{\mathcal{X}} f(\mu_1 - \mu_2)(dx)\right| : \|f\|_{BL} \leq 1\}$$

and the Kantorovich distance *by*

$$d_K(\mu_1, \mu_2) = \sup\{\left|\int_{\mathcal{X}} f(\mu_1 - \mu_2)(dx)\right| : \|f\|_L \leq 1\}.$$

Definition A.2. Prokhorov distance *$d_P(\mu_1, \mu_2)$ between two probability measures μ_1, μ_2 on $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ is defined to be the infimum of the $\delta > 0$ that satisfies*

$$\mu_1(K) < \mu_2(K_\delta) + \delta \quad \textit{and} \quad \mu_2(K) < \mu_1(K_\delta) + \delta$$

for all compact subset K of $\mathcal{X}$, where $K_\delta = \{x \in \mathcal{X} : d(K, x) < \delta\}$.

For more on Prokhorov distance, see [27] and [40].

Theorem A.3. *[19] For any probability measures μ_n and μ on $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$, the following are equivalent.*

(i) μ_n *converges weakly to* μ.

(ii) $d_W(\mu_n, \mu) \to 0$.

(iii) $d_P(\mu_n, \mu) \to 0$.

Definition A.4. Total variation distance *of probability measures μ_1 and μ_2 on $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ is defined by*

$$\|\mu_1 - \mu_2\|_{TV} = \sup\{|\mu_1(A) - \mu_2(A)| : A \in \mathfrak{B}(\mathcal{X})\}.$$

APPENDIX B: GIBBS SAMPLING

Gibbs sampling is an MCMC method that is used to get approximate samples from a multivariate distribution $\pi(x_1, ..., x_n)$ in cases where sampling from the full conditionals can be easily implemented. It may be seen as a special case of Metropolis-Hastings algorithm. Since this technique is used in the thesis, we shortly describe the simulation procedure. For a detailed treatment, see [4] and [6].

To use Gibbs sampling, one needs to sample from the conditional distributions of each component given the remaining components (full conditional distributions), i.e. to simulate from

$$\pi_{X_i|X_{-i}}(x_i|x_{-i}) \quad \text{where} \quad X_{-i} = (X_1, ..., X_i, X_{i+1}, ..., X_n).$$

Letting $\mathbf{x}^{(t)} = (x_1^{(t)}, x_2^{(t)}, ..., x_n^{(t)})$ be the state of the chain at time t, we may describe the two common forms of Gibbs sampler as follows.

1. Random Gibbs Sampler At time $t+1$, choose a coordinate i uniformly from $\{1, 2, ..., n\}$. Then draw $x_i^{(t+1)} \sim \pi_{X_i|X_{-i}}(x_i|x_{-i})$ and leave all other coordinates unchanged.

2. Deterministic Gibbs Sampler At time $t+1$, draw $x_i^{(t+1)}$ from the conditional distribution

$$\pi(x_i|x_1^{(t+1)}, ..., x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, ..., x_n^{(t)})$$

for $i = 1, ..., n$.

REFERENCES

1. Eckhardt, R., *Stan Ulam, John von Neumann, and the Monte Carlo method*, Los Alamos Sci. no. 15, Special Issue, pp. 131–137, 1987.

2. Casella, G. and C. P. Robert, *A History of Markov Chain Monte Carlo*, Unpublished notes, 2008.

3. Metropolis, N., *The beginning of the Monte Carlo method*, Los Alamos Sci., No. 15, Special Issue, pp. 125–130, 1987.

4. Liu, J. S., *Monte Carlo strategies in scientific computing*, Springer Series in Statistics, Springer, New York, 2001.

5. Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, *Equations of state calculations by fast computing machines*, J. Chem. Phys., no. 21, 1087–1092, 1953.

6. Casella, G. and C. P. Robert, *Monte Carlo Statistical Methods*, Springer Texts in Statistics, Springer-Verlag, New York, 2004.

7. Diaconis, P., *The Markov chain Monte Carlo revolution*, Bull. Amer. Math. Soc., 46, pp. 179–205, 2009.

8. Haggström, O., *Problem solving is often a matter of cooking up an appropriate Markov chain*, Scand. J. Statist., 34, pp. 768–780, 2007.

9. Diaconis, P. and L. Saloff-Coste, *What do we know about the Metropolis algorithm?*, J. Comput. System Sci. 57, no. 1, 20–36, 1998.

10. Propp, J. G. and D. B. Wilson, *Exact sampling with coupled Markov chains and applications to statistical mechanics*, Proceedings of the Seventh International Conference

on Random Structures and Algorithms, 1995.

11. Murdoch, D. J. and P. J. Green, *Exact sampling from a continuous state space*, Scand. J. Statist., 25, pp. 483-502, 1998.

12. Fill, J. A., *An interruptible algorithm for perfect sampling via Markov chains*, Ann. Appl. Probab., 8, no. 1, pp. 131–162, 1998.

13. Letac, G., *A contraction principle for certain Markov chains and its applications*, Contemp. Math., 50, pp. 263-273, 1984.

14. Diaconis, P. and D. Freedman, *Iterated random functions*, SIAM Rev., 41, pp. 45–76, 1999.

15. Stenflo, Ö., *Ergodic theorems for Markov chains represented by iterated function systems*, Bull. Polish Acad. Sci. Math., 49, pp. 27–43, 2001.

16. Stenflo, Ö., *Markov chains in random environments and random iterated function systems*, Trans. Amer. Math. Soc. 353, no.9, pp. 3547–3562.

17. Athreya, K. B. and Ö. Stenflo, *Perfect sampling for Doeblin chains*, Sankhya 65, no. 4, 763–777, 2003.

18. Krylov, N. V., *Introduction to the theory of random processes*, Graduate Studies in Mathematics, 43, American Mathematical Society, Providence, RI, 2002.

19. Dudley, R. M., *Real analysis and probability*, Cambridge Studies in Adv. Mathematics, 74, Cambridge University Press, Cambridge, 2002.

20. Borovkov, A. A. and S. G. Foss, *Stochastically recursive sequences and their generalizations*, Siberian Adv. Math., 2, pp. 16–81, 1992.

21. Kallenberg, O., *Foundations of modern probability*, Springer-Verlag, New York, 2002.

22. Dubischar, D., *The representation of transition probabilities by random maps*, Doctoral thesis, Bremen University, Germany, 1999.

23. Folland, G. B., *Real analysis*, John Wiley and Sons, 1999.

24. Rogers, L. C. G. and D. Williams, *Diffusions, Markov processes, and martingales. Vol. 1.*, Cambridge Mathematical Library, Cambridge University Press, 2000.

25. Breiman, L., *Probability*, Classics in Applied Mathematics, 7, SIAM, 1992.

26. Meyn, S. P. and R. L. Tweedie, *Markov chains and stochastic stability*, Communications and Control Engineering Series, Springer-Verlag London, Ltd., London, 1993.

27. Shiryaev, A. N., *Probability*, Graduate Texts in Mathematics, Springer-Verlag, New York, 1996.

28. Fismen, M., *Exact simulation using Markov chains*, Master thesis, The Norwegian University of Science and Technology, 1998.

29. Bremaud P., *Markov chains. Gibbs fields, Monte Carlo simulation, and queues*, Texts in Applied Mathematics, 31, Springer-Verlag, New York, 1999.

30. Haggström, O., *Finite Markov chains and algorithmic applications*, Cambridge University Press, Cambridge, 2002.

31. Wilson B., *How to couple from the past using a read-once source of randomness*, Random Structures Algorithms, 16, pp. 85-113, 2000.

32. Foss, S.G. and R. L. Tweedie, *Perfect simulation and backward coupling*, Comm. Statist. Stochastic Models, 14, pp. 187-203, 1998.

33. Lindvall, T., *Lectures on the Coupling Method*, Dover Publications, 2002.

34. Petersen, K., *Ergodic Theory*, Cambridge Studies in Adv. Mathematics, 2, Cambridge University Press, Cambridge, 1989.

35. Krengel, U., *Ergodic Theorems*, de Gruyter Studies in Mathematics, 6, Walter de Gruyter Co., Berlin, 1985.

36. Bougerol, P. and N. Picard, *Strict stationarity of generalized autoregressive processes*, Ann. Probab., 20, no. 4, pp. 1714–1730, 1992.

37. Vervaat, W., *On a stochastic difference equation and a representation of nonnegative infinitely divisible random variables*, Adv. in Appl. Probab, 11, pp. 750–783, 1979.

38. Wu, W. B. and X. Shao, *Limit theorems for iterated random functions* J. Appl. Probab., 41, no.2, pp. 425–436, 2004.

39. Barnsley, M. F. and J. H. Elton, *A new class of Markov processes for image encoding* Adv. in Appl. Probab., 20, no.1, pp. 14-32, 1988.

40. Billingsley, P., *Convergence of probability measures*, John Wiley and Sons, Inc., 1968.

REFERENCES NOT CITED

1. Levin, D. A., Y. Peres and E. Wilmer, *Markov chains and mixing times*, American Mathematical Society, 2009.

2. Djuric, P. M., Y. Huang and T. Ghirmai, *Perfect sampling: a review and applications to signal processing*, IEEE Trans. Signal Process., 50, pp. 345–356, 2002.

3. Goldie, C. M., *Implicit renewal theory and tails of solutions of random equations*, Ann. Appl. Probab., 1, pp. 126–166, 1991.

4. Schneider, U., *Advances and Applications in Perfect Sampling*, Doctoral Thesis, University of Colorado, 2003.

5. Goswami, A., *Random continued fractions: a Markov chain approach*, Econom. Theory, 23, pp. 85–105, 2004.

6. Barnsley, M. F., J. E. Hutchinson and Ö. Stenflo, *V-variable fractals: fractals with partial self similarity* Adv. Math., 218 , pp. 2051-2088, 2008.

7. Chamayou, J. and G. Letac, *Explicit stationary distributions for compositions of random functions and products of random matrices*, J. Theoret. Probab., 4, no. 1, pp. 3–36, 1991.

8. Chassaing, P., G. Letac and M. Mora, *Brocot sequences and random walks in* SL$(2, R)$, Lecture Notes in Math., Springer, Berlin, 1984.

9. Corcoran, J. N. and R. L. Tweedie, *Perfect sampling of ergodic Harris chains* Ann. Appl. Probab., 11, pp. 438–451, 2001.