ORDERING, SLICING AND SPLITTING MONTE CARLO MARKOV CHAINS

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

ANTONIETTA MIRA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Luke J. Tierney, Adviser

October 1998

UNIVERSITY OF MINNESOTA

This is to certify that I have examined this copy of a doctoral thesis by

ANTONIETTA MIRA

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Luke J. Tierney

Name of Faculty Adviser(s)

Signature of Faculty Adviser(s)

Date

GRADUATE SCHOOL

Con hostinato rigore

Leonardo da Vinci

Acknowledgments

Peskun, a student of Hastings, published a very interesting but not well known paper [64] which introduces the idea of ordering transition matrices. This paper together with that of Besag and Green [6], discussing the slice sampler, provided the inspiration for this dissertation.

The paper by Peskun was the second one that my advisor, Luke Tierney, advised me to read. The first one, of course, was the work by Hastings [35]. Since then my discussions with Luke have proven to be very valuable to me as have his suggestions. Thank you Luke for putting up with my unpredictable schedule and for the freedom given to me in my research while still providing precious academic and intellectual guidance that have been essential for achieving the results described in the following pages.

I was introduced to the slice sampler by Steve Walker while I was visiting the Imperial College in London. I am very thankful to Steve because he accelerated the often very slowly converging process of getting started on ones thesis.

A year later I met Peter Green at the University of Pavia where he was giving a short course on MCMC. This course and further instruction from Peter enlightened me on the slice sampler and other tricks of the MCMC game.

If we really want to follow this long chain back to its very beginning (no, this one is not a Markovian chain - it has a long memory!) I should thank Pietro Muliere for not having given in to my attempts to avoid the cold winters in Minneapolis and insisting that I would apply for the Ph.D. program here. His guide throughout my whole career, he is the one responsible for my choice to become a statistician in the first place, has been very precious to me: I am getting there, Pietro!

Luckily enough a lot of friends warmed up my stay in Minneapolis and without their help and support I would have certainly frozen. Piercesare and Elena, George, Francesca, Efstathia, the Allbrights, Florin, Igor, Laura and Giampaolo, Alessia and Jonathan, Andres and Jackie, John, Michael, Erik and James kept me going on this side of the Ocean while Clelia, Antonio, Fabio, Alberto, and James were always patiently waiting for my return on the other side of the Ocean. No, James is not omnipresent, we still have to properly synchronize!

I would like to thank the School of Statistics of the University of Minnesota for providing the perfect environment where to form, develop and discuss ideas. I am also deeply in debt to the numerous friends and colleagues of the department of Quantitative Methods and the department of Informatics of the University of Pavia and the new faculty of Economics in Varese. In particular to Raffaella and Claudio who have taken up my teaching duties in Varese: this work would still be in its burn-in phases if it were not for you.

I would like to acknowledge the substantial help and the numerous suggestions of two patient readers of this thesis: John Baxter and Charlie Geyer.

John taught me to look at statistics from a different angle and with a wider view. He lent me his mathematical glasses that helped me see things more clearly and often gave color and depth to otherwise black and white flat images. Thank you John, for providing the mathematical insight that I often lack, for asking the simple question that I didn't think about, and for being a close friend and mentor.

My research has strongly benefited from the fact that the light in Charlie's office is always on and I feel welcome to simply walk in and talk, not only about statistics. He has open my eyes on a new world, or better yet, on a new space, $L_0^2(\pi)$, and taught me all the functional analysis used in my thesis and more. He often pulled out of his hat the right book with the right theorem and if the theorem was not there he helped me to get it out of my head. Thank you Charlie for being there, 24 hours

a day, 7 days a week - well, not quite, but this is the limiting distribution of the time we spent discussing and learning. And we were very close to convergence!

I have stolen from Gareth some of his enthusiasm for the research and hopefully I have repaid him with some skiing advice. The last chapter of this thesis is joint work with him and was conceived on the ski runs in Aussois. I hope to soon continue this line of research at some other ski resort.

I have received enormous encouragement and support from my family, a voi dedico questa tesi, con un pensiero particolare al NONNO che da piccola mi ha insegnato a contare dividendo le monetine della mancia domenicale.

A special thank you to my TWA friends, without their cheap stand-by tickets I would have not been able to afford all this traveling across the Ocean!

I gratefully acknowledge the Graduate School of the University of Minnesota for supporting my research with a Doctoral Dissertation Fellowship.

Contents

List of Figures

Ordering, Slicing and Splitting Monte Carlo Markov Chains

Antonietta Mira

October 14, 1998

Abstract

Markov chain Monte Carlo is a method of approximating the integral of a function f with respect to a distribution π. A Markov chain that has π as its stationary distribution is simulated producing samples $X_1, X_2, \ldots$. The integral is approximated by taking the average of $f(X_n)$ over the sample path. The standard way to construct such Markov chains is the Metropolis-Hastings algorithm.

The class $\mathcal{P}$ of all Markov chains having π as their unique stationary distribution is very large, so it is important to have criteria telling when one chain performs better than another. The *Peskun ordering* is a partial ordering on $\mathcal{P}$. If two Markov chains are Peskun ordered, then the better chain has smaller variance in the central limit theorem for every function f that has a variance. Peskun ordering is sufficient for this but not necessary. We study the implications of the Peskun ordering both in finite and general state spaces.

Unfortunately there are many Metropolis-Hastings samplers that are not comparable in the Peskun sense. We thus define a more useful ordering, the *covariance ordering*, that allows us to compare a wider class of sampling schemes. Two Markov chains are covariance ordered if and only if the better one has smaller variance in the central limit theorem for every function f that has a variance. Unlike the Peskun ordering, the covariance ordering is therefore necessary and sufficient.

In the second part of this work we show that it is always possible to beat the Metropolis-Hastings algorithm in terms of the Peskun ordering. Two new algorithms will be introduced and studied for this purpose. The *slice sampler* outperforms the independence Metropolis-Hastings algorithm. The *splitting rejection sampler* outperforms the more general Metropolis-Hastings algorithm.

Chapter 1

Introduction

Markov chain Monte Carlo (MCMC) algorithms allow the approximation of the expectation of a function f with respect to a complicated density function π often only known up to a normalizing constant. The underlying idea is to construct a discrete time ergodic Markov chain with stationary distribution π which is easy to simulate without knowing the normalizing constant of π. Then, in order to approximate $\mu = E_\pi[f(x)]$, we typically compute the empirical mean of f along a sample path of the chain, $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. When n is large, ergodicity ensures that $\hat{\mu}_n$ is an unbiased estimate of μ. In practice it is appropriate to disregard an initial part of the sample path in order to avoid strong dependence on the initial distribution of the chain and consider instead the estimator $\frac{1}{n}\sum_{i=n_0+1}^{n_0+n} f(X_i)$. In this case we have a burn-in period of length n_0. The asymptotic variance of the MCMC estimates is a measure of the quality of the approximation.

In this work we make a contribution to answering some of the open questions that arise when implementing MCMC. In particular we focus on the following issues.

Given a target distribution π, there are various transition kernels that give rise to different Markov chains having π as their stationary distribution. The practitioner is often faced with the problem of choosing one of them or an efficient combination of them. Among the possible criterion that can drive this choice we focus on the asymptotic variance of the resulting MCMC estimates. In Chapter 2 we discuss partial orderings of transition kernels with respect to this criterion; in particular we study the implications of the *Peskun ordering* [64] and propose a generalization of it. Some of the results are extended to non-reversible transition kernels.

In Chapter 3 we compare various transition kernels existing in the literature with respect to the ordering previously studied. We focus on different versions of the *slice sampler*, a modification of the independence Metropolis-Hastings algorithm via auxiliary variables. The performance of the slice sampler is compared with that of the original independence Metropolis-Hastings algorithm. We show that the former outperforms the latter since not only does it produce MCMC estimates with a smaller asymptotic variance, but it also has a faster rate of convergence to stationarity in total variation distance. Other properties of the slice sampler are studied. In particular a sufficient condition for uniform ergodicity is given. Furthermore an upper bound to the rate of convergence to stationarity is provided.

In Chapter 4 various slice samplers are analyzed in detail. The target distributions considered are the exponential, the Cauchy and the witch's hat distribution.

In Chapter 5 we propose a new method to construct a Markov chain having a specified stationary distribution, the *splitting rejection sampler*. This method is a modification of the Metropolis-Hastings algorithm in order to obtain better performance in terms of Peskun ordering. The new idea is exploited to implement efficient slice samplers.

Chapter 2

Ordering Monte Carlo Markov Chains

Can you label three six sided dice with any subset of the numbers 1-18, each die labeled differently, such that: die A beats die B at least 21/36 of the time, die B beats die C at least 21/36 of the time, die C beats die A at least 21/36 of the time?

2.1 Introduction

In statistics, both Bayesian and classical, we are often faced with the problem of integrating over high-dimensional distributions to make inferences or predictions. Let π be a density function, possibly specified up to a normalized constant, which, due to its complexity, cannot be easily studied by standard analytical or approximation techniques. The key intuition behind Markov chain Monte Carlo (MCMC) methods is to design a transition kernel $P(x, A) = P(X_n \in A | X_{n-1} = x)$ that has π as its unique stationary distribution,

$$\pi(A) = \int P(x, A)\pi(dx) \tag{2.1}$$

for all measurable sets A.

There are different strategies to construct such a transition law by combining, via composition or mixing, elementary Gibbs or Metropolis-Hastings update mechanisms ([28], Chapter 1). In composition one update mechanism is followed by another

in a predefined fixed order. When mixing elementary updates we choose at random among them according to some predefined probability distribution.

Consider the class $\mathcal{P}$ of all transition kernels which have π as their stationary distribution. The researcher is often faced with the problem of choosing one transition kernel within $\mathcal{P}$ according to some criterion: ease of implementation, speed of convergence to stationarity, mixing properties of the Markov chain, asymptotic variances of the resulting estimates. We will focus on this last criterion to define an ordering on $\mathcal{P}$ in an attempt to assist the choice of the researcher among different transition kernels. To justify the choice of this criterion consider that, in classical statistics, estimates are compared in terms of their asymptotic relative efficiency; likewise here we will prefer a transition kernel if it produces estimates that are asymptotically more efficient. If the chain is irreducible and the state space is finite the corresponding MCMC estimates are strongly consistent and normally distributed, therefore efficiency will be measured by the asymptotic variance of ergodic averages. For general state spaces several conditions have been stated that guaranty the existence of a central limit theorem, such as uniform, geometric ergodicity or other mixing conditions [83, 55, 73]. A detailed discussion on this issue is available in [83] and [8].

The set of reversible transition kernels with respect to π is a subset of $\mathcal{P}$. In the first part of this chapter we restrict our attention to this subset while in the second part (Sections 2.5 - 2.6 - 2.7) we try to extend the results to non-reversible transition kernels. The difficulty lies in the fact that, for non-reversible transition kernels, we cannot use spectral theory or classical functional analysis tools. Moreover, intuition often fails to support the reasoning or, even worse, intuition can be misleading.

Contrary to what is often done in classical statistical inference when looking for minimum variance estimates, we do not assume any prior knowledge of the function that we want to estimate. Therefore in Section 2.3, given two transition kernels P and Q in $\mathcal{P}$, we consider orderings of the form P better than Q that imply that, for

all functions that obey the central limit theorem, the asymptotic variance of MCMC estimates obtained via P is smaller than the asymptotic variance of estimates obtained via Q (Peskun ordering).

In Section 2.4 we provide a necessary and sufficient conditions for a transition kernel to have smaller asymptotic variance for all functions to be estimated (covariance ordering).

Section 2.3.1 shows that, in finite state spaces, the Peskun ordering induces an ordering on the eigenvalues of the corresponding transition matrices but not on the absolute values of the eigenvalues. The distinction is quite relevant: fast convergence to stationarity in total variation distance is reached by having small eigenvalues in absolute value, while small asymptotic variance of MCMC estimates is achieved by having small eigenvalues. Therefore, unless the operators used are positive (i.e. have positive eigenvalues or a positive spectrum), we are faced with conflicting goals.

In Section 2.3.2 we try to extend to general state spaces the result on ordering the eigenvalues. Here the difficulty lies in the fact that we cannot talk about eigenvalues anymore but we need to introduce the concept of a spectrum.

In the final part of this chapter (Sections 2.8 and 2.9) using the instruments previously developed, we compare the performance of reversible and non-reversible transition kernels.

2.2 Preliminaries

In this section we set up the notation needed in the sequel and review some of the theory on MCMC. Let $\{X_n\}_{n=1}^{\infty}$ denote a Markov chain generated with the transition kernel $P \in \mathcal{P}$ and having initial distribution equal to the stationary distribution π. This gives us a *stationary Markov chain* that is, the distribution of X_n does not

depend on n. Say we are interested in estimating the expectation of some function f:

$$E_\pi[f(X)] = \int f(x)\pi(dx) = \mu. \tag{2.2}$$

Then, the MCMC estimate of μ will be the sample average

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Given a nonzero measure on the state space, φ, a Markov chain is said to be φ-*irreducible* if, for any point x and any measurable set A such that $\varphi(A) > 0$, there exists an integer n such that $P^n(x, A) > 0$. For a φ-irreducible Markov chain, conditional on the starting position $X_1 = x$, $\hat{\mu}_n$ converges almost surely to μ for π-almost all x (Birkhoff ergodic theorem, [21]). If furthermore the chain is *Harris recurrent*, then almost sure convergence holds for any initial distribution (Proposition 17.1.6 [55]). The same principle is true for the central limit theorem. If the chain is Harris recurrent the central limit theorem holds for all initial distributions if it holds for the invariant distribution. So, without loss of generality, we can work with stationary Markov chains when dealing with the strong law of large number or the central limit theorem.

In Section 2.3.2 we state a general sufficient condition for a *central limit theorem* to hold for a function f that is square integrable with respect to π. For now let us recall that the theorem says

$$\sqrt{n}\,(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} N(0, v(f, P)) \tag{2.3}$$

where the notation $\xrightarrow{\mathcal{D}}$ means convergence in distribution and $v(f, P)$ is the limit,

as n tends to infinity, of

$$\begin{aligned}\sigma_n^2 &= n \operatorname{Var}_\pi[\hat{\mu}_n] \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}_\pi[f(X_i), f(X_j)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}_\pi[f(X_i)] + \frac{2}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\operatorname{Cov}_\pi[f(X_i), f(X_j)].\end{aligned}$$

Since all the expectations are taken under the stationary distribution π, $\operatorname{Var}_\pi[f(X_n)]$ does not depend on n and $\operatorname{Cov}_\pi[f(X_n), f(X_{n+k})]$ does not depend on n for fixed k. Hence

$$\sigma_n^2 = \operatorname{Var}_\pi[f(X_i)] + \frac{2}{n}\sum_{k=1}^{n-1}(n-k)\operatorname{Cov}_\pi[f(X_i), f(X_{i+k})]. \tag{2.4}$$

To simplify the notation let

$$\gamma_0 = \operatorname{Var}_\pi[f(X_i)]$$

and

$$\gamma_k = \operatorname{Cov}_\pi[f(X_i), f(X_{i+k})] \tag{2.5}$$

which is the lag k autocovariance of the stationary time series $\{f(X_n)\}_{n=1}^{\infty}$. If the central limit theorem holds, we might expect the limiting variance to be the limit of (2.4) as $n \to \infty$. If this is the case we have ([22], Chapter 3)

$$v(f, P) = \gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k. \tag{2.6}$$

The asymptotic variance in the central limit theorem given in (2.6) defines our criterion for ranking transition kernels.

A transition kernel is *reversible* with respect to π if, for all bounded functions f and g,

$$\iint f(y)g(x)\pi(dx)P(x,dy) = \iint f(x)g(y)\pi(dx)P(x,dy). \tag{2.7}$$

Condition (2.7) is also known as the detailed balance condition. Let $\mathcal{D}$ be the class of transition kernels for which (2.7) is satisfied. If (2.7) holds, then by taking $g(x) = 1$ we have

$$\iint f(y)\pi(dx)P(x,dy) = \iint f(x)\pi(dx)P(x,dy) = \int f(x)\pi(dx)$$

and thus π is the stationary distribution for P so that $\mathcal{D} \subset \mathcal{P}$.

A transition kernel $P \in \mathcal{P}$ defines an *operator* on the Hilbert space $L^2(\pi)$ of square integrable functions with respect to π. The operator corresponding to P is described by the way P acts on a generic element $g \in L^2(\pi)$:

$$(Pg)(x) = E[g(X_t)|X_{t-1} = x] = \int g(y)P(x,dy). \tag{2.8}$$

The inner product on $L^2(\pi)$ is

$$(f,g) = \int f(x)g(x)\pi(dx) = E_\pi[f(x)g(x)].$$

Let $L_0^2(\pi) = \{g \in L^2(\pi) : \int g d\pi = 0\}$ be the subspace of L^2 of zero mean functions. One of the reasons we often restrict to this subspace is that, for zero mean functions f and g, the inner product (f,g) is equal to the covariance of f and g under stationarity

$$(f,g) = \mathrm{Cov}_\pi[f(x), g(x)].$$

The other reason why $L_0^2(\pi)$ is relevant for our purposes is related to its spectrum and will become clear in Section 2.3.2. Another way to describe $L_0^2(\pi)$ is the subspace of $L^2(\pi)$ orthogonal (inner product equal to zero) to the constant functions.

Let P^* be the *adjoint* of P, that is, the unique operator such that

$$(f, Pg) = (P^* f, g), \qquad \forall f, g \in L^2(\pi).$$

An operator is said to be *self-adjoint* if $(f, Pg) = (Pf, g)$, for any function f and g in $L^2(\pi)$. If a transition kernel is reversible then the corresponding operator is self-adjoint.

An operator is *positive* on $L^2(\pi)$, $P \geq 0$, if

$$(Pf, f) = \iint f(x) f(y) P(x, dy) \pi(dx) \geq 0, \qquad \forall f \in L^2(\pi). \tag{2.9}$$

This is not the standard definition of positive operators; more comments on this issue will be given in Section 2.3.2. When referring to a general state space we mean a state spaces equipped with a countably generated σ-field, i.e. generated by a countable collection of subsets of the state space. On finite state spaces an irreducible chain is called *aperiodic* if for some i (and hence for all) the greated common divisor of $\{t > 0 : P(X_t = i | X_0 = i)\}$ is equal to one. On general state spaces an m-cycle for an irredicible chain is a collection $\{E_0, \cdots, E_{m-1}\}$ of disjoint sets such that $P(x, E_j) = 1$ for $j = i + 1 \mod m$ and all $x \in E_i$. The period of the chain is the largest m for wich an m-cycle exists. The chain is aperiodic if $d = 1$.

2.3 Peskun Ordering and Generalizations

In order to compare transition kernels of different Markovian schemes it is useful to refer to the partial ordering introduced by Peskun [64] for discrete state spaces and extended by Tierney [84] to general state spaces.

Definition 2.3.1.
If P and Q are transition kernels on a measurable space with stationary distribution π, then P dominates Q off the diagonal, $P \succeq Q$, if for $\pi-$almost all x in the state space we have

$$P(x, B \setminus \{x\}) \geq Q(x, B \setminus \{x\})$$

for all measurable B.

For a better understanding of this definition let us restrict our attention to finite state spaces. In this setting, P dominates Q off the diagonal if each of the off-diagonal elements of P is greater than or equal to the corresponding off-diagonal elements in Q. This means that P has higher probability of moving around in the state space than Q and therefore the corresponding Markov chain will explore the space in a more efficient way (better mixing). Thus, we expect that the resulting MCMC estimates will be more precise than the ones obtained by averaging along a Markov chain generated via Q. This intuition is stated more rigorously in the next theorem by Tierney [84] which holds on general state spaces.

Theorem 2.3.1.
Let P and Q be reversible transition kernels with stationary distribution π and suppose

$f \in L_0^2(\pi)$. Let

$$v(f,P) = \lim_{n\to\infty} \frac{1}{n} \operatorname{Var}\Big\{ \sum_{i=1}^{n} f(X_i) \Big\} = \lim_{n\to\infty} n \operatorname{Var}[\hat{\mu}_n]$$

where $X_0, X_1, \ldots$ is a Markov chain with initial distribution π and transition kernel P. If $P \succeq Q$, then

$$v(f,P) \leq v(f,Q), \qquad \forall f \in L_0^2(\pi). \tag{2.10}$$

The next Theorem also appears in [84]:

Theorem 2.3.2.
If P and Q have stationary distribution π and $P \succeq Q$, then $Q - P$ is a positive operator on $L^2(\pi)$.

In the next section we show how the Peskun ordering, Definition 2.3.1, implies an ordering on the eigenvalues of the corresponding transition matrices in finite state spaces. We extend the result to general state spaces in Section 2.3.2.

2.3.1 Finite State Spaces

Let $\{\lambda_{0P}, \lambda_{1P} \cdots\}$ be the eigenvalues of P, arranged in decreasing order, and let $\{e_{0P}, e_{1P}, \cdots\}$ be the corresponding normalized right eigenvectors, so that $Pe_{jP} = \lambda_{jP}\, e_{jP}, \;\; j = 0, 1, \cdots$. For $P \in \mathcal{P}$ there is an eigenvalue equal to one, λ_{0P}, which is associated with the constant eigenvector. Since this is always the case let us restrict our attention to the eigenvalues associated with non-constant eigenvectors. Reversibility of a transition kernel ensures that the eigenvalues and eigenvectors are real.

Theorem 2.3.3.

For $P, Q \in \mathcal{D}$, if $Q - P \geq 0$, then $\lambda_{iP} \leq \lambda_{iQ}$ for all i.

Proof. Consider the following definition of eigenvalues [4]:

$$\lambda_{iP} = \min_{\substack{(g_j, g_j)=1 \\ j=1,\cdots,i}} \left\{ \max_{\substack{(f,g_j)=0 \\ j=1,\cdots,i}} \frac{(f, Pf)}{(f,f)} \right\}.$$

If $Q - P \geq 0$ then

$$\frac{(f, Qf)}{(f,f)} \geq \frac{(f, Pf)}{(f,f)}, \qquad \forall f \in L^2(\pi)$$

and the result follows since the eigenvalues of a transition matrix and of the corresponding operator in $L^2(\pi)$ are the same (because the defining equation is the same). Q.E.D.

The previous theorem is a known fact for symmetric matrices. In our setting neither P nor Q need to be symmetric but if we consider them as operators on $L^2(\pi)$ they are indeed self-adjoint operators, provided that the detailed balance condition holds.

By Theorem 2.3.2, $P \succeq Q$ implies that $Q - P \geq 0$, thus the Peskun ordering induces an ordering on all the eigenvalues of the two transition matrices. This proof can be generalized to compact operators on Hilbert spaces since their spectra are either empty, finite, or countable with zero as the only limit point [9]. But, as noticed in [8], not many Markov chains have compact transition operators.

In [20] Frigessi at al. identify the subset of matrices in $\mathcal{P}$ which minimize $v(f, P)$ for all possible functions f. They begin by describing the structure of the matrices in $\mathcal{P}$ that have the smallest possible second largest eigenvalue. The procedure is then repeated in order to build a matrix with lowest third eigenvalue, given that the second is already the smallest possible. By iterating, the matrix which is minimal

with respect to the lexicographic order of the eigenvalues within $\mathcal{P}$ is obtained. This matrix gives rise to a Monte Carlo method with smaller asymptotic variance compared with independent sampling since all eigenvalues (except the largest one) turn out to be negative.

2.3.2 General State Spaces

In this section we extend to general state spaces, the results obtained for finite state spaces in Section 2.3.1. The difficulty lies in the fact that, while in finite state spaces we have a finite number of eigenvalues and it makes sense to compare and order eigenvalues of two transition matrices, in general state spaces we cannot talk about eigenvalues anymore but we need to introduce the concept of a spectrum. Let $\sigma(P)$ be the *spectrum* of P considered as an operator on $L^2(\pi)$, that is, the set of λ's such that $\lambda I - P$ is not invertible, where I denotes the identity operator on $L^2(\pi)$. The spectrum includes the eigenvalues, the λ's for which $\lambda I - P$ is not one-to-one. But it also includes the values λ such that $\lambda I - P$ is not onto. For linear operators on finite dimensional vector spaces, one-to-one and onto are equivalent so that $\sigma(P)$ is the set of the eigenvalues of P.

The *norm* of a linear operator on $L^2(\pi)$ is defined by

$$\|P\| = \sup_{\substack{u \in L^2(\pi) \\ u \neq 0}} \frac{\|Pu\|}{\|u\|}$$

where $\|u\|^2 = (u, u)$. The spectrum is a non-empty closed subset of the interval $[-1, +1]$ since the norm of P is less than or equal to one by Jensen's inequality and the norm of an operator bounds the spectrum (Proposition 1.11 (e) p. 239 in [9]). In this setting it does not make sense to say that the spectrum of one operator is smaller than the spectrum of another operator, we can at most compare the suprema of the

spectra and this is what we will do. For reversible geometrically ergodic chains, all the eigenvalues but the principal eigenvalue, $\lambda_{0P} = 1$, are bounded away from ± 1 [73].

When considering a transition kernel as an operator on the subset $L_0^2(\pi)$ of $L^2(\pi)$ of zero mean functions, we eliminate from its spectrum the eigenvalue one associated with constant functions. Unless otherwise stated a transition kernel will be considered as an operator on $L_0^2(\pi)$.

Let $l_P = I - P$ be the *Laplacian operator* of the chain. An operator is invertible if it is one-to-one and onto. In our setting, the Laplacian l_P is invertible if it has a trivial null space when considered as an operator on $L_0^2(\pi)$ (one-to-one) and if its range is the entirety of $L_0^2(\pi)$ (onto). By the definition of spectrum l_P is invertible if and only if the spectrum of P does not contain the point 1. By the open mapping theorem ([9] Chapter 3, Theorem 12.5) invertible means that there exists an inverse which is a bounded operator on $L_0^2(\pi)$.

If a Markov chain has an invertible Laplacian, then the *central limit theorem* (2.3) holds for the stationary chain for every function $f \in L_0^2(\pi)$ [30].

A weaker requirement on the Laplacian is that it is injective (one-to-one). In terms of the spectrum of P this is equivalent to the fact that one is not an eigenvalue. In this case the central limit theorem holds for every function in the range of l_P [30]. For every such function we can still talk about the inverse Laplacian if we restrict the domain of l_P^{-1} to be the range of l_P. In other words, for any f in the range of l_P there exists a $g \in L_0^2(\pi)$ such that $f = l_P g$ so that $g = l_P^{-1} f$.

For a recurrent Markov chain, the only functions in $L^2(\pi)$ satisfying $Pf = f$ or equivalently $l_P f = 0$ (harmonic functions) are π-almost surely constant (Proposition 17.4.1 in [55] and [30]). Thus the operator P is injective and the Laplacian, as an operator on $L_0^2(\pi)$, has a trivial null space $= \{0\}$. Since our chains are recurrent, in our setting the Laplacian is an injective operator.

Let $E_P(\cdot)$ be the *resolution of the identity* associated with P in the spectral theorem [9], that is

$$P = \int \lambda E_P(d\,\lambda).$$

As in [9], for every bounded Borel measurable function g on $\sigma(P)$ define

$$g(P) = \int g(\lambda) E_P(d\,\lambda).$$

In general our integrals will not involve the resolution to the identity but the spectral measure which is defined below. Given a function g in $L_0^2(\pi)$ define $E_{g,P}(\cdot) = (g, E_P(\cdot)g)$ to be the *spectral measure* associated with g.

If P is irreducible, then $E_{g,P}$ is a positive measure on $[-1,+1)$ because atoms in the spectrum are eigenvalues (Proposition 12.29 (c) in [76]) and, as noted before, 1 is not an eigenvalue when considering P as an operator on $L_0^2(\pi)$. Using the above notation we have that $(g, f(P)g) = \int f(\lambda) E_{g,P}(d\,\lambda)$. Let $\lambda_{max,P} = \sup\{\ \lambda : \lambda \in \sigma(P)\ \}$ and $\Lambda_{max,P} = \sup\{\ |\lambda| : \lambda \in \sigma(P)\ \}$. $\Lambda_{max,P}$ is also called the spectral radius. The quantity $1-\Lambda_{max,P}$ is the spectral gap. If a transition kernel P has $1-\Lambda_{max,P} > 0$ we say that is has a spectral gap. Roberts and Rosenthal in [73] show that a Markov chain is geometrically ergodic if and only if it has a spectral gap.

A reversible transition kernel P as an operator on $L_0^2(\pi)$ is a self-adjoint contraction, so the Laplacian is a positive operator and has a square root, $l_P^{\frac{1}{2}}$. If the chain is irreducible then $l_P^{\frac{1}{2}}$ is also injective and self-adjoint and therefore its range is dense and $l_P^{-\frac{1}{2}}$ is also self-adjoint ([9] p. 309). As proved in [44] the range of $l_P^{\frac{1}{2}}$ is indeed the set of functions that have a finite asymptotic variance. Another interesting result from [44] is that, for a stationary, irreducible, reversible transition kernel P,

the variance of a function g in the central limit theorem can be written as

$$v(g,P) = \int_{-1}^{1} \frac{1+\lambda}{1-\lambda} E_{g,P}(d\lambda). \tag{2.11}$$

Denote the *domain* and *range* of an operator A by $D(A)$ and $R(A)$, respectively. An operator on $L_0^2(\pi)$ is said to be *densely defined* if $D(A)$ is dense in $L_0^2(\pi)$. An operator is *positive*, $A \geq 0$, if $(g, Ag) \geq 0,\ \forall g \in L_0^2(\pi)$. Notice that, if we restrict ourselves to the space of real-valued functions in $L_0^2(\pi)$, then the fact that an operator is positive does not imply that the operator is self-adjoint. If, on the other hand, we consider also complex-valued functions, then $A \geq 0$ implies $A = A^*$, where A^* is the adjoint of A. There are functional analysis books such as [9], Theorem 3.8, that claim that the only positive operators are self-adjoint. This is because they are considering complex-valued spaces but this is not explicitly stated in the theorem (but 50 pages before). This fact can be quite misleading. In [47], the authors explicitly consider the space of complex valued functions. Since real valued functions are the only functions we are interested in, from a statistical point of view, we restrict ourselves to such functions when dealing with non-reversible transition kernels so that when we require a non-self-adjoint operator to be positive we do not contradict ourselves. The next lemma and corollary will be used in Section 2.4.

Lemma 2.3.1.

Let $A \geq 0$ be a self-adjoint, injective, bounded operator. Then, for every $g \in D(A)$:

$$(g, Ag) = \sup_{f \in D(A^{-\frac{1}{2}})} \left[2(f,g) - (A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}f)\right].$$

Proof. Since A is positive A^{-1} is also positive. This allows us to take square roots of both A and A^{-1}. Let $h = Ag$ so $g = A^{-1}h$. For every $f \in D(A^{-\frac{1}{2}}) \supset D(A^{-1})$

$$
\begin{aligned}
0 &\leq (A^{-\frac{1}{2}}(f-h), A^{-\frac{1}{2}}(f-h)) \\
&= (A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}f) - 2(A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}h) + (A^{-\frac{1}{2}}h, A^{-\frac{1}{2}}h).
\end{aligned}
$$

Now substitute $h = Ag$ and use the fact that $(f,g) = (g,f)$, which is true in a real Hilbert space but not true in complex Hilbert spaces. Thus

$$
(g, Ag) \geq \left[2(f,g) - (A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}f)\right], \qquad \forall f \in D(A^{-\frac{1}{2}}) \tag{2.12}
$$

and the supremum is achieved by taking $f = h$ since, in this case, the right hand side equals the left hand side in (2.12). Q.E.D.

Corollary 2.3.1.

Suppose A and B are everywhere defined self-adjoint, injective, bounded operators. If the two conditions

$$
(B^{-\frac{1}{2}}f, B^{-\frac{1}{2}}f) \leq (A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}f), \ \ \forall f \in D(A^{-\frac{1}{2}}) \ \ \textit{and} \ \ D(B^{-\frac{1}{2}}) \supset D(A^{-\frac{1}{2}})
$$

are satisfied, then $A \leq B$.

Proof. By Lemma 2.3.1 we have that, for every $g \in D(A) = D(B)$:

$$
\begin{aligned}
(g, Bg) &= \sup_{f \in D(B^{-\frac{1}{2}})} \left[2(f,g) - (B^{-\frac{1}{2}}f, B^{-\frac{1}{2}}f)\right] \\
&\geq \sup_{f \in D(A^{-\frac{1}{2}})} \left[2(f,g) - (A^{-\frac{1}{2}}f, A^{-\frac{1}{2}}f)\right] \\
&= (g, Ag).
\end{aligned}
$$

Q.E.D.

In this chapter we will make extensive use of the following result:

Lemma 2.3.2.

For a transition kernel P with stationary distribution π, the asymptotic variance can be written as:

$$v(g,P) = (g,[2l_P^{-1} - I]g), \qquad \forall g \in D(l_P^{-1}). \tag{2.13}$$

Proof. For any $g \in D(l_P^{-1})$ there exists an $f \in L_0^2(\pi)$ such that $g = l_P f$ so that $Pf = f - g$. Using a result in [30] we can write the asymptotic variance as

$$\begin{aligned} v(g,P) =& \| f \|^2 - \| Pf \|^2 \\ =& \| f \|^2 - \| f - g \|^2 \\ =& (f,f) - (f-g, f-g) \\ =& 2(g,f) - (g,g) \\ =& 2(g, l_P^{-1} g) - (g,g) \\ =& (g,[2l_P^{-1} - I]g). \end{aligned}$$

Q.E.D.

The previous result generalizes the representation of the asymptotic variance given in [39] for finite state spaces. Notice that the transition kernel does not need to be reversible for this lemma to hold.

The next theorem extends Theorem 2.3.3 to general state spaces.

Theorem 2.3.4.

Given reversible transition kernels P and Q with stationary distribution π, suppose $P \succeq Q$, then

$$\lambda_{max,P} \leq \lambda_{max,Q}. \tag{2.14}$$

Proof. It follows directly from Theorem X.4.2 of [14] that, for any bounded self-adjoint operator A on a Hilbert space, we have

$$\lambda_{max,A} = sup_{||f||=1}(f, Af).$$

Thus (2.14) holds whenever $Q-P \geq 0$, and Theorem 2.3.2 finishes the proof. Q.E.D.

2.3.3 A Counterexample

The rate of convergence in total variation distance of P^n (and of weak convergence of X_n) to $\pi(x)$ is governed by the spectral radius, $\Lambda_{max,P}$, which, in finite state spaces is the second largest eigenvalue in absolute value [6, 22]. We thus have conflicting requirements: fast total variation convergence to equilibrium is obtained by having all eigenvalues small in absolute value while good properties in terms of asymptotic variance of ergodic averages are obtained by having small positive and large negative eigenvalues, as (2.11) indicates. Only if the transition kernels are positive operators, that is if the eigenvalues are all positive, are the two goals not in conflict. It has been shown by Liu et al. that the independence Metropolis-Hastings algorithm [50] and the random scan Gibbs sampler [47] are positive operators. In Section 3.7 we prove that the slice sampler is also a positive operator.

The next example shows that the Peskun partial ordering does *not* imply an ordering on the largest eigenvalue in absolute value. Consider the following two transition matrices:

$$A = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \end{pmatrix}$$

and

$$B = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

Matrix A refers to a symmetric random walk with reflecting barriers at the end points. With B, no matter where you are there is equal probability to move to any of the other 2 states. Both transition matrices, being doubly stochastic, have $\pi = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right]$ as their stationary distribution.

Clearly $B \succeq A$. The ordered eigenvalues of the two matrices are

$$\lambda_A = [1,\ 0.5,\ -0.5]$$

and

$$\lambda_B = [1,\ -0.5,\ -0.5].$$

As expected from Theorem 2.3.2,

$$\lambda_{iB} \leq \lambda_{iA}, \qquad \forall i.$$

Consider now the transition matrix given by $C = 0.8\,A + 0.1\,B + 0.1\,I_3$ where I_n is

the identity matrix of dimension n. We have

$$C = \begin{pmatrix} 0.5 & 0.45 & 0.05 \\ 0.45 & 0.1 & 0.45 \\ 0.05 & 0.45 & 0.5 \end{pmatrix}$$

and $\lambda_C = [1,\ 0.45,\ -0.35]$. Again the stationary distribution of C is π, $B \succeq C$ and

$$\lambda_{iB} \leq \lambda_{iC}, \qquad \forall i$$

but $\Lambda_{max,B} > \Lambda_{max,C}$.

2.4 A New Ordering

The Peskun criterion and its generalization given by Tierney order only a limited number of transition kernels. For example, the ordering does not allow comparing two distinct transition matrices having all zeros on the main diagonal or two transition kernels for which $P(x, \{x\}) = 0$ for every x in the state space. The latter includes all Gibbs samplers with continuous full conditional distributions. Furthermore, if only one of the entries of $P - Q$ is "out of order" then P and Q are incomparable. If you have tried the dice puzzle at the beginning of this chapter you have certainly realized how hard it can be to order finite dimensional object, also in this case it is sufficient that the face of one dice is mislabeled to mess up the ordering. With this in mind imagine the difficulties that one would encounter when trying to order infinite dimensional objects. Here is a labeling that works to solve the puzzle $A\ A\ C\ C\ C\ B\ B\ B\ A\ C\ B\ B\ B\ A\ A\ A\ C\ C$ that is, die A has the six faces labeled $1, 2, 9, 14, 15, 16$ etc.

A natural way to define a weaker ordering for comparing more transition kernels is the following.

Definition 2.4.1.
P dominates Q in the covariance ordering, $P \succeq_1 Q$, if $Q - P$ is a positive operator on $L_0^2(\pi)$, that is, if $(f, (Q-P)f) \geq 0$, for every $f \in L_0^2(\pi)$.

To understand the meaning of weaker let us go back to the dice puzzle: it is as if I were to ask now for a labeling such that die A beats die B at least $X/36$ of the time, die B beats die C at least $X/36$ of the time and die C beats die A at least $X/36$ of the time. By allowing X to be less than 21 we make the ordering easier, weaker, in the sense that there are more labeling that satisfy this requirement.

Restricting ourselves to $L_0^2(\pi)$ does not reduce the generality of the previous definition, since

$$(f, Qf) \geq (f, Pf), \qquad \forall f \in L_0^2(\pi)$$

if and only if

$$(f, Qf) \geq (f, Pf), \qquad \forall f \in L^2(\pi).$$

One implication is obvious. For the other, let f in $L^2(\pi)$, then $f_0 = f - \mu$ with $f_0 \in L_0^2(\pi)$ and $(f, Pf) = (f_0, Pf_0) + \mu^2$. Similarly we have $(f, Qf) = (f_0, Qf_0) + \mu^2$ and this gives what we want.

The binary relation $\succeq_1$ defines a partial ordering on the space $\mathcal{D}$ of reversible transition kernels with respect to π, since the following properties hold:

1. **Reflexive**. $P \succeq_1 P$. This follows from the fact that $(f, (P-P)f) \geq 0$ for all $f \in L_0^2(\pi)$.

2. **Antisymmetric**. $P \succeq_1 Q$ and $Q \succeq_1 P$ imply $P = Q$. This means that $(f, (Q-P)f) \geq 0$ and $(f, (P-Q)f) \geq 0$ for all $f \in L^2(\pi)$ imply $P = Q$. In

order to prove this, it is sufficient to show that $(f, Af) = 0$ for all $f \in L^2(\pi)$ implies that A is the zero operator. This in turn is equivalent to

$$(a + b, A(a + b)) = 0, \qquad \forall a, b \in L^2(\pi). \tag{2.15}$$

But since $A \in \mathcal{D}$

$$(a + b, A(a + b)) = (a, Aa) + (b, Ab) + 2(a, Ab).$$

By assumption the first two terms on the right hand side are equal to zero, thus condition (2.15) is equivalent to

$$(a, Ab) = 0, \qquad \forall a, b \in L^2(\pi)$$

which implies that A is the zero operator as required.

3. **Transitive**. $P \succeq_1 Q$ and $Q \succeq_1 R$ implies $P \succeq_1 R$. This is easy to verify since, if $(f, (Q - P)f) \geq 0$ and $(f, (R - Q)f) \geq 0$ for all $f \in L_0^2(\pi)$, then, $(f, (Q - P)f) + (f, (R - Q)f) = (f, (R - P)f) \geq 0$ for all $f \in L_0^2(\pi)$.

Notice that, if we consider also non-self-adjoint operators and move from $\mathcal{D}$ to $\mathcal{P}$, then $\succeq_1$ is not a partial ordering anymore since the antisymmetry property fails. To see this consider a non-reversible transition kernel P and let P^* be its adjoint. Then $(f, Pf) = (f, P^*f)$ so that $P \succeq_1 P^*$ and $P^* \succeq_1 P$ but it is not true that $P^* = P$ unless P is self-adjoint.

The condition $P \succeq_1 Q$ is equivalent to

$$\mathrm{Cov}_\pi(f, Qf) \geq \mathrm{Cov}_\pi(f, Pf), \qquad \forall f \in L_0^2(\pi)$$

where

$$\mathrm{Cov}_\pi(f, Qf) = E_\pi[f(X_0)f(X_1)] = \gamma_1, \qquad \forall f \in L_0^2(\pi)$$

is the lag one autocovariance.

Covariance order does not imply Peskun order, as the next example shows, but Peskun order does imply covariance order (Theorem 2.3.2). Hence covariance order is a more general (weaker) criterion.

Consider the following matrices:

$$P = \begin{pmatrix} 0.3 & 0.3 & 0.2 & 0.2 \\ 0.3 & 0.3 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.3 & 0.3 \\ 0.2 & 0.2 & 0.3 & 0.3 \end{pmatrix}$$

and

$$A = x^T x = \begin{pmatrix} 0.1 & 0.1 & -0.1 & -0.1 \\ 0.1 & 0.1 & -0.1 & -0.1 \\ -0.1 & -0.1 & 0.1 & 0.1 \\ -0.1 & -0.1 & 0.1 & 0.1 \end{pmatrix}$$

where $x = [\ \sqrt{0.1},\ \sqrt{0.1},\ -\sqrt{0.1},\ -\sqrt{0.1}\ ]$. The matrix P is doubly stochastic therefore it is a transition matrix with uniform stationary distribution. The matrix A is positive definite. Let $Q = P + A$. Since the row and column sums of A are zero, Q is again a doubly stochastic matrix so that both P and Q have the same stationary distribution. We have that $Q - P = A \geq 0$ therefore $P \succeq_1 Q$, but it is not true that $P \succeq Q$ because this would imply that the matrix $Q - P$ has all negative off diagonal

elements which is not true.

The ordering we have introduced is equivalent to Löwner partial ordering, ($\geq_L$), on positive, bounded, linear operator on a Hilbert space [51], [5]. Löwner ordering is defined on positive operators therefore we need to consider the Laplacian of P, $l_P = I - P$, instead of P. Since $P \succeq I$ for every $P \in \mathcal{P}$, we have that $l_P \geq 0$.

Definition 2.4.2.

Given two positive, bounded, self-adjoint, linear operators on a Hilbert space, l_P, l_Q, we say that l_P dominates l_Q in the Löwner sense, $l_P \geq_L l_Q$, if $l_P - l_Q \geq 0$.

The following conditions are trivially equivalent:

1. $P \succeq_1 Q$ i.e. $Q - P \geq 0$;

2. $l_P \geq_L l_Q$ i.e. $l_P - l_Q \geq 0$.

A variety of inequalities are obtainable, for any partial ordering, once the order-preserving functions are identified. For the Löwner ordering or better for a generalization of it that does not require the operators to be positive, the following theorem characterizes the class of order preserving functions [51], [5]. Let $f(x)$ be a bounded real-valued function of a real variable x defined in an interval I. Consider a bounded self-adjoint operators A in a Hilbert space H whose spectrum lies in the domain of $f(x)$. Then by $f(A)$ we mean the self-adjoint operator defined as $f(A) = \int f(\lambda) E_A(d\lambda)$.

Theorem 2.4.1.

A necessary and sufficient condition for a continuous real-valued function f on (I_1, I_2) to have the property that $f(A) \leq f(B)$ for all pairs of bounded, self-adjoint operators A and B with $\sigma(A), \sigma(B) \subseteq (I_1, I_2)$ and $A \leq B$ is that f is analytic in (I_1, I_2), can be analytically continued into the whole upper half-plane, and represents there an analytic function with the property $(\mathrm{Im}\ f) \geq 0$ for all z with $(\mathrm{Im}\ z) > 0$.

Further characterizations of such classes of functions can be found in [45]. A function that satisfies the conditions of Theorem 2.4.1 is

$$h(x) = \frac{ax+b}{cx+d} \qquad \text{with } \; ad-bc > 0$$

either in $x > -\frac{d}{c}$ or $x < -\frac{d}{c}$.

Take $a = b = d = 1$ and $c = -1$, then $ad - bc = 2 > 0$, and

$$h(x) = \frac{1+x}{1-x}$$

preserves the ordering for $x < 1$. Thus

$$P \succeq_1 Q \quad \text{if and only if} \quad Q \geq P \quad \text{if and only if} \quad \frac{I+Q}{I-Q} \geq \frac{I+P}{I-P}.$$

The reason for introducing the new ordering defined in (2.4.1) is given in the next Theorem:

Theorem 2.4.2.
Let P and Q be reversible and irreducible transition kernels with stationary distribution π. Then

$$v(f,P) \leq v(f,Q), \qquad \forall f \in L_0^2(\pi) \tag{2.16}$$

if and only if $P \succeq_1 Q$.

Proof. Let us consider two cases depending on whether the Laplacian is an invertible operator on $L_0^2(\pi)$.

Case (1) : l_P invertible. Let $h(l_P) = \frac{2}{l_P} - I = \frac{I+P}{I-P}$. Using Lemma 2.3.2 we have that (2.16) holds if and only if, for all $f \in L_0^2(\pi)$,

$$(f, h(l_P)f) \leq (f, h(l_Q)f) \tag{2.17}$$

which, by definition is equivalent to

$$h(l_P) \leq h(l_Q) \tag{2.18}$$

and by Theorem 2.4.1, this is true if and only if

$$Q - P \geq 0. \tag{2.19}$$

Case (2): if l_P is not invertible consider the following two sub-cases.

Case (2.1): (2.19) $\rightarrow$ (2.16). Let $K_{\epsilon P} = I-(1-\epsilon)P$ for $0 < \epsilon < 1$. $K_{\epsilon P}$ is invertible since its spectrum $\sigma(K_{\epsilon P}) \subseteq (\epsilon, 2-\epsilon)$ does not contain zero. Furthermore $h(K_{\epsilon P})$ is also invertible since its spectrum is

$$\sigma(h(K_{\epsilon P})) = h(\sigma(K_{\epsilon P})) \subseteq \left(\frac{\epsilon}{2-\epsilon}, \frac{2-\epsilon}{\epsilon}\right).$$

Then, for all $0 < \epsilon < 1$, $Q - P \geq 0$ implies $K_{\epsilon Q} \leq K_{\epsilon P}$ and from case (1) this is true if and only if

$$(f, h(K_{\epsilon,Q})f) \geq (f, h(K_{\epsilon,P})f), \qquad \forall f \in L_0^2(\pi). \tag{2.20}$$

We now want to take the limit as $\epsilon \to 0$. Consider

$$(f, h(K_{\epsilon,P})f) = \int \frac{1+(1-\epsilon)\lambda}{1-(1-\epsilon)\lambda} E_{fP}(d\lambda).$$

The derivative of the integrand with respect to ϵ is

$$\frac{-2\lambda}{[1-(1-\epsilon)\lambda]^2}$$

thus, for $\lambda \in [-1, 0)$ the integrand is increasing in ϵ while for $\lambda \in [0, +1)$ the integrand is decreasing. This suggests that we break the integral over these two subsets of the spectrum

$$(f, h[K_{\epsilon,P}]f) = \int_{-1}^{0} \frac{1+(1-\epsilon)\lambda}{1-(1-\epsilon)\lambda} E_{fP}(d\lambda) + \int_{0}^{1} \frac{1+(1-\epsilon)\lambda}{1-(1-\epsilon)\lambda} E_{fP}(d\lambda).$$

For every $\lambda \in \sigma(P)$ and every $\epsilon \in (0, 1)$ the integrals are finite by construction, therefore a modified version of the standard monotone convergence theorem ([21] p. 50) can be used to take the limit inside the integral and we get that (2.20) implies (2.16).

Case (2.2): (2.16) $\to$ (2.19).

Suppose (2.16) holds. Then from the properties of the Laplacian recalled in Section 2.3.2 and in particular from the fact that the range of $l_Q^{\frac{1}{2}}$ is the set of functions that have a finite asymptotic variance [44], it follows

$$v(f, P) \leq v(f, Q) < \infty, \qquad \forall f \in R(l_Q^{\frac{1}{2}})$$

and

$$R(l_Q^{\frac{1}{2}}) \subseteq R(l_P^{\frac{1}{2}}).$$

It follows that

$$(l_P^{-\frac{1}{2}} f, l_P^{-\frac{1}{2}} f) \leq (l_Q^{-\frac{1}{2}} f, l_Q^{-\frac{1}{2}} f), \qquad \forall f \in R(l_Q^{\frac{1}{2}}) = D(l_Q^{-\frac{1}{2}}) \tag{2.21}$$

and since the hypothesis of Corollary 2.3.1 are satisfied we have $l_Q \leq l_P$, that is, $P \succeq_1 Q$. Q.E.D.

One possible application of Theorem 2.4.2 is given in the following corollary. If we have two transition kernels P and Q having the same stationary distribution there are different possible strategies to run our Markov chain. We could choose one of the two transition kernels and iterate it, obtaining P^n or Q^n respectively. Otherwise we could combine the two basic steps via composition, obtaining a hybrid sampler. If we know that one of the two original kernels is better than the other in the covariance ordering, then the next corollary gives guidelines on how to combine to two kernels in an efficient way.

Corollary 2.4.1.
If $P \succeq_1 Q$ then $P^3 \succeq_1 PQP$ and $QPQ \succeq_1 Q^3$. The first inequality follows from

$$I - P \geq I - Q$$

by multiplying on both sides by Q. The second inequality follows by multiplying on both sides by P.

Another interesting theorem related to Löwner ordering is the following [32]:

Theorem 2.4.3.
If $A \geq 0$ and $B \geq 0$ are Hermitian matrices, then $A \leq B$ if and only if $R(A) \subseteq R(B)$ and $\lambda_{max}(AB^+) \leq 1$.

Here B^+ is any generalized inverse (not necessarily the Moore-Penrose generalized inverse). If P and Q are reversible transition kernels with respect to π, then $A = l_Q$ and $B = l_P$, are positive self-adjoint operators and

$$P \succeq_1 Q$$

$$\Updownarrow$$

$$v(f,P) \leq v(f,Q), \qquad \forall f \in L_0^2(\pi)$$

$$\Updownarrow$$

$$l_P^{-1} \leq l_Q^{-1}$$

$$\Updownarrow$$

$$l_Q \leq l_P$$

$$\Updownarrow$$

$$l_Q^{\frac{1}{2}} \leq l_P^{\frac{1}{2}}.$$

These implications follow from the fact that the function $h(x) = x^a$ preserves the Löwner ordering when $0 \leq a \leq 1$ while the function $h(x) = \frac{1}{x}$ reverses the ordering [5]. By Theorem 2.4.3 it follows that $P \succeq_1 Q$ implies

$$R(l_P^{\frac{1}{2}}) \subseteq R(l_Q^{\frac{1}{2}}).$$

From [44] recall that if a function f is in the range of $l_P^{\frac{1}{2}}$ then the MCMC estimate of $E_\pi[f(X)]$ obtained using P as the transition kernel, has finite asymptotic variance in the central limit theorem. The previous chain of equivalence relations tells us that if the estimate of a function has finite asymptotic variance under P, than it also

has finite asymptotic variance under Q whenever P dominates Q in the covariance ordering. Theorem 2.4.3 is only stated for finite dimensional state spaces but we believe that an equivalent theorem for operators on Hilbert spaces holds. It is true that the part of the theorem that we have actually used holds on general Hilbert spaces, that is $0 \leq A \leq B$ imply $R(A) \subseteq R(B)$.

We finally report another result related to Löwner ordering. We have not made much use of it but we believe it could lead to interesting results when comparing transition matrices in terms of the covariance ordering. In [2] the authors characterize the class of functions of more than one variable that preserve Löwner ordering. That is functions f such that $A \leq A_1$ and $B \leq B_1$ imply

$$f(A, B) + f(A_1, B_1) \geq f(A, B_1) + f(A_1, B)$$

where the matrices involved in the comparison are Hermitian matrices. Functions such that

$$f(p(A, B) + (1 - p)(A_1, B_1)) \leq pf(A, B) + (1 - p)f(A_1, B_1)$$

for $p \in [0, 1]$ are also characterized in the same paper.

2.5 Non-Reversible Transition Kernels

Reversibility of a transition kernel with respect to π implies that π is the stationary distribution of the corresponding Markov chain, but reversibility is a much stronger condition than (2.1). While (2.1) places restrictions only on the marginal distribution of X_t, (2.7) places restrictions on the joint distribution of (X_t, X_{t+1}) by requiring that, when X_t has the distribution π, then (X_t, X_{t+1}) has the same joint distribution as

(X_{t+1}, X_t).

Reversibility is not necessary for MCMC, only having the correct stationary distribution is. However, reversibility of a transition kernel ensures that the corresponding operator on $L_0^2(\pi)$ is self-adjoint and this is a very appealing property when studying the behavior of our Markov chain. We can use spectral theory and this makes the analysis much easier. Moreover the only simple way to show that an update mechanism has a specified stationary distribution is to show that it is reversible with respect to that distribution. However, it is a very common practice to construct a Markov chain for Monte Carlo that is non-reversible by combining reversible elementary update steps by composition. If P and Q are reversible and have the same stationary distribution, then PQ also has the same stationary distribution but is reversible only if P and Q are commuting operators, which very rarely holds. Recently there has been a growing interest in non-reversible transition kernels since [38], [60] and [12] constructed non-reversible Markov chains and showed that they have better properties in terms of convergence to stationarity in total variation distance than other reversible operators.

In this section we restrict our attention to transition kernels P that are not self-adjoint but for which l_P is invertible. One important fact is that Lemma 2.3.2 still holds in this setting and thus we have

Corollary 2.5.1.

Let P and Q be irreducible transition kernels with stationary distribution π such that both l_P and l_Q are invertible. Then

$$v(f, P) \leq v(f, Q), \qquad \forall f \in L_0^2(\pi)$$

if and only if $l_P^{-1} \leq l_Q^{-1}$.

Proof. The proof follows directly from the representation of the asymptotic variance

in the central limit theorem that appears in equation (2.13). Q.E.D.

Lemma 2.5.1.

Let A be a injective positive linear operator defined on a subspace $V = D(A)$ with inverse A^{-1} defined on $R(A) = D(A^{-1})$. For every $g \in D(A^{-1})$

$$(g, A^{-1}g) = \sup_{f \in D(A)} \big[(f, g) + (Af, A^{-1}g) - (f, Af)\big].$$

Proof. Since $g \in D(A^{-1})$ there exists $h = A^{-1}g \in D(A)$. For every $f \in D(A)$:

$$\begin{aligned} 0 &\leq (f - h, A(f - h)) \\ &= (f, Af) - (f, Ah) - (h, Af) + (h, Ah). \end{aligned}$$

It follows, substituting back $g = Ah$, that:

$$(g, A^{-1}g) \geq \big[(f, g) + (Af, A^{-1}g) - (f, Af)\big], \qquad \forall f \in D(A)$$

and the supremum is achieved by taking $f = h$. Q.E.D.

Corollary 2.5.2.

Let A and B be positive and invertible operators such that

$$B^*B^{-1} = A^*A^{-1}. \tag{2.22}$$

Then $A - B \geq 0$ implies $B^{-1} - A^{-1} \geq 0$.

Proof. Apply Lemma 2.5.1 with A replaced by B. The condition $B^*B^{-1} = A^*A^{-1}$ is needed so that $(Bf, B^{-1}g) = (Af, A^{-1}g)$ for all f and g. Q.E.D.

Notice that $B^*B^{-1} = A^*A^{-1}$ or, equivalently, $(BA^{-1})^* = A^{-1}B$, automatically holds if A and B are self-adjoint. Moreover if (2.22) holds, either both A and B are

self-adjoint or neither is.

Corollary 2.5.3.

Let A and B be positive and invertible operators such that (2.22) *holds. Then $B^{-1} - A^{-1} \geq 0$ implies $A - B \geq 0$.*

Proof. Apply Lemma 2.5.1 with A replaced by A^{-1}. Q.E.D.

Theorem 2.5.1.

Let P and Q be irreducible transition kernels with stationary distribution π such that both l_P and l_Q are invertible. Assume that $(l_Q)^* l_Q^{-1} = (l_P)^* l_P^{-1}$. Then

$$v(f, P) \leq v(f, Q), \qquad \forall f \in L_0^2(\pi)$$

if and only if $P \succeq_1 Q$.

Proof. From Corollary 2.5.1 $v(f, P) \leq v(f, Q)$ for every $f \in L_0^2(\pi)$ if and only if $l_P^{-1} \leq l_Q^{-1}$. By Corollary 2.5.2 and 2.5.3, if $(l_Q)^* l_Q^{-1} = (l_P)^* l_P^{-1}$ this is equivalent to $P \leq Q$. Q.E.D.

Proposition 1. *Let $P^\diamond = \frac{P+P^*}{2}$, where P^* is the adjoint of P. Then, for all $f \in L^2(\pi)$*

$$\begin{aligned}(f, P^\diamond f) &= \left(f, \frac{P + P^*}{2} f\right) \\ &= \frac{1}{2}[(f, Pf) + (f, P^* f)] \\ &= \frac{1}{2}[(f, Pf) + (Pf, f)] \\ &= (f, Pf).\end{aligned}$$

This means that every statement about (f, Pf) is actually is a statement involving only the self-adjoint part $P^\diamond$ of P.

The following example shows that the implication

$$l_P \geq l_Q \rightarrow l_Q^{-1} \geq l_P^{-1} \tag{2.23}$$

does not hold in general. Consider the following transition matrices:

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

$$P^* = P^{-1} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Then

$$P^\diamond = \frac{P + P^*}{2} = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

All transition matrices being doubly stochastic have the uniform distribution as their stationary distribution. The state space can be viewed as a circle, with states labeled from 0 to $n-1$ and the n-th state coincides with the origin, zero. The given matrices refer to the case where $n = 5$. Both P and P^* are non-symmetric (non-self-adjoint) and represent deterministic walks along the circle in anti-clockwise and clockwise

directions respectively. They produce estimates with zero asymptotic variance for every function in $L_0^2(\pi)$. On the other hand, $P^\diamond$ represents a symmetric random walk on the circle. By letting n increase we can make the asymptotic variance of ergodic averages obtained using a symmetric random walk as large as we wish. This means that:

$$(f, l_P^{-1} f) < (f, l_{P^\diamond}^{-1} f), \qquad \forall f \in L^2(\pi).$$

Because of Remark 1 we also have

$$(f, l_P f) = (f, l_{P^\diamond} f), \qquad \forall f \in L^2(\pi)$$

therefore (2.23) with $Q = P^\diamond$ does not hold. Notice that condition (2.22) is not verified in this setting since $P^\diamond$ is self-adjoint while P is not.

2.6 Constructing the Inverse

In this section we provide guidelines on how to construct an $n \times n$ matrix representing the inverse Laplacian, l_P^-. The spectrum of the Laplacian, considered as an operator on $L^2(\pi)$, contains the point zero, therefore it is not invertible. We want an operator $l_P^- : L^2(\pi) \to L^2(\pi)$ such that its restriction to $L_0^2(\pi)$ behaves like l_P inverse, that is, such that $l_P^- l_P f_0 = f_0$ and $l_P l_P^- f_0 = f_0$ for every $f_0 \in L_0^2(\pi)$. We will then take the matrix representation of this operator. Define

$$\tilde{P}(x, y) = \frac{\sqrt{\pi(x)} P(x, y)}{\sqrt{\pi(y)}}. \tag{2.24}$$

$\tilde{P}(x, y)$ is the matrix representation of a linear map $\tilde{P}$. If P is a self-adjoint operator, that is, if the detailed balance condition holds, then $\tilde{P}(x, y) = \tilde{P}(y, x)$. Assume that $\pi(x) > 0$ for all x, which follows from the fact that P is φ-irreducible if φ is the

counting measure (which is the default measure when talking about irreducibility on finite state spaces). Define

$$(Tf)(x) = \sqrt{\pi(x)}f(x). \tag{2.25}$$

T is another linear map whose matrix representation is a diagonal matrix with $t_{i,i} = \sqrt{\pi(x_i)}$. If P is irreducible, then $\pi(x) > 0$ for all x, thus T is invertible and

$$(T^{-1}f)(x) = \frac{f(x)}{\sqrt{\pi(x)}}. \tag{2.26}$$

The matrix representation of the linear map T^{-1} is again a diagonal matrix with diagonal elements equal to $\frac{1}{\sqrt{\pi(x_i)}}$. We can now rewrite $\tilde{P}$ as

$$\tilde{P} = TPT^{-1}. \tag{2.27}$$

Another way to indicate (2.27) is by the commutative diagram

$$\begin{array}{ccc} L^2(\pi) & \xrightarrow{P} & L^2(\pi) \\ {\scriptstyle T^{-1}}\uparrow & & \downarrow{\scriptstyle T} \\ E & \xrightarrow{\tilde{P}} & F \end{array}$$

where we write E and F for the Hilbert spaces that are the domain and codomain of $\tilde{P}$. They are, of course, finite-dimensional vector spaces, the question is what inner product they have. The diagram shows that

$$E \xrightarrow{T^{-1}} L^2(\pi) \xrightarrow{T} F$$

thus E and F are the same Hilbert space. Denote temporarily the inner product on E by $(\cdot,\cdot)_E$. Then, since T is a Hilbert space isomorphism, [9], for $f, g \in E$

$$(f,g) := (T^{-1}f, T^{-1}g) = \sum_x \frac{f(x)}{\sqrt{\pi(x)}} \frac{g(x)}{\sqrt{\pi(x)}} \pi(x).$$

So E is $\mathbb{R}^n$ with the usual inner product

$$(f,g)_E = \sum_x f(x)g(x).$$

We might also denote E as $L^2(\nu)$ where ν is the counting measure, $\nu(x) = 1$ for all x. Thus our commutative diagram becomes

$$\begin{array}{ccc} L^2(\pi) & \xrightarrow{P} & L^2(\pi) \\ {\scriptstyle T^{-1}}\uparrow & & \downarrow{\scriptstyle T} \\ L^2(\nu) & \xrightarrow{\tilde{P}} & L^2(\nu) \end{array}$$

Let us now see what properties $\tilde{P}$ inherits from P. Since P is a stochastic matrix we have that $P\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ denotes the column vector of ones. Thus

$$\tilde{P}\sqrt{\pi} = TPT^{-1}\sqrt{\pi} = TP\mathbf{1} = T\mathbf{1} = \sqrt{\pi}. \tag{2.28}$$

Similarly, $\pi^T P = \pi^T$, since π is the stationary distribution for P. It follows that

$$\sqrt{\pi^T}\tilde{P} = \sqrt{\pi}^T TPT^{-1} = \pi^T PT^{-1} = \pi^T T^{-1} = \sqrt{\pi^T}. \tag{2.29}$$

Thus, $\tilde{P}$ has the same right and left eigenvectors, namely $\sqrt{\pi}$, corresponding to the eigenvalue 1. Notice that $\sqrt{\pi} = T\mathbf{1}$, so $T : L^2(\pi) \to L^2(\nu)$ maps the constants to multiples of $\sqrt{\pi}$. This suggests that, in order to construct the inverse of l_P, we perform

a singular value decomposition on $l_{\tilde{P}} = I - \tilde{P} = UDV^T$. Then $l^-_{\tilde{P}} = VD^-U^T$ is the Moore-Penrose generalized inverse of $l_{\tilde{P}}$ [67]. From (2.28) it follows that $l_{\tilde{P}}\sqrt{\pi} = 0$, thus $DV^T\sqrt{\pi} = 0$. Similarly, from (2.29), $\sqrt{\pi}l_{\tilde{P}} = 0$, thus $DU^T\sqrt{\pi} = 0$. If we denote by V_j and U_j the j^{th} column of V and U respectively, we have that, for all j

$$d_{jj}(V_j, \sqrt{\pi}) = 0$$

and

$$d_{jj}(U_j, \sqrt{\pi}) = 0$$

thus, for the j^* such that $d_{j^*j^*} \neq 0$,

$$(V_{j^*}, \sqrt{\pi}) = 0$$

and

$$(V_{j^*}, \sqrt{\pi}) = 0$$

There are $n - 1$ non-zero $d_{j^*j^*}$, and the corresponding collection of V_{j^*} spans the subspace of $L^2(\nu)$ orthogonal to $\sqrt{\pi}$. A similar reasoning holds for U_{j^*}. Using the maps T and T^{-1} we now move everything back to $L^2(\pi)$

$$l^-_P = T^{-1}VD^-U^TT \tag{2.30}$$

and

$$l_P = T^{-1}UDV^TT \tag{2.31}$$

Some of the properties of the operators defined in (2.30) are studied in the sequel. First notice that, since $UU^T = VV^T = I$, l_P^- and l_P commute, that is

$$l_P^- l_P = T^{-1}VD^-DV^TT = T^{-1}UD^-DU^TT = l_P l_P^-.$$

From $T\mathbf{1} = \sqrt{\pi}$ and $DV^T\sqrt{\pi} = 0$ it follows that l_P, $l_P^- l_P$ (and hence also $l_P l_P^-$) annihilate constant vectors. Furthermore, for every $f_0 \in L_0^2(\pi)$, $l_P^- l_P f_0 = f_0$ and $l_P l_P^- f_0 = f_0$ as requested. This is easy to verify if we pick the columns of V as a basis for $L^2(\nu)$. One column is $\sqrt{(\pi)}$ and it is the image of $\mathbf{1}$ under T. Take the images under T of the other columns as a basis for L_0^2. Now, $l_P^- l_P f_0 = T^{-1}VD^-DV^TTf_0$, and D^-D is a diagonal matrix with all ones except for a zero on the main diagonal. Furthermore $VV^T = T^{-1}T = I$, and the result follows.

Consider finally the operator $I - l_P^- l_P$, that, with an excess of notation, we could define to be $l_{l_P^- l_P}$. For every $f \in L^2(\pi)$, $l_{l_P^- l_P} f = E_\pi(f) = \sum_x f(x)\pi(x)$. This follows from the the fact that a generic element f in $L^2(\pi)$ can be written as $f = f_0 + \pi^T f$ where $f_0 \in L_0^2(\pi)$, and $\pi^T f = \mu$ is the mean of f with respect to π. Thus

$$l_{l_P^- l_P} f = (I - l_P^- l_P)f = f - (l_P^- l_P)(f_0 + \pi^T f) = f - (l_P^- l_P)f_0 = f - f_0 = \pi^T f.$$

2.7 Examples

In this section we analyze some non-reversible transition kernels by means of the tools developed in the previous part of the chapter. The first example is the same

one studied in [12]. Consider a finite state space, $\{1, 2, \cdots, n\}$, with the uniform distribution, $\pi(x) = \frac{1}{n}$ as the target distribution. The nearest-neighbor symmetric random walk with holding probabilities of $\frac{1}{2}$ at each end is a reversible Markov chain converging to π. In order to avoid the diffusive behavior of the random walk, Diaconis et al. [12] propose to enlarge the state space by introducing an additional copy of each state. We relabel state s as $(+, s)$ and label its copy $(-, s)$, for $s = 1, \cdots, n$. The transition matrix considered in [12] switches between copies at rate $\frac{c}{n}$ for some value of $0 \leq c \leq n$. The other possible moves allowed are to the left and they happen $1 - \frac{c}{n}$ of the times. This Markov chain has $\frac{\pi(x)}{2}$ as its stationary distribution on each half of the enlarged state space and hence the marginal distribution on the second component of the state (ignoring the $+$ or $-$ sign) is $\pi(x)$, as required. Notice that, since the stationary distribution is uniform, it does not really matter which two states we collapse to go back to the original state. Any sort of grouping of the states two by two preserves the stationary distribution on the original state space.

Let us relabel the state space in the following way: $(+, s) = s$, and $(-, s) = -s$. For $n = 3$ the Markov chain we study can be represented as in Figure (2.1) and the corresponding transition matrix on the enlarged state space is

$$P_c = \begin{pmatrix} 0 & 1-\frac{c}{3} & 0 & 0 & \frac{c}{3} & 0 \\ 0 & 0 & 1-\frac{c}{3} & \frac{c}{3} & 0 & 0 \\ 0 & 0 & \frac{c}{3} & 1-\frac{c}{3} & 0 & 0 \\ 0 & \frac{c}{3} & 0 & 0 & 1-\frac{c}{3} & 0 \\ \frac{c}{3} & 0 & 0 & 0 & 0 & 1-\frac{c}{3} \\ 1-\frac{c}{3} & 0 & 0 & 0 & 0 & \frac{c}{3} \end{pmatrix} \tag{2.32}$$

The rows and the colums of the matrix are labeled as in Figure (2.1) starting

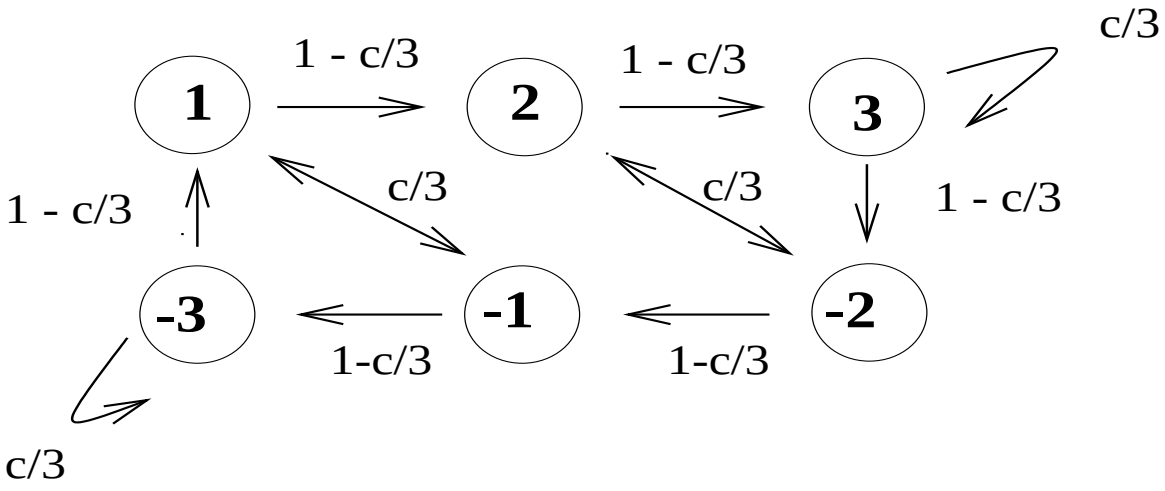

Figure 2.1: Enlarged state space

from state 1 and proceeding clockwise all the way up to state -3. We can think of the state space as a circle and the Markov chain either moves around the circle, with probability $1 - \frac{c}{n}$, or jumps across the circle, with probability $\frac{c}{n}$.

The operator corresponding to P_c is not self-adjoint because P_c does not satisfy the detailed balance condition. Since the stationary distribution is uniform, an operator on $L^2(\pi)$ is self-adjoint if and only if its transition matrix is symmetric.

By taking the inverse of l_{P_c} as described in Section 2.6, and letting the value of c vary over the interval $[0, n]$ we can study the properties of P_c in terms of asymptotic variance of the corresponding MCMC estimates. Using Mathematica 3.0 we find that the eigenvalues of $(I - P_c)^- - (I - P_{c'})^-$ are

$$\lambda_0 = \lambda_1 = \lambda_2 = 0$$
$$\lambda_3 = \frac{6(c - c')}{(c' - 3)(c - 3)}$$
$$\lambda_4 = \frac{2(c - c')}{(c' - 3)(c - 3)}$$
$$\lambda_5 = \frac{3(c - c')}{2(c' - 3)(c - 3)}$$

Since $c \leq 3$ and $c' \leq 3$ these eigenvalues are non-negative if $c \geq c'$. This means that $v(f, P_c) \geq v(f, P_{c'})$ for every $f \in L_0^2(\pi)$ if $c \geq c'$. In other words, the performance of the transition matrix P_c in terms of asymptotic variance of *any function* of interest improves as c decreases towards 0, that is as the probability of moving around the circle increases while the probability of jumping across the circle decreases.

The inverse Laplacian, as a function of the parameter c, is

$$l^-_{P_c} = \frac{1}{36(c-3)} \begin{pmatrix} -45+4c & -27+16c & -9+16c & 9+4c & 27-20c & 45+20c \\ 45-8c & -45+4c & -27+4c & -9-8c & 9+4c & 27+4c \\ 27+4c & 45+20c & -45-20c & -27+4c & -9+16c & 9+16c \\ 9+4c & 27-20c & 45+20c & -45+4c & -27+16c & -9+16c \\ -9-8c & 9+4c & 27+4c & 45-8c & -45+4c & -27+4c \\ -27+4c & -9+16c & 9+16c & 27+4c & 45+20c & -45-20c \end{pmatrix} \tag{2.33}$$

Let $c = 0$ and compute $2(I - P_c)^- - I$ which is the quantity that matters when computing the asymptotic variance of MCMC estimates

$$2l^-_{P_0} - I = \begin{pmatrix} -\frac{1}{6} & \frac{1}{2} & \frac{1}{6} & -\frac{1}{6} & -\frac{1}{2} & -\frac{5}{6} \\ -\frac{5}{6} & -\frac{1}{6} & \frac{1}{2} & \frac{1}{6} & -\frac{1}{6} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{5}{6} & -\frac{1}{6} & \frac{1}{2} & \frac{1}{6} & -\frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{2} & -\frac{5}{6} & -\frac{1}{6} & \frac{1}{2} & \frac{1}{6} \\ \frac{1}{6} & -\frac{1}{6} & -\frac{1}{2} & -\frac{5}{6} & -\frac{1}{6} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{6} & -\frac{1}{6} & -\frac{1}{2} & -\frac{5}{6} & -\frac{1}{6} \end{pmatrix}$$

The previous matrix is not self-adjoint. Its self-adjoint part is

$$[(2l^-_{P_0} - I) + (2l^-_{P_0} - I)] = -\frac{1}{6}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Thus, the asymptotic variance of any function which is square integrable with respect to π, and has zero mean is zero. For this choice of the parameter the Markov chain on the enlarged state space moves around the state space in a deterministic fashion, that is, it circles around clock wise.

The eigenvalues of P_c are

$$\begin{aligned} \lambda_0 &= 1 \\ \lambda_1 &= \frac{2c-3}{3} \\ \lambda_2 &= \frac{3-c-\sqrt{-27+18c+c^2}}{6} \\ \lambda_3 &= \frac{-3+c-\sqrt{-27+18c+c^2}}{6} \\ \lambda_4 &= \frac{3-c+\sqrt{-27+18c+c^2}}{6} \\ \lambda_5 &= \frac{-3+c+\sqrt{-27+18c+c^2}}{6}. \end{aligned}$$

Setting $c = 0$ we get that the only real eigenvalues are $+1$ or -1 thus this transition matrix gives rise to a periodic Markov chain that does not converge to stationarity in total variation distance but produces MCMC estimates with zero asymptotic variance for any function of interest.

Diaconis et al. [12] compute the optimal value of c in terms of convergence to stationarity in χ^2 distance. Roughly this is $c = \sqrt{\log\ n}$ where n is the number of states for the original problem.

2.8 Comparing the Performance of Reversible and Non-Reversible Kernels

Since reversible transition kernels are much easier to analyze it would be nice to have a way of transforming a non-reversible kernel, say P, into a reversible one, say Q, with the same performance in terms of asymptotic variance of MCMC estimates. Recall that the performance of P in terms of asymptotic variance is related to $(I - P)^-$. Furthermore, as we have commented earlier, an operator P and its adjoint have the same asymptotic variance and so does $\frac{P+P^*}{2}$. Therefore $l_Q^{-1} = \frac{(I-P)^{-1}+[(I-P)^{-1}]^*}{2}$ is the inverse Laplacian of an operator with the same performance as P and to obtain Q from l_Q^{-1} we take the inverse and subtract what we get from the identity operator:

$$Q = I - \left\{ \frac{(I-P)^{-1} + [(I-P)^{-1}]^*}{2} \right\}^{-1} \tag{2.34}$$

$$= I - \left\{ \frac{(I-P)^{-1} + (I-P^*)^{-1}}{2} \right\}^{-1}. \tag{2.35}$$

Thus we only need to check if Q is a self-adjoint transition kernel with the proper stationary distribution. Notice that if P is self-adjoint then $Q = P$ as it would be sensible to require.

Q is self-adjoint since the identity is self-adjoint with respect to any stationary distribution, furthermore, given and operator A, $(\frac{A+A^*}{2})^{-1}$ is self-adjoint with respect to the stationary distribution of A and the sum of self-adjoint operators is self-adjoint.

Q has the same performance in terms of asymptotic variance as P since, as

noted before, $(f,(I-Q)^{-1}f) = (f,\frac{(I-P)^{-1}+(I-P^*)^{-1}}{2}f) = (f,(I-P)^{-1}f)$ for all $f \in L_0^2(\pi)$.

Let us now focus on the requirement that Q is a Markov transition kernel with the proper stationary distribution. We need to verify that

1. $\pi Q = \pi$;

2. $Q1 = 1$;

3. $f \geq 0$ implies $Qf \geq 0$;

where 1 represents the constant unitary function. For a transition operator A, the condition $\pi A = \pi$ holds if and only if $\pi Af = \pi f$ for all f in $L^2(\pi)$. Using the inner product notation this means $(1, Af) = (1, f)$, for all f in $L^2(\pi)$. By the definition of adjoint, this is equivalent to $(A^*1, f) = (1, f)$, for all f in $L^2(\pi)$, which implies $A^*1 = 1$ by the Ritz representation theorem. Thus the condition $\pi A = \pi$ is equivalent to $A^*1 = 1$. Notice that the matrix condition $\pi^T A = \pi^T$ and $A\mathbf{1} = \mathbf{1}$ do not represent restrictions on A and A^* as operators on $L_0^2(\pi)$, though they determine the fact that these operators map into $L_0^2(\pi)$. This is easy to see since $A1 = 1$ is equivalent to $l_A 1 = 0$ or $l_A : L^2(\pi) \to L_0^2(\pi)$. Similarly $A^*1 = 1$ is equivalent to $l_{A^*} : L^2(\pi) \to L_0^2(\pi)$. If we start with the operator l_A defined only on $L_0^2(\pi)$ there is a unique extension to $L^2(\pi)$ that behaves like the Laplacian of a Markov transition operator, that is, satisfies $l_P 1 = 0$. A generic function $f \in L^2(\pi)$ can be written as $f_0 + c1$ where $f_0 \in L_0^2(\pi)$ and $c \in R$. By linearity $l_A f = l_A f_0$, and similarly $l_{A^*} f = l_{A^*} f_0$. Hence the only thing to be shown about Q in order to verify requirements (1) and (2) is that $Q : L_0^2(\pi) \to L_0^2(\pi)$, which is true by definition, and we then extend Q and Q^* to $L^2(\pi)$ so that $Q1 = 1$ and $Q^*1 = 1$. More in detail we have that $l_P : L_0^2(\pi) \to L_0^2(\pi)$ and, assuming l_P is invertible, $l_P^{-1} : L_0^2(\pi) \to L_0^2(\pi)$. This implies $(l_{P^*})^{-1} = (l_P^{-1})^* : L_0^2(\pi) \to L_0^2(\pi)$, hence $\frac{1}{2}(l_P^{-1} + (l_{P^*})^{-1}) : L_0^2(\pi) \to L_0^2(\pi)$, again assume invertibility and call l_Q the

inverse. It follows that $l_Q : L_0^2(\pi) \to L_0^2(\pi)$. Extend now l_Q from $L_0^2(\pi)$ to $L^2(\pi)$ by "reconstruction":

$$l_Q f = l_Q[f - (1, f)1], \qquad \forall f \in L^2(\pi)$$

where $(1, f)1$ is nothing but the mean of f under the stationary distribution. Define now $Q = I - l_Q$ where l_Q here is the extension to $L^2(\pi)$. Then Q is an operator on $L^2(\pi)$ and because of the "reconstruction" we have operated to obtain Q, it follows that $Q1 = 1$. Extend now the codomain of l_Q to $L^2(\pi)$ by simply using the fact that $L_0^2(\pi) \subset L^2(\pi)$. Since $l_Q f \in L_0^2(\pi), \qquad \forall f \in L^2(\pi)$ it follows that $(1, l_Q f) = (l_{Q^*}1, f) = 0$. This implies $l_{Q^*}1 = 0$ and $Q^*1 = 1$ as required.

Unfortunately nothing guarantees that conditions (3) is verified so, in general, Q is not a proper transition kernel. For finite state spaces condition (3) is equivalent to requiring all the entries of the transition matrix to be non-negative.

2.9 Comparing the Performance of Kernels Taking CPU Time into Account

From a practitioner's point of view, it can be misleading to compare transition kernels in terms of asymptotic variance. This is due to the fact that different transition kernels take different amounts of time to complete one iteration, that is, to move from X_t to X_{t+1}. Furthermore, the time needed to write the computer code to implement different Markov chains can be different.

Hammersley and Handscomb [34] proposed that "the efficiency of a Monte Carlo process may be taken as inversely proportional to the product of the sampling variance and the amount of labour expended in obtaining this estimates", where the word labour is used with a very broad meaning.

Let τ_P and τ_Q be the CPU time needed to complete one iteration for transition

kernel P and Q respectively. Then P performs better than Q in terms of asymptotic variance, given a fixed amount of CPU time, if

$$\tau_P v(f,P) \leq \tau_Q v(f,Q), \qquad \forall f \in L_0^2(\pi).$$

Because of (2.13), this condition is equivalent to

$$\tau_P(f,[2l_P^{-1}-I]f) \leq \tau_Q(f,[2l_Q^{-1}-I]f), \qquad \forall f \in L_0^2(\pi). \tag{2.36}$$

The comparison in (2.36) requires the computation of the inverse Laplacian and this is often not an easy task. Due to the multiplicative factors τ_P and τ_Q even for self-adjoint operators we cannot find a condition, equivalent to (2.36), that only involves the Laplacian.

The second problem that arises is related to the definition of CPU time needed to complete one iteration. In theory finding τ_P requires that an "optimal" computer program is written in order to run a Markov chain having P as its transition kernel. Since most experimental encodings are less than ideal, τ_P is very hard to measure.

It is true that from a practical point of view, the researcher is really only interested in the time per iteration of the software that he or she has available to run the Markov chains under comparison. These provide good surrogates of τ_P and τ_Q but require that the practitioner writes the computer programs for both transition kernels to be compared, unless externally provided software is available. On the other hand it is desirable that the comparison between transition kernels could be made without having to go through the extra amount of work of implementing all of them. After the comparison is made and a kernel is chosen, the computer program to run only that particular one will be written.

2.10 Harmonic Functions

In the sequel we only assume that our Markov chain $\{X_n\}_{n=1}^{\infty}$ has a stationary distribution π and it is stationary, that is, the marginal distribution of X_n does not depend on n, and in particular, it is equal to π. Let P_π be the probability measure induced by the sequence $\{X_n\}_{n=1}^{\infty}$ ([55], p.66). Such a measure will be our reference measure so almost surely means almost surely with respect to P_π.

Definition 2.10.1.
A function f is said to be harmonic if $Pf = f$, that is if $E[h(X_{n+1})|X_n] = h(X_n)$ almost surely.

Notice that if h is a harmonic function then $h(X_n)$ forms a martingale.

Theorem 2.10.1.
If h is a harmonic function for a stationary Markov chain and $h(X_n)$ has finite expectation, then $h(X_m) = h(X_n)$ almost surely for all m and n.

Proof. Since the sequence $\{X_n\}_{n=1}^{\infty}$ is stationary and $h(X_n)$ has finite expectation

$$\sup_{n\geq 0} E[|h(X_n)|] = E_\pi[|h(X)|] < \infty$$

thus the martingale convergence theorem [82] gives $h(X_n) \to Y$ almost surely for some random variable Y. More precisely, there is a null set A such that, for every $\epsilon > 0$ there is an N_ϵ such that

$$|h(X_n(\omega)) - Y(\omega)| \leq \epsilon, \qquad n \geq N_\epsilon,\ \omega \notin A$$

Define

$$B_{\epsilon,n} = \{\omega \in \Omega : |h(X_n(\omega)) - Y(\omega)| > \epsilon\}$$

Clearly $B_{\epsilon,n} \subset A$ for $n \geq N_\epsilon$. Also, by stationarity of the Markov chain, $P_\pi(B_{\epsilon,n})$ does not depend on n. Hence $P_\pi(B_{\epsilon,n}) = 0$ for all n. Since a countable union of null sets is a null set,

$$B = A \cup \bigcup_{n=1}^{\infty} \bigcup_{k=1}^{\infty} B_{\frac{1}{k},n}$$

is also a null set. For $\omega \notin B$

$$|h(X_n(\omega)) - Y(\omega)| \leq \epsilon$$

holds for all n and all $\epsilon > 0$. Hence

$$h(X_n(\omega)) = Y(\omega), \qquad n \in \mathbb{N}, \omega \notin B$$

which is what was to be proved. Q.E.D.

Corollary 2.10.1.

If there exists a harmonic function in $L^1(\pi)$ that is not almost surely constant, then the Markov chain is decomposable and so is the time-reversed chain.

Proof. By the theorem, with probability one, $h(X_n) = h(X_m)$ for all m and n. Hence for the doubly infinite chain

$$h(X_n) = h(X_0), \qquad n \in \mathbb{Z}$$

with probability one. Thus the chain stays forever in sets of the partition formed by h, i.e. sets of the form

$$\{x : h(x) = c\}$$

for some constant value c. Unless one such set has π-probability one, the chain and the time-reversed chain are both decomposable. If one such set does have π-probability

one, then h is almost surely constant. Q.E.D.

Recall that if a Markov chain is irreducible then it is indecomposable. It follows that if P is irreducible then the range of l_P is dense and for every function in $R(l_P)$ there exists a central limit theorem according to Gordin and Lifšic [30] thus we can compare variances on a dense set in a meaningful way.

Chapter 3

Auxiliary Variables and Slice Samplers

How would you cut a circular cake with no frosting in 8 equal slices with 3 straight cuts?

3.1 Introduction

The slice sampler is a method of constructing a reversible Markov transition kernel with a specified invariant distribution. Auxiliary variables are introduced within the independence Metropolis-Hasting algorithm to ensure that the acceptance rejection probability is always equal to one. This makes the slice sampler better than the corresponding independence Metropolis-Hastings sampler relative to the Peskun ordering.

The idea of introducing auxiliary (latent) variables in Markov chain Monte Carlo sampling arose in statistical physics [81, 15], and was brought into the mainstream statistical literature by Besag and Green [6]. Higdon [36] discusses the use of latent variables to combat slow mixing in sampling from the Ising model and explores applications in Bayesian image analysis. More recently Damien et al. [11] made new contributions to the strategic use of auxiliary variables to set up a Gibbs algorithm having easily sampled full conditionals.

Auxiliary variable techniques exploit the general principle that it is often the case that an apparently complicated n-dimensional problem becomes easier and more

tractable if embedded in an higher dimensional framework. Once the high dimensional solution is found it is projected on the original state space and the original problem is thus solved. This projection procedure is, in our case, obtained when we disregard the value of the auxiliary variable(s) thus obtaining a sample from the target distribution. The same general principle can be used to solve the puzzle: cut the pie into four equal slices with two vertical cuts. Now recall that a pie is a 3 dimensional object! The last cut is an horizontal one, parallel to the table, in the middle of the pie.

In this chapter we first offer an overview of the use of auxiliary variables in Markov chain Monte Carlo sampling (Section 3.2). Some generalizations and variations on the basic idea will also be considered (Section 3.3). A few examples of interesting implementations of auxiliary variables taken from the literature are examined in Section 3.4.

We then concentrate on one of the possible schemes that allows the introduction of auxiliary variables, the slice sampler, and compare its performance with the performance of the independence Metropolis-Hastings algorithm (Section 3.5). Finally, we provide some results on convergence rates of the described sampler (Sections 3.6 - 3.9).

3.2 The Main Idea

Suppose $\pi(x)$ is an unnormalized density with respect to the measure μ and let ν_π be the corresponding probability measure, defined by

$$\nu_\pi(A) = \frac{\int_A \pi(x)\mu(dx)}{\int \pi(x)\mu(dx)}$$

for all measurable A. The key intuition behind the slice sampler is to introduce a new variable, u and to construct the joint distribution of u and x by taking the

marginal distribution for x unchanged and defining the conditional distribution of u given x in a "convenient" way: $\pi(x,u) = \pi(x)\pi(u|x)$. Here convenient has different meanings: the latent variable can be introduced to allow better mixing or faster convergence of the process defined on the enlarged state space with respect to the original process. Alternatively the use of an auxiliary variable can result in an easier to implement or faster algorithm in terms of CPU time. Some of these performance indicators are easily measured, some are not; moreover, it is hard to compare them on a uniform scale. We would like to answer questions like how fast a well programmed implementation can be for a given accuracy, or how much accuracy you can get for a given amount of CPU time. Measuring the CPU time is hard since most experimental encodings we use are less than ideal as noted in Section 2.9. So we look at surrogates such as number of iterations, number of rejections, and so on. These performance indicators should be used with care, keeping in mind that they are surrogates and therefore might be misleading.

The auxiliary variable may have an interpretation in terms of the original problem such as missing values, temperature, regeneration times or link variables for neighboring pair of pixels (refer to Section 3.4 for details), but this is not needed.

In the sequel we will give guidelines on how to specify $\pi(u|x)$. Having done this, a Markov chain $\{X_n, U_n\}_{n=0}^{\infty}$ with $\pi((x,u))$ as its equilibrium distribution is set up using our favorite MCMC method. We will use the short notation $\pi((x,u)) = \pi(x,u)$ since it is not confusing. To define our Markov chain we need to specify a transition kernel $K_{x,u}((x,u)),(x',u'))$ having $\pi(x,u)$ as its stationary distribution. If the MCMC procedure constructed can be shown to be irreducible and aperiodic, than $\pi(x,u)$ is the unique invariant distribution and is also the equilibrium distribution of the chain [83]. It follows that $x_t \xrightarrow{\mathcal{D}} x^* \sim \pi$. The marginal X-chain is therefore the one of interest for inference on $\pi(x)$; the U-chain is an auxiliary construction. Typically the joint transition kernel is given by specifying the two kernels $K_x((x,u),(x',u))$ and

$K_u((x,u),(x,u'))$. Such a choice implies that x is updated conditionally on u and u is updated conditionally on x. This is however not necessary and sometimes it might be useful to update x and u jointly. The *hybrid Monte Carlo algorithm* (Section 3.4) is an interesting example of this.

The simplest choice for the conditional transition functions is

$$K_x((x,u),(x',u)) = \pi(x'|u) \quad \text{and} \quad K_u((x,u),(x,u')) = \pi(u'|x) \tag{3.1}$$

i.e. both u and x are drawn from their full conditional distribution with a Gibbs step.

Once K_x and K_u have been specified, they are combined via systematic scan, random scan or other suitable updating scheme in order to provide realizations $(x_1,u_1),\cdots,(x_N,u_N,)$ from the Markov chain.

It appears clear from the previous description that various degrees of freedom are left to the researcher:

1. the conditional distribution of the auxiliary variable given the variable of interest, $\pi(u|x)$;
2. the transition kernels that define the Markov chain, $K_{x,u}$ or K_x and K_u;
3. the updating schedule between the transition kernels.

All these choices should be made on a case by case basis. To our knowledge there is no general formulation that performs uniformly better in every setting or even in limited classes of problems.

Following [6], guidelines to the specification of $\pi(u|x)$ are given in the sequel. Suppose a factorization of the target distribution, possibly up to a constant of proportionality, is available: $\pi(x) \propto q(x)l(x)$, where $l(x)$ is a nonnegative function. In Bayesian inference $l(\cdot)$ might be the likelihood and $q(\cdot)$ the prior. The integral of $q(x)$

does not need to be finite, in particular the constant function (improper prior) is an acceptable choice for $q(x)$. Given the factorization $\pi(x) \propto q(x)l(x)$ we can take the conditional distribution of u given x to be uniform on the interval $(0, l(x))$. This leads to a joint distribution for u and x which is proportional to $\pi(x, u) \propto q(x)I_{\{u<l(x)\}}(x, u)$, where $I_A(x)$ is the indicator function of the set A. Choosing our transition kernels as in (3.1) we would then iteratively generate

- u given x from a uniform distribution on the interval $(0, l(x))$;
- x given u from q restricted to the set $A_{u,l} = \{x : l(x) > u\}$.

We will refer to this scheme as the *slice sampler algorithm* for $\pi(\cdot) \propto q(\cdot)l(\cdot)$. In the notation A_u the subscript l is omitted, so that $A_u = A_{u,l}$. In the sequel of this chapter we will mostly concentrate our attention on this simple idea which has proven to be very effective in a large class of Bayesian non-conjugate models [11] and on some variations of this idea. As pointed out in [6], this algorithm is also very appealing in multidimensional problems when $q(x)$ has a simpler structure than $\pi(x)$, for example with independence for the component of x, while $l(x)$ contains the factors that make the interaction among the variables strong.

We stress that the above construction adds a degree of freedom for the researcher. For each given target distribution π, a number of sensible decompositions are often available. The natural question that arises is: "which is the 'best' one?". We do not have an answer but would point out that if $l(x)$ is constant the auxiliary construction produces i.i.d. samples from π which is best in some minimax sense but cannot be globally best because, for certain functionals of π, the "right" negative correlation will reduce the variance of the estimates (antithetic variates, [69]).

Finally notice that, in going from the marginal distribution of x to the conditional distribution of x given u, the interaction term $l(x)$ is replaced by a constraint term $I_{\{u<l(x)\}}(x, u)$. A significant improvement will be reached only if sampling $q(x)$

restricted to the set A_u is more "convenient" than sampling $\pi(x)$ directly. When random observations from q are available, sampling from q restricted to the set A_u can always be done by rejection, but this is a last resort technique. The optimal situation arises when direct sampling from the full conditional distribution of x is possible. This is often the case when $l(x)$ is an invertible function. In such situations Damien at al. [11] show that there is a large class of models for which direct simulation from the full conditional distribution of x is achieved. In Section 5.7 guidelines on how to efficiently sample the full conditional distribution $\pi(x|u)$ are given.

3.3 Variations on the Main Idea

Variations on the general scheme described in the previous section are available. The first possible generalization was introduced in Edwards and Sokal [15]. Suppose

$$\pi(x) \propto q(x) \prod_{l=1}^{L} g_l(x),$$

then we introduce a collection of latent variables $u = (u_1, ..., u_L)$, with each u_l defined on $(0, \infty)$. The latent variables are conditionally independent given x and have joint density with x given by

$$\pi_M(x, u) \propto q(x) \prod_{l=1}^{L} I_{\{u_l < g_l(x)\}}(x, u_l). \tag{3.2}$$

It is easy to verify that the marginal density for x is $\pi(x)$. A Gibbs sampler can now be implemented where the full conditionals for each u_l is the uniform density on $(0,\ g_l(x)),\ l = 1, ..., L$. The full conditional for x is given by q restricted to the set $M_u = \{x : g_l(x) > u_l, l = 1, ..., L\}$. We will refer to this scheme as the *product slice sampler*.

Alternatively, instead of introducing multiple auxiliary variables $u_1, ..., u_L$, we can introduce a single one, say z, and sample $z \mid x$ from a uniform on $(0, \prod_{l=1}^{L} g_l(x))$ and $x \mid z$ from q restricted to the set $P_z = \{x : \prod_{l=1}^{L} g_l(x) > z\}$. This amounts to the following factorization of the distribution of interest

$$\pi_P(x, z) \propto q(x) I_{\{z < \prod_{l=1}^{L} g_l(x)\}}(x, u_1, \cdots u_L). \tag{3.3}$$

We will refer to this scheme as the *multiple slice sampler*. This is nothing but the simple slice sampler where we replace $l(x)$ with $\prod_{l=1}^{L} g_l(x)$, thus all the results we will prove for the slice sampler algorithm also apply to the multiple slice sampler scheme, mutatis mutandis.

Both factorizations in (3.2) and (3.3) have the nice property that the possibly complicated factor $\prod_{l=1}^{L} g_l(x)$ present in the marginal distribution of x, disappears in the joint distribution of x and the auxiliary variables and is replaced by a constraint (or a collection of constraints).

This is another instance where a choice between two possible factorizations, (3.2) and (3.3), is given. Besag and Green [6] show that the equilibrium mean number of attempts before acceptance, when sampling from $\pi(x|u)$ by rejection, is $h \prod_l \sup_x g_l(x)$, where $h = \int q(x) \prod_{l=1}^{L} g_l(x) dx$. Since a product of suprema is greater than, or equal to, the supremum of the product (if the functions are non-negative, as in our case), the equilibrium mean number of attempts is usually reduced by using fewer auxiliary variables. In particular the factorization in (3.3) dominates (3.2) in this sense. Still, it is not at all clear if this is the right comparison to be made: sampling uniforms may be quite cheap compared to performing the comparisons that might be needed in order to implement (3.3).

The use of multiple uniform variables in the above scheme can be viewed as choosing a rectangular shaped region A_u such that its area, given x, is equal to

$G(x) = \prod_{l=1}^{L} g_l(x)$. A possible generalization in this direction consists in allowing A_u to be something other than a rectangle, all that is needed is for its area to have the right volume. The crucial fact, from a practical standpoint, is that the extra cost in drawing random variables uniformly distributed on a complicated region, provides some benefit in terms of making the sampling from the full conditional of x easier.

So far we have designed the conditional distribution of the auxiliary variable so that its normalizing constant cancels with the factor $\prod_{l=1}^{L} g_l(x)$. This eliminates all awkward interactions among the variables of interest. Following [36] and [31], we could instead diminish the strength of such interactions by making the normalizing constant equal to $\prod_{l=1}^{L} g_l^a(x)$. This results in a joint distribution proportional to

$$\pi_a(x,u) \propto q(x) \prod_{l=1}^{L} g_l(x)^{1-a} \prod_{l=1}^{L} I_{\{u<g_l^a(x)\}}(x,u).$$

This method was first introduced by Higdon [36] in the context of the Ising model and was named *partial decoupling*. For guidelines on the choice of the working parameter, a, refer to [36, 57]. Further work along this is contained in [37, 57].

Another possible variation on the product slice sampler scheme consists of allowing a structure of dependence among the uniform random variables. This should be done again in an attempt to simplify sampling from $\pi(x|u)$. In this case the joint distribution of x and u can be specified as

$$\pi(x,u) \propto \pi(x)\pi(u_1|x)\pi(u_2|x,u_1)\cdots\pi(u_L|x,u_1,\cdots u_{L-1}).$$

A special case of the slice sampler is when q only takes values zero and one thereby reducing all the simulation to uniform distributions on various shaped regions. Following the terminology in [74] we will refer to this scheme as to the *uniform slice sampler*. The Gibbs steps to implement the uniform slice sampler are

- sample u given x from a uniform distribution on the interval $(0, \pi(x))$;
- sample x given u from π restricted to the set $A_{u,\pi} = \{x : \pi(x) > u\}$.

The uniform slice sampler can be seen as sampling from the uniform density on the region bounded above by the density of interest and then disregarding the vertical coordinate of the sample points. In more detail, suppose the original distribution is defined over some subset of $\mathcal{R}^n$ and has density proportional to $\pi(x)$. We can sample from this target distribution by sampling from the $n+1$ dimensional region that lies under the curve $\pi(x)$. This idea is formalized by introducing the auxiliary variable u that indeed represents the additional dimension added to the problem for a more efficient solution.

3.4 Examples of Auxiliary Variables

In this section we give a few instances of how latent variables can be exploited to speed up mixing (Swendsen-Wang algorithm, hybrid Monte Carlo, simulated tempering), to simplify calculations (missing values), or to get better output analysis (regeneration times), in Markov chain sampling.

In the *Ising model* with external field [43], or in the *Potts model* [65], there are cases where single site methods move very slowly through the sample space. Swendsen and Wang [81] use binary auxiliary variables that represent links for each neighboring pair of pixels. These variables partition the lattice sites into like colored clusters. When the latent variable is updated, clusters are formed. When the original variable is updated, clusters are colored independently. Variations on this idea are given in [36] and [57].

The *hybrid Monte Carlo algorithm* is another example of how auxiliary variable can be introduced to develop computationally feasible procedures. As before we

indicate with $\pi(x), x \in R^n$, the distribution of interest. If $\pi(x) > 0$ for every x, define the potential energy as $E(x) = -\log \pi(x) - \log z$, for any convenient constant z, so that $\pi(x) \propto \exp(-E(x))$. Define the auxiliary variable $u \in R^n$ by specifying its distribution, $\exp(K(u))$, possibly up to a constant of proportionality. $K(u)$ is called kinetic energy. Assume that x and u are independent so that their joint distribution is $\pi(x,u) \propto \exp(H(x,u))$, where $H(x,u) = E(x) + K(u)$, is the total energy. A Markov chain having $\pi(x,u)$ as its limiting distribution is set up and the marginal realizations of the X-chain are used for inference on relevant functionals of π. In order to set up the Markov chain, Neal [61], suggests to first sample uniformly from values of x and u having the same total energy, $H(x,u)$, and then sample states with different energy value. This is achieved by iteratively updating u with a Gibbs step and (x,u) with a Metropolis step. Given the independence structure imposed on the problem, the full conditional distribution of u coincides with its marginal distribution. In particular, if the kinetic energy is taken to be $K(u) = \sum_{i=1}^{n} \frac{U_i^2}{2m_i}$, with $m_i > 0, \forall i$, then u is updated by drawing u_i i.i.d. $N(0, m_i)$, $i = 1, \cdots, n$. As a result of these updates of u, the value of the total energy changes. When updating (x,u) jointly, the candidate value is proposed by simulating a discretization of the Hamiltonian dynamics of the system. Due to the discretization the value of the total energy is not preserved as we wished. To correct for such a systematic error a Metropolis step is performed and the candidate state is accepted with a probability depending on the change of the total energy function caused by such move. Variations on the hybrid Monte Carlo algorithm described above and a comparison of its performance with other MCMC methods are given in [61].

A typical way of dealing with *missing values* (or future data) in MCMC, is to include them in the state space together with model parameters. Say the data set y is partitioned into y_o, observed data, and y_m, missing data (or future values). We are interested in the posterior distribution of the parameters, $\pi(\theta|y_o) \propto$

$q(\theta|y_o)\int l(y_o, y_m|\theta)dy_m$. It is often easier to treat y_m as latent variables and run a Markov chain sampler on the augmented state space, having $\pi(\theta, y_m|y_o) \propto q(\theta|y_o)$ $l(y_m|\theta, y_o)$ as limiting distribution. Conditional independence structures in the model can be exploited to improve the efficiency of the sampler. Suppose, for example, that actual data and future data are conditionally independent given θ, so that $l(y_m|\theta, y_o) = l(y_m|\theta)$. In such situations MCMC should be used only on the first factor, $q(\theta|y_o)$. Once the algorithm has reached convergence and a plausible sample of θ's is available, direct forward simulation should be performed for the second factor, $l(y_m|\theta)$. Thus, the future observations, y_m, are never used in updating the parameter θ, which seems a reasonable thing to do. This strategy is proven to be more efficient than running a Markov chain on both θ and y_m.

Auxiliary variables can also be exploited to facilitate the analysis of the output of a Markov chain rather than to improve the performance of the sampler. Mykland et al. [58] use latent variables to identify *regeneration times* at which a Markov chain restarts itself. Again the state space is augmented to include a Bernoulli variable, s_n, with probability of success depending on the current and next value of the original variable of interest, x_n. The times at which $s_n = 1$ are regeneration times for the chain $\{x_n, s_n\}$. At each regeneration time the future of the process is independent of the past and identically distributed. This fact can be used to allow variance estimates to be computed based on i.i.d observations. Useful information about the performance of the sampler can also be obtained by analyzing the sequence of regeneration times.

Simulated tempering, [52] and [25], is another instance where a latent variable is introduced to improve mixing. In the general framework discussed in Section 3.2 the auxiliary variable is set up so that the marginal distribution of the variable of interest, x, is preserved. In simulated tempering, instead, the target distribution, $\pi(x)$, coincides with the conditional distribution of x given some specified value of the auxiliary variable (which is usually discrete), say $U = 1$. At the end of the

simulation all sample values (x_t, u_t) for which $u_t \neq 1$ are disregarded. Inference on $\pi(x)$ is based only upon those values x_t for which $u_t = 1$, after an initial burn-in. In other words, in the simulated tempering scheme, instead of specifying the joint distribution of x and u as $\pi(x, u) = \pi(x)\pi(u|x)$ we take $\pi(x, u) = \pi(u)\pi(x|x)$ and typically the target distribution is taken to be $\pi(x) = \pi(x|u = 1)$.

The working parameter introduced in the *partial decoupling* can be treated as the temperature in the Metropolis-coupled MCMC (MCMCMC, [23]) or in the *simulated tempering* algorithm [52, 25]. In the first case we would run a number of chains in parallel, each one having different joint limiting distribution $\pi_a(x, u)$, but all having the target distribution as the marginal limiting distribution for x. At each iteration an attempt is made to swap the states of two of the chains using a Metropolis step. The advantage of this variation over the more standard MCMCMC algorithm is that, at the end of the run, the marginal output from all the chains can be retained and used for inference on $\pi(x)$. The same idea applies in simulated tempering where the multiple samplers having $\pi_a(x, u)$ as limiting distribution, are run in series and randomly interchanged.

3.5 Comparison with the Independence Metropolis-Hastings Algorithm

In this and the following sections we dedicate our efforts to a better understanding of the slice sampler for $\pi(x) \propto q(x)l(x)$ and its variations (3.2) and (3.3). We will start comparing the performance of these schemes with the independence Metropolis-Hastings algorithm. To this aim we will use the Peskun ordering and its generalizations given in Chapter 2.

In order to apply Theorem 2.3.1 we need to find the transition kernel of the slice sampler for $\pi(x) \propto q(x)l(x)$. Since only the marginal X-chain is relevant for inference on π, we are interested in the transition kernel of this sub-chain

$$K(x, B) = \int_B \int_U P(y|u)\, P(u|x)\, du\, dy,$$

leading to

$$K(x, B) = \frac{1}{l(x)} \int_{u=0}^{l(x)} \frac{\pi(B \cap A_u)}{\pi(A_u)}\, du.$$

The next theorem shows that the slice sampler algorithm for $\pi(\cdot) \propto q(\cdot)l(\cdot)$ performs better, in terms of asymptotic variance, than the "corresponding" independence Metropolis-Hastings where $q(\cdot)$ is used as the proposal distribution.

Theorem 3.5.1.

Let $K_S(x, A)$ be the transition kernel for the slice sampler for $\pi(\cdot) \propto q(\cdot)l(\cdot)$ with $\int q(dx) = 1$ and $K_I(x, A)$ be the transition kernel for the independence Metropolis-Hastings algorithm with $q(\cdot)$ as the proposal distribution. Then

$$K_S \succeq K_I. \tag{3.4}$$

Furthermore, on finite state spaces

$$\lambda_{i,K_S} \leq \lambda_{i,K_I}, \ \ \forall i \tag{3.5}$$

where $\lambda_{i,K}$ is the i-th largest eigenvalue of the transition kernel K. On general state spaces

$$\lambda_{max,K_S} \leq \lambda_{max,K_I} \tag{3.6}$$

where $\lambda_{max,K}$ is the supremum of the spectrum of K.

Proof. Consider first the transition kernel for the slice sampler

$$\begin{aligned} K_S(x, A \setminus \{x\}) &= \tfrac{1}{l(x)} \int_{u=0}^{l(x)} \int_{A\setminus\{x\}} q(y \mid A_u)\, dy\, du \\ &= \tfrac{1}{l(x)} \int_{A\setminus\{x\}} \int_{u=0}^{l(x)} q(y \mid A_u)\, du\, dy \\ &= \tfrac{1}{l(x)} \int_{A\setminus\{x\}} q(y) \int_{u=0}^{l(x)} \tfrac{I_{A_u}(y)}{q(A_u)}\, du\, dy \end{aligned}$$

where $q(A_u) = \int_{A_u} q(y)\, dy$. For the independence Metropolis-Hastings algorithm we

have

$$\begin{aligned} K_I(x, A \setminus \{x\}) &= \int_{A\setminus\{x\}} q(y) \min\left[\tfrac{l(y)}{l(x)}, 1\right] \, dy \\ &= \tfrac{1}{l(x)} \int_{A\setminus\{x\}} q(y) \min[l(x), l(y)] \, dy. \\ &= \tfrac{1}{l(x)} \int_{A\setminus\{x\}} q(y) \int_{u=0}^{\min[l(x),l(y)]} \, dy. \end{aligned}$$

It is therefore sufficient to compare the inside integrands to get (3.4):

$$\int_{u=0}^{l(x)} \frac{I_{A_u}(y)}{q(A_u)} \, du = \int_{u=0}^{\min[l(x),l(y)]} \frac{1}{q(A_u)} \, du \geq \int_{u=0}^{\min[l(x),l(y)]} \, du = \min[l(x), l(y)].$$

From Theorem 2.3.3 we have that if two transition matrices are Peskun ordered then the corresponding eigenvalues are also ordered, therefore (3.5) follows. On general state spaces this translates into (3.6), that is, the supremum of the spectrum of K_S is less then or equal to the corresponding supremum for K_I as proved in Theorem 2.3.4. Q.E.D.

A similar result can be proved for the multiple slice sampler that follows from the factorization given in equation (3.3).

Corollary 3.5.1.

Let $P_1(x, A)$ be the transition kernel for the multiple slice sampler and $P_2(x, A)$ be the transition kernel for the independence Metropolis-Hastings algorithm with $q(\cdot)$ as the proposal distribution. We have

$$P_1 \succeq P_2 \tag{3.7}$$

Furthermore on finite state spaces

$$\lambda_{i,P_1} \leq \lambda_{i,P_2}, \ \ \forall i.$$

and on general state spaces

$$\lambda_{max,P_1} \leq \lambda_{max,P2}.$$

Proof. This is a special case of Theorem 3.5.1 with $l(x) = \prod_l g(x)$. Q.E.D.

Theorem 3.5.1 states that, given any independence Metropolis-Hastings algorithm we can construct a corresponding slice sampler that has a smaller asymptotic variance of sample path averages. Namely, if $q(x)$ is the proposal distribution for the independence Metropolis-Hastings sample and $\pi(x)$ is the stationary distribution, let $l(x) = \frac{\pi(x)}{q(x)}$ so that $l(x)q(x)$ provides a factorization for the distribution of interest. The problem with the slice sampler is that it might be expensive to sample from the conditional distribution of $x|u$ as pointed out in Section 3.2. We will see in the sequel that if $l(x)$ is a bounded function both the independence Metropolis-Hastings algorithm and the corresponding slice sampler are uniformly ergodic. If $l(x)$ is not bounded then the independence Metropolis-Hastings algorithm is not geometrically ergodic, but there are conditions under which the slice sampler can still be geometric ergodic (see Section 3.9 and [71]).

When using an independence Metropolis-Hasting sampler it is a good idea to look for a proposal distribution that is heavier-tailed than the target distribution in order to avoid long periods stuck in the tails. This is also a valid suggestion if we use a slice sampler and sample the conditional distribution of x given u by rejection. To

see this suppose $q(\cdot)$ is lighter-tailed than $\pi(\cdot)$ and that X_t is currently in the tails of $\pi(\cdot)$. Then $l(X_t)$ will be large and the set $A_{u_t} = \{x : l(x) > u_t\}$ will likely be small. This gives a high rejection rate when using q as the proposal distribution to sample by rejection from q restricted on A_u.

3.6 Irreducibility, Aperiodicity and Detailed Balance

It is useful to note that if $\pi(\cdot) \propto q(\cdot)l(\cdot)$ then the slice sampler is well defined, π-irreducible and aperiodic. π-irreducibility follows from the fact that $\pi(\cdot)$ is absolutely continuous with respect to $q(\cdot)$ [83]. Furthermore recall that a π-irreducible Gibbs sampler is aperiodic [26], therefore the slice sampler, which is nothing but a Gibbs algorithm defined on an augmented state space, is aperiodic.

If $l(x) > 0, \quad \forall x \in \mathcal{X}$ then the transition kernel for the slice sampler is absolutely continuous with respect to $\pi(\cdot)$. It follows that the slice sampler is Harris recurrent [83]. If, on the other hand, $l(\cdot)$ is not strictly positive, let $\mathcal{T} = \{x : l(x) > 0\}$, then on $\mathcal{X} \setminus \mathcal{T}$ the slice sampler is Harris recurrent.

The detailed balance condition (2.7)

$$\iint f(y)g(x)\pi(dx)K_S(x,dy) = \iint f(x)g(y)\pi(dx)K_S(x,dy). \tag{3.8}$$

for all bounded f and g, holds for the slice sampler. In fact, the only way (2.7) can hold for all bounded functions f and g is that

$$\pi(dx)K_S(x,dy) \quad = q(dx)l(x)\tfrac{q(dy)}{l(x)} \int_{u=0}^{\min[l(x),l(y)]} \tfrac{1}{q(A_u)}\, du =$$

$$= \pi(dy)K_S(y,dx) \quad = q(dy)l(y)\tfrac{q(dx)}{l(y)} \int_{u=0}^{\min[l(y),l(x)]} \tfrac{1}{q(A_u)}\, du.$$

This allows us to rewrite the transition kernel as a symmetric linear operator on the Hilbert space $L^2(\pi)$

$$\begin{aligned}\tilde{K}_S(x,y) &= \frac{K_S(x,y)}{\pi(y)} \\ &= \frac{1}{l(x)l(y)} \int_{u=0}^{\min[l(y),l(x)]} \frac{1}{q(A_u)}\, du \\ &= \tilde{K}_S(y,x).\end{aligned}$$

3.7 Positive Operators

An operator P is called *positive* if for any f in the space where P is defined,

$$(f, Pf) \geq 0.$$

Suppose P is the operator on $L_0^2(\pi)$ of a reversible Markov chain having π as its stationary distribution. Then P is positive if and only if all the lag one autocovariances are non-negative.

A stationary Markov chain $\{X_n\}_{n=1}^{\infty}$ is said to have the *interleaving Markov property* if there exists a conjugate Markov chain $\{Y_n\}_{n=1}^{\infty}$ such that

1. X_k and X_{k+1} are conditionally independent given Y_k, $\forall k$;
2. Y_k and Y_{k+1} are conditionally independent given X_k, $\forall k$;
3. (Y_{k-1}, X_k), (Y_k, X_k) and (Y_k, X_{k+1}) are identically distributed.

As pointed out in [48] the marginal chains in a two-state Gibbs sampler have the interleaving Markov property. It follows that the slice sampler with a single auxiliary variable has the same property since it is nothing but a two-state Gibbs sampler. But even if we consider a slice sampler with more than one auxiliary variable, say $U_1, \cdots, U_l$, again the interleaving Markov property holds for the two marginal chains $\{X_n\}_{n=1}^{\infty}$ and $\{Y_n\}_{n=1}^{\infty}$, where X is the variable of interest and $Y = (U_1, \cdots, U_l)$ is a vector valued random variable whose components are the single auxiliary variables. It is obvious why (1) and (2) are satisfied. As for (3), it is true because the full-conditional updates of X and Y preserve the joint stationary distribution. To be more precise, in the product slice sampler we implicitly define a joint distribution on the enlarged state space, which is given in (3.2). We then construct our sampler by updating X and Y to preserve that joint distribution, π_M, which thus becomes our stationary distribution. So, under stationarity the distribution of (Y_{k-1}, X_k) is π_M. Since the update of Y given X preserves this joint distribution, it follows that (Y_k, X_k) also has distribution π_M. The update of X given Y also preserves π_M, thus also (Y_k, X_{k+1}) has distribution π_M. Similar reasoning holds for the multiple slice sampler with, of course, a different joint distribution.

The interleaving Markov property implies that the lag one autocovariances of the marginal chains are non-negative [48]. It follows that the operators of the marginal Markov chains are non-negative. As pointed out by Liu [48], the interleaving Markov property as defined here implies reversibility, furthermore, reversibility implies that the even-lag autocovariances are non-negative [47]. By using induction we obtain that the n-lag autocovariances of the marginal X and Y chains are non-negative monotone decreasing with n [47].

The above observations can be summarized in the following theorem that appears in Liu's doctoral dissertation ([48], p. 20).

Theorem 3.7.1.

Suppose P is the operator of a general Markov chain $\{X_n\}_{n=1}^{\infty}$. A necessary and sufficient condition for $\mathrm{Cov}[f(X_0), f(X_n)]$ *to be non-negative and monotone decreasing with n for all $f \in L_0^2(\pi)$ is that P is a positive and self-adjoint operator.*

But we can prove an even stronger result using the representation given in [24] of the lag k autocovariance $\gamma_k = \mathrm{Cov}_\pi[f(X_i), f(X_{i+k})]$ in terms of the spectral measure (Section 2.3.2)

$$\gamma_k = \int \lambda^k E_{f,P}(d\lambda), \qquad \forall k.$$

For a positive operator P the spectrum, $\sigma(P)$, is a subset of $[0, 1]$. Unless P produces independent samples we have that $\sigma(P) \neq \{0\}$. Thus, the lag k autocovariances are strictly positive. Furthermore, γ_k strictly decreases as k increases since, on the spectrum, for fixed λ, λ^k is a decreasing function in k. Finally

$$\gamma_{k-1} - \gamma_k \geq \gamma_k - \gamma_{k+1}$$
$$\gamma_k \leq \frac{1}{2}(\gamma_{k-1} + \gamma_{k+1})$$

is implied by $\lambda^k \leq \frac{1}{2}(\lambda^{k-1} + \lambda^{k+1})$, which is implied by $\lambda^2 = 2\lambda + 1 \geq 0$, which is implied by $(\lambda - 1)^2 \geq 0$. Thus, γ_k is strictly convex in k if the Markov chain is irreducible (the reasoning to get these results follows [24]). These properties of the autocovariances of a chain generated using a self-adjoint positive operator can be used to construct adaptive window estimators for the variance in the central limit theorem along the line of what has been proposed in [24].

In the special case of the two-variable Gibbs sampler and of the slice sampler, the interleaving Markov property of the marginal chains leads to positive marginal operators. Whether the interleaving Markov property is a necessary condition for the operator to be positive is not clear. Other examples of positive operators are the

random scan Gibbs sampler, and the independence Metropolis-Hastings algorithm [47, 49]. Since both transition kernels K_S (simple slice sampler) and K_I (corresponding independence Metropolis-Hastings algorithm) give rise to positive operators, this implies that their spectra are concentrated on $[0, 1]$. We have shown that K_S dominates K_I in the Peskun sense (Section 3.5). On a finite state space this implies that all the eigenvalues of K_S are less then or equal to the corresponding eigenvalues of K_I (Section 2.3.1). On a general state space this implies that the supremum of the spectrum of K_S is less than or equal to the supremum of the spectrum of K_I (Section 2.3.2). It follows that, on finite state spaces, the simple slice sampler has better performance than the corresponding independence Metropolis-Hastings algorithm both in terms of asymptotic variance or MCMC estimate (since it has smaller eigenvalues) and in terms of rate of convergence to stationarity in total variation distance (since it has smaller eigenvalues in absolute value). The same conclusion can be reached for general state spaces by comparing the supremum of the spectrum and of the absolute spectrum for the two operators (Theorem 2.3.4).

3.8 A Counterexample

The next example will show that a theorem similar to 3.5.1 does not hold for the product slice sampler. Let $K_3(x, A)$ be the transition kernel for the product slice sampler as defined in (3.2) and $K_2(x, A)$ be the transition kernel for the independence Metropolis-Hastings algorithm with $q(\cdot)$ as the proposal distribution. We will show that there are situation in which

$$K_3(x, A) \succeq K_2(x, A) \tag{3.9}$$

as expected, but there are also cases where the two kernels are not comparable in terms of the Peskun ordering. For a positive random variable X, consider the following target distribution π and corresponding factorization $\pi(x) = q(x)g_1(x)g_2(x)$:

$$\begin{aligned}
\pi(x) &= e^{-x} \\
q(x) &= (1 - q_1 - q_2)e^{-x(1-q_1-q_2)} \\
g_1(x) &= \tfrac{1}{1-q_1-q_2}e^{-xq_1} \\
g_2(x) &= e^{-xq_2}.
\end{aligned}$$

Take $q_1 > 0$ and $q_2 < 0$. Let $k = \frac{1}{1-q_1-q_2} > 0$, which implies $q_1 + q_2 < 1$. With this choice for the parameters we have that g_1 is a decreasing function with $g_1 < k$, while g_2 is increasing and $g_2 > 1$ for $x > 0$. The transition kernel for the product slice

sampler satisfies

$$K_3(x, A \setminus \{x\}) = \frac{1}{g_1(x)g_2(x)} \int_{A\setminus\{x\}} q(y) \int_{u_1=0}^{\min[g_1(x),g_1(y)]} \int_{u_2=0}^{\min[g_2(x),g_2(y)]} \frac{1}{q(M_u)}\, d\,u_2\, d\,u_1$$

where $M_u = \{\, x : g_1(x) > u_1 \text{ and } g_2(x) > u_2 \,\}$. The transition kernel for the independence Metropolis-Hastings algorithm is

$$K_2(x, A \setminus \{x\}) = \frac{1}{g_1(x)g_2(x)} \int_{A\setminus\{x\}} q(y)\ \min[g_1(x)g_2(x), g_1(y)g_2(y)]$$

We can therefore limit ourself to the comparison of

$$\begin{aligned} K_3^*(x, y) &= \int_{u_1=0}^{\min[g_1(x),g_1(y)]} \int_{u_2=0}^{\min[g_2(x),g_2(y)]} \frac{1}{q(M_u)}\, d\,u_2\, d\,u_1 \\ &\geq \min[g_1(x), g_1(y)]\ \min[g_2(x), g_2(y)] \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} K_2^*(x, y) &= \min[g_1(x)g_2(x), g_1(y)g_2(y)] \\ &= \min[ke^{-x(q_1+q_2)}, ke^{-y(q_1+q_2)}] \end{aligned} \tag{3.11}$$

for $x \neq y$. The way in which (3.10) and (3.11) are ordered depends on $(q_1 + q_2)$ being positive or negative. Consider first the case $(q_1 + q_2) \leq 0$ so that $k \leq 1$. Then

$$K_2^* = ke^{-(x\wedge y)(q_1+q_2)} \tag{3.12}$$

and

$$K_3^* \geq k e^{-(x\vee y)q_1 - (x\wedge y)q_2} \tag{3.13}$$

Thus

$$\frac{K_3^*}{K_2^*} \geq e^{(x\wedge y)2q_1} \geq 0$$

therefore (3.9) holds. Consider now the case when $(q_1 + q_2) > 0$. We want to find a pair (x, y) such that $K_3^*(x, y) < K_2^*(x, y)$. Without loss of generality assume $y = 0$ since, by continuity of $K_2^*(x, y)$ and $K_3^*(x, y)$ we know that, if $(x, 0)$ is a pair such that $K_3^*(x, 0) < K_2^*(x, 0)$, then also (x, ϵ) will preserve the ordering for small values of ϵ. Setting $y = 0$ gives

$$K_3^*(x, 0) = \int_0^{g_1(x)\vee k} du_1 \int_0^{g_2(x)\vee 1} \frac{1}{q(M_u)} du_2$$

and

$$K_2^*(x, 0) = [g_1(x)g_2(x)] \vee k.$$

Next, write K_3^* and K_2^* in terms of $g = g_1(x)$ so we can examine the behavior of this function when $x \to \infty$. Notice that $g_2(x) = (\frac{g}{k})^{\frac{q_2}{q_1}}$ and for small positive values of g (that is for large values of x) we have

$$g_1(x) \vee k = g_1(x) = g$$

$$g_2(x) \vee 1 = 1$$

$$g_1(x) * g_2(x) \vee k = g_1(x) * g_2(x) = g(\frac{g}{k})^{\frac{q_2}{q_1}}$$

Thus (again for small g)

$$K_2^*(g) = g * \left(\frac{g}{k}\right)^{\frac{q_2}{q_1}} = \left(\frac{1}{k}\right)^{\frac{q_2}{q_1}} g^{\frac{q-1+q_2}{q_1}}$$

and

$$K_3^*(g) = \int_0^g du_1 \int_0^1 du_2 \frac{1}{q(M_u)}.$$

From these expressions it is clear that $K_3^*(g=0) = K_2^*(g=0) = 0$. We now compute the first derivative with respect to g to see what happens for small values of g.

$$\begin{aligned} K_2^{*'}(0) &= \lim_{g\to 0+} \left(\frac{1}{k}\right)^{\frac{q_2}{q_1}} \frac{q-1+q_2}{q_1} g^{\frac{q-1+q_2}{q_1}-1} \\ &= \lim_{g\to 0+} \left(\frac{1}{k}\right)^{\frac{q_2}{q_1}} \frac{q-1+q_2}{q_1} g^{\frac{q_2}{q_1}} \\ &= +\infty \end{aligned} \tag{3.14}$$

since $\frac{q2}{q1} < 0$ and $\left(\frac{1}{k}\right)^{\frac{q_2}{q_1}} \frac{q1+q2}{q1} > 0$. For K_3^* we have

$$\begin{aligned} K_3^{*'}(0) &= \int_0^1 \frac{1}{q(M_{(0,u_2)})} du_2 \\ &= \int_0^1 u_2^{-1/(k*q2)} du_2 \\ &= \frac{-1}{kq_2} < \infty. \end{aligned} \tag{3.15}$$

Thus, in a neighborhood of $g = 0$ we have $K_3^* < K_2^*$ Hence for large enough x we have $K_3^*(x,0) < K_2^*(x,0)$ for suitable values of q_1 and q_2.

3.9 Uniform and Geometric Ergodicity

In this section we discuss the rate of convergence to stationarity in total variation norm of the slice sampler for $\pi(\cdot) \propto q(\cdot)l(\cdot)$ and of its variations (3.2) and (3.3). We will prove that, under mild regularity conditions, these Markovian schemes are uniformly ergodic. An example will show that if these conditions are not met, geometric ergodicity can sometimes still be achieved.

Let us recall some definitions and facts that hold for a π-irreducible Markov chain defined on a state space $(E, \mathcal{E})$, [83, 54]. Following Nummelin [62], the *total variation distance* of a bounded signed measure λ on $(E, \mathcal{E})$ is defined as

$$||\lambda|| = \sup_{A \in \mathcal{E}} \lambda(A) - \inf_{A \in \mathcal{E}} \lambda(A).$$

A Markov chain is *ergodic* if it is positive recurrent and aperiodic. An ergodic Markov chain with invariant distribution π is *geometrically ergodic* if there exists a nonnegative extended real-valued function V with $\pi|V| < \infty$ and a positive constant $r < 1$ such that

$$||P^n(x, \cdot) - \pi(\cdot)|| \leq V(x)r^n \tag{3.16}$$

for all x. The chain is *uniformly ergodic* if there is a positive constant V and a positive constant $r < 1$ such that

$$\sup_{x \in \mathcal{E}} -\pi(\cdot)||P^n(x, \cdot)|| \leq Vr^n. \tag{3.17}$$

Uniform ergodicity implies geometric ergodicity. The constant r is called rate of convergence and equals the spectral radius, $\Lambda_{max,P}$ (Chapter 3, p.48 [28]).

Definition 3.9.1.

A set $C \in \mathcal{E}$ is a small set for the Markov chain, if for some $\delta > 0$ and $n > 0$ and some probability measure ν,

$$P^n(x, \cdot) \geq \delta\nu(\cdot) \qquad x \in C \tag{3.18}$$

where $P^n(x, \cdot)$ is the n-step transition probability for the Markov chain.

For uniform ergodicity a simple necessary and sufficient condition is available (Nummelin, [62], Theorem 6.15 and Tierney, [83], Proposition 2).

Theorem 3.9.1.

A transition kernel P is uniformly ergodic if and only if the state space E is a small set. Furthermore, if P satisfies condition (3.18) *then the convergence rate r satisfies $r^n \leq (1 - \delta)$.*

Definition 3.9.2.

A set $C \in \mathcal{E}$ is a proper atom for the Markov chain, if $\pi(C) > 0$ and $P(x, \cdot) = P(y, \cdot)$ for all $x, y \in C$.

Clearly a proper atom is a small set. For any set $A \in \mathcal{E}$, let τ_A be the first return time to A. The next theorem appears in [54].

Theorem 3.9.2.

A necessary and sufficient condition for geometric ergodicity is that for some small set C with $\pi(C) > 0$, there exists $K > 1$ such that

$$\sup_{x \in C} E_x[K^{\tau_C}] < \infty. \tag{3.19}$$

Theorem 3.9.3.

The slice sampler for $\pi(\cdot) \propto q(\cdot)l(\cdot)$ is uniformly ergodic if $l(\cdot)$ is a bounded function and the rate of convergence to stationarity in total variation distance, r, is such that

$$r \leq \{1 - h[\sup_{x \in X} l(x)]^{-1}\} \tag{3.20}$$

where $h = \int_X q(x)l(x)\,dx$.

Proof. We will show that

$$K_1(x, \cdot) \geq \pi(\cdot)h \left[\sup_{x \in X} l(x)\right]^{-1}$$

i.e. the entire space X is a small set for K_1. It follows that K_1 is uniformly ergodic and (3.20) holds.

$$\begin{aligned} K_1(x, \cdot) &= \frac{q(\cdot)}{l(x)} \int_{u=0}^{\min[l(x), l(\cdot)]} \frac{1}{q(A_u)}\,du \\ &\geq \frac{q(\cdot)}{l(x)} \min[l(x), l(\cdot)] \\ &\geq \frac{h\pi(\cdot)}{\sup_{x \in X} l(x)}. \end{aligned}$$

Q.E.D.

Corollary 3.9.1.

The multiple slice sampler (3.3) *is uniformly ergodic if $\prod_l g_l(x)$ is bounded and the rate of convergence to stationarity in total variation distance, r, is such that*

$$r \leq \{1 - h[\sup_{x \in X} \prod_l g_l(x)]^{-1}\} \tag{3.21}$$

where $h = \int_X q(x) \prod_l g_l(x)\, dx$.

Proof. It follows from Theorem 3.9.3 by setting $l(x)$ equal to $\prod_l g_l(x)$. Q.E.D.

Theorem 3.9.4.

The product slice sampler (3.2) *is uniformly ergodic if* $g_l(x)$ *are bounded functions and the rate of convergence* r *is such that*

$$r \leq \{1 - h[\prod_l \sup_{x \in X} g_l(x)]^{-1}\} \tag{3.22}$$

Proof. We will show that the transition kernel for the product slice sampler, $K_M(x, \cdot)$ is such that

$$K_M(x, \cdot) \geq \pi(\cdot) h[\prod_l \sup_{x \in X} g_l(x)]^{-1}$$

i.e. the entire space X is a small set for K_M. It follows that K_M is uniformly ergodic

and (3.22) holds. Let $B_x = \{i : l_l(x) > l_i(\cdot)\}$.

$$\begin{aligned}
K_M(x,\cdot) &= \frac{q(\cdot)}{\prod_l g_l(x)} \int_{u_1=0}^{\min[g_1(x),g_1(\cdot)]} \cdots \int_{u_L=0}^{\min[g_L(x),g_L(\cdot)]} \frac{1}{q(M_u)}\, du_L \cdots du_1 \\
&\geq \frac{q(\cdot)}{\prod_l g_l(x)} \prod_l \min[g_l(x), g_l(\cdot)] \\
&= \frac{q(\cdot)}{\prod_l g_l(x)} \prod_{l\in B_x} g_l(\cdot) \prod_{l\in B_x^C} g_l(x) \\
&= \frac{q(\cdot)}{\prod_{l\in B_x} g_l(x)} \prod_{l\in B_x} g_l(\cdot) \frac{\prod_{l\in B_x^C} g_l(x)}{\prod_{l\in B_x^C} g_l(x)} \\
&= \frac{\pi(\cdot)h}{\prod_{l\in B_x} g_l(x) \prod_{l\in B_x^C} g_l(\cdot)} \\
&\geq \frac{\pi(\cdot)h}{\prod_l \sup_{y\in X} g_l(y)}.
\end{aligned}$$

Q.E.D.

A result similar to Theorem 3.9.3 holds for the independence Metropolis-Hastings algorithm [54]: if the weight function $w(x) = \frac{\pi(x)}{q(x)}$ is bounded then the independence Metropolis-Hastings algorithm with proposal $q(\cdot)$ is uniformly ergodic. On the other hand if the $\operatorname{ess\,inf}[\frac{q(x)}{\pi(x)}] = 0$, then the algorithm is not even geometrically ergodic.

This is not true for the slice sampler algorithm: even if $l(\cdot)$ is not bounded the algorithm can still be geometrically ergodic as the exponential example in Chapter 4 shows. A detailed discussion on geometric convergence of the slice sampler and of the product slice sampler is given in [71].

3.10 Related Work

There are a few recent papers where theoretical convergence properties and possible applications of the various slice samplers are studied. In this section a brief review of the results contained in these papers is given.

3.10.1 Theoretical Properties of Slice Samplers

Fishman in [17] focuses on discrete state spaces and gives a useful eigenvalue analysis of the slice sampler. This study can be exploited to choose among different available factorizations of the target distributions in order to select the best one in terms of minimizing the asymptotic variance of MCMC estimates or in terms of obtaining better rate of convergence to stationarity.

Roberts and Rosenthal in [74] give interesting results on convergence rates of the slice sampler in general state spaces. Using Foster-Lyapunov drift condition techniques [62, 55] the authors provide conditions under which the one dimensional slice sampler is geometrically ergodic: Theorem 7 (for bounded target distribution) and Theorem 8 (for unbounded target distributions). Furthermore in Theorem 10, Roberts and Rosenthal give useful rigorous quantitative bounds on the total variation distance to stationarity of an arbitrary one dimensional slice sampler satisfying the condition that $uQ'(u)$ is non-increasing, where $Q'(u)$ is the derivative of the function defined in (4.3). They also provide a rather general statement that the uniform slice sampler, on any target distribution satisfying a condition similar to log-concavity, converges to stationarity in less than 530 iterations (Theorem 12).

In the last part of their paper Roberts and Rosenthal analyze convergence properties of the product slice sampler. They focus on two special cases of (3.2). The first is the case when the functions $g_l(x)$ are all decreasing in the same direction (Theorem 14). They then study the *opposite slice sampler* which is a one dimensional

product slice sampler as in (3.2) with $L = 2$ and and g_1 and g_2 decreasing in opposite directions (Theorem 16).

Finally in [74] it is noted that the Markov chain generated via the uniform slice sampler is *stochastically monotone* with respect to the following partial ordering defined on the $\mathcal{X}$ portion of the state space:

$$x \prec x' \text{ if and only if } \pi(x) \leq \pi(x'). \tag{3.23}$$

This means that, for all fixed $z \in \mathcal{X}$, we have:

$$P(X_1 \prec z | X_0 = x) \geq P(X_1 \prec z | X_0 = x') \text{ whenever } x \prec x'.$$

This ordering can be naturally extended to an ordering on the enlarged state space by simply disregarding the value of u.

Neal [59, 61] discusses both theoretical issues and applications of various slice samplers. In [59] he proposes different clever approaches to the conditional sampling of the variable of interest given the auxiliary variable in a more efficient way than by rejection in cases where these conditional distributions cannot be sampled by standard techniques. An over-relaxed version of the slice sampler is also proposed which sometimes greatly improves sampling efficiency by suppressing random walk behavior. In [61] hybrid Monte Carlo methods that take advantage of the use of latent variables are discussed. A brief review of these techniques is given in Section 3.4.

3.10.2 Applications of the Slice Sampler

Damien et al. in [11] give a variety of examples of successful uses of slice samplers within a Bayesian framework. They analyze Bayesian non-conjugate models (Poisson/lognormal models, Bernoulli/logistic regression, Probit models, Weibull proportional hazards models); generalized linear mixed models (random effects binomial models and Poisson models) and nonlinear mixed models (pharmacokinetic models and logistic models).

In all their examples the authors choose a factorization of the target distribution such that the conditional sampling of the variable of interest given the auxiliary variable(s) is easily implemented by standard techniques. The resulting effect is to set up a Gibbs sampler having a set of full conditionals that can be sampled directly thus avoiding rejection based methods, adaptive algorithms or other sampling approaches which can be difficult to implement since they require careful tuning to achieve satisfactory performance.

Hurn [37] gives a careful discussion on the use of auxiliary variables in image reconstruction problems. Partial decoupling and multi-grid implementations of slice sampler techniques are discussed.

3.11 Conclusions

One of the major problems with the use of MCMC methods is the assessment of convergence. Theoretical results on rates of convergence are currently available only for a very limited number of applications. We state an easy to verify, sufficient condition for the uniformly ergodicity of the slice sampler and of the multiple slice sampler. Furthermore we provide an upper bound for the rate of convergence to stationarity for both samplers.

As a final remark recall that is often a good idea, when combining transition kernels via mixtures or cycles, to include an independence Metropolis-Hastings kernel with bounded weight function. This makes the whole mixture or cycle uniform ergodic [83]. We would suggest instead to include the "corresponding" slice sampler kernel: again the resulting mixture or cycle will gain uniform ergodicity (since the whole space is proven to be a small set) with the advantage of having a smaller asymptotic variance of sample path averages.

Chapter 4

Examples

In this chapter various slice samplers will be analyzed in detail. The target distributions considered are the exponential, the Cauchy and the witches hat distribution [53].

4.1 The Exponential Case

We analyze in detail the convergence properties of the slice sampler when the target distribution is exponential. For $q > 0$ and $x > 0$ let

$$\pi(x) = e^{-x}$$
$$q(x) = qe^{-qx}$$
$$l(x) = \frac{1}{q}e^{-x(1-q)}$$

The slice sampler for this example works as follows: iteratively sample u from a uniform distribution on the interval $[0, l(x)]$ and x from its full conditional distribution which is proportional to $q\, e^{-qx}\, I_{\{l(x)>u\}}(x, u)$.

If $q > 1$ then $l(x)$ is unbounded. Nevertheless, if $1 < q < 2$ the marginal chain of the auxiliary variable U is geometrically ergodic. This is enough to ensure that, for these values of the parameter q, the whole Markov chain is also geometrically ergodic.

To prove geometric ergodicity of the U-chain in the next two lemmas we will show that

1. the set $S = \{u : 0 < u < \frac{1}{q}\}$ is a proper atom for all $q > 1$;

2. condition (3.9.2) holds on the set S for all $1 < q < 2$.

Lemma 4.1.1.

The set S is a proper atom and therefore a small set for the U-chain.

Proof. For all $u \in S$ we have $q(A_u) = 1$, thus

$$\begin{aligned} P(u,v) &= \int_X f(x \mid u) f(v \mid x)\, dx \\ &= \int_X \tfrac{q(x)}{l(x)} I_{(0,l(x))}(v)\, dx \\ &= \int_X q^2 e^{x(1-2q)} I_{(\frac{log(qv)}{q-1},\infty)}(x)\, dx \\ &= \begin{cases} \frac{q^2}{2q-1}(qv)^{\frac{1-2q}{q-1}} & v \geq \frac{1}{q} \\ \frac{q^2}{2q-1} & v < \frac{1}{q} \end{cases} \end{aligned}$$

therefore the one step transition probability $P(u,v)$ does not depend on the starting point u as long as $u \in S$ and we can conclude that S is a proper atom for the U-chain. Q.E.D.

Lemma 4.1.2.

There exists $K > 1$ such that $\sup_{u \in S} E_u[K^{\tau_S}] < \infty$.

Proof. Consider the one step transition probability, $P(u,v)$, for $u \in S^C$

$$P(u,v) = \begin{cases} \frac{q}{2q-1}\frac{1}{u}\left(\frac{v}{u}\right)^{\frac{1-2q}{q-1}} & v \geq u \\ \frac{q}{2q-1}\frac{1}{u} & 0 < v < u \end{cases} \tag{4.1}$$

Perform the change of variable $x = \frac{v}{u}$ and derive, from $P(u, v)$, the density of x conditional on the value of u

$$f(x) = \begin{cases} \frac{q}{2q-1} x^{\frac{1-2q}{q-1}} & x \geq 1 \\ \\ \frac{q}{2q-1} & 0 < x < 1 \end{cases} \tag{4.2}$$

which is independent of u. We can therefore rewrite the U-chain as follows

$$U_{n+1} = \max(U_n, \frac{1}{q}) X_n$$

where $X_1, X_2 \cdots$ are i.i.d. random variables with density given in (4.2). Notice that $E(X) = \frac{2q^2-q}{2(2q-1)}$.

Let $\tau = \tau_S$ and $K = \frac{1}{E(X)}$. By recurrence of the Markov chain we have $P_u(\tau < \infty) = 1$. Let $W_n = \prod_{i=1}^{n} X_i$ and $V_n = K^n W_n$. Being the product of i.i.d, mean one, non negative random variables, V_n is a non negative martingale. Thus, for each n

$$1 = E[K^{\tau \wedge n} W_{\tau \wedge n}] \geq E[K^{\tau \wedge n} W_\tau].$$

Since τ is almost surely finite, by monotone convergence

$$1 \geq E[K^\tau W_\tau] = E[K^\tau E[W_\tau | \tau]].$$

To be precise, we need $K > 1$ to be able to apply the monotone convergence theorem while, if $K \leq 1$, we should instead use the dominated convergence theorem. But, as we will see, $K \geq 1$ is the only case we are really interested in.

Consider now

$$\begin{aligned} E[W_\tau|\tau = n] &= E[W_n|\tau = n] \\ &= E[E[W_n|\tau = n, W_{n-1}]|\tau = n] \\ &= E[E[W_n|W_n \leq \frac{1}{q}, W_{n-1}]|\tau = n] \\ &= E[E[X_n W_{n-1}|X_n W_{n-1} \leq \frac{1}{q}, W_{n-1}]|\tau = n]. \end{aligned}$$

We will study

$$E[Xw|Xw \leq \frac{1}{q}] = wE[X|X \leq \frac{1}{qw}]$$

as a function of $w \in (\frac{1}{q}, \infty)$

$$\begin{aligned} wE[X|X \leq \frac{1}{qw}] &= w \int_0^{\frac{1}{qw}} x \frac{q}{2q-1} \frac{1}{P(x < \frac{1}{qw})}\, dx \\ &= w\frac{q}{2q-1} w(2q-1) \int_0^{\frac{1}{qw}} x\, dx \\ &= \frac{1}{2q} > 0. \end{aligned}$$

It follows that

$$E[K^\tau] \leq 2q < \infty.$$

Q.E.D.

Condition (3.9.2) is hence verified as long as $K = \frac{1}{E(X)} > 1$ and this is true for $\frac{1}{2} < q < 2$. We can thus conclude that for $1 < q < 2$ the U-chain, and therefore also the X-chain, is geometrically ergodic.

We stress that this example does not fit in any of the theorems in [71] where

various conditions for geometric ergodicity of the slice sampler are given. In order to see this we first need to make a change of variable and transform $q(x)$ to the indicator function of a (possibly infinite) subset of the state space $\mathcal{X}$. Performing the change of variable $Z = T(X) = 1 - e^{-qX}$, we get

$$\begin{aligned} \pi_Z(z) &= \frac{1}{q}(1-z)^{\frac{1-q}{q}} I_{(0,1)}(z) \\ q_Z(z) &= I_{(0,1)}(z) \\ l_Z(z) &= \pi_Z(z). \end{aligned}$$

Following the notation in [71] we have

$$L(u) = \{z : \pi_Z(z) > u\} = \{z : z > 1 - (qu)^{\frac{q}{1-q}}\}$$

and the Lebesgue measure of this set is given by

$$Q(u) = (qu)^{\frac{q}{1-q}} \tag{4.3}$$

Since $\pi_Z(z)$ in unbounded, Theorem 7 in [71] does not apply. Furthermore, there is no constant $\alpha > 1$ such that the function

$$Q'(u)u^{1+\frac{1}{\alpha}} = \frac{q^{\frac{1}{1-q}}}{1-q} \; u^{\frac{q}{1-q}+\frac{1}{\alpha}}$$

is non-increasing in u (when $q > 1$ as in the case we are considering) therefore Theorem 8 in [71] does not apply either. Thus, geometric ergodicity for the slice sampler under study cannot be derived from any of the theorems given in [71] but must be derived directly by applying one of the equivalent conditions to geometric ergodicity as we did.

4.2 The Cauchy Case

Consider the Cauchy target distribution

$$\pi(x) = \frac{1}{\pi(1+x^2)}.$$

In order to sample from $\pi(x)$ we will use the uniform slice sampler

$u|x \sim$ Uniform on the interval $[0, \pi(\mathbf{x})]$

and

$x|u \sim$ Uniform on the set $L(u) = \{\ x : \pi(x) > u\ \}$.

Following the notation in [71] we have

$$\begin{aligned} Q(u) &= m^1[L(u)] \\ &= m^1[\{\ x : -\sqrt{\frac{1}{\pi u} - 1} < x < \sqrt{\frac{1}{\pi u} - 1}\}] \\ &= 2\sqrt{\frac{1}{\pi u} - 1} \end{aligned}$$

where m^d is the d-dimensional Lebesgue measure. Notice that $u < \frac{1}{\pi}$ so that $\frac{1}{\pi u} - 1 > 0$.

Since $\pi(x)$ is a bounded function, Theorem 7 in [71] applies: if there exists a constant $\alpha > 1$ such that $Q'(u)u^{1+\frac{1}{\alpha}}$ is non-increasing at least in an open set containing zero, then this uniform slice sampler is geometrically ergodic. We have

$$Q'(u)u^{1+\frac{1}{\alpha}} = \left[\left(\frac{1}{\pi u} - 1\right)^{-\frac{1}{2}} \left(\frac{-1}{\pi}\right) u^{\frac{1}{\alpha}-1}\right] \tag{4.4}$$

and

$$\frac{d}{d\,u}\left[Q'(u)u^{1+\frac{1}{\alpha}}\right] = \frac{-1}{\pi}\left(\frac{1}{\pi u} - 1\right)^{-\frac{1}{2}} u^{\frac{1}{\alpha}-2}\left[\frac{1}{2\pi}\left(\frac{1}{\pi u} - 1\right)^{-1} u^{-1} + \left(\frac{1}{\alpha} - 1\right)\right].$$

The mode of our target distribution is $\frac{1}{\pi}$ thus u is always less than $\frac{1}{\pi}$. Therefore, the function (4.4) is non-increasing if

$$\frac{1}{2\pi}\left(\frac{1}{\pi u}-1\right)^{-1}u^{-1}+\left(\frac{1}{\alpha}-1\right)\geq 0$$

or, equivalently, if

$$\alpha \leq \left[1-\frac{1}{2\pi}\left(\frac{1}{\pi u}-1\right)^{-1}u^{-1}\right]^{-1}=\frac{2(1-\pi u)}{1-2\pi u}. \tag{4.5}$$

The right hand side of (4.5) is strictly greater than 2 for all $0 < u < \frac{1}{2\pi}$, therefore there exists some $\alpha > 1$ that satisfies the hypothesis of Theorem 7 in [71] and we can conclude that this sampler is geometrically ergodic.

4.3 The Witch's Hat Example

In this section we analyze the slice sampler algorithm for a simplified version of the "witch's hat distribution" [53]. The target distribution, (Figure 4.3), is

$$\pi(x) = \begin{cases} h & a < |x| < a + b \\ t + h & |x| \leq a \\ 0 & otherwise \end{cases}$$

where $a,\ b,\ h,\ t$ are non negative unknown parameters such that

$$2(a + b)h + 2at = 1. \tag{4.6}$$

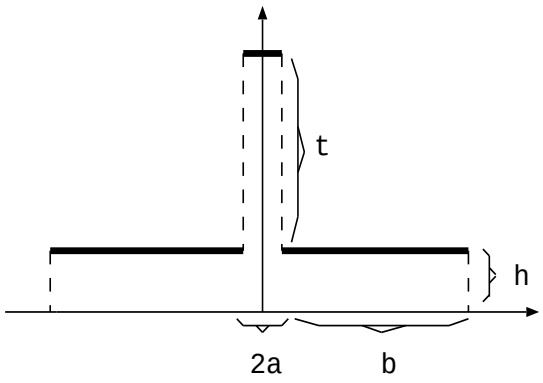

Figure 4.1: Modified witch's hat distribution

Consider the following uniform slice sampler: iteratively update

$$u|x \sim \begin{cases} U[0, h] & if \ \ |x| > a \\ \\ U[0, h + t] & if \ \ |x| \ \ \leq a \end{cases}$$

and

$$x|u \sim \begin{cases} U[-(a+b), a+b] & if \;\; 0 < u < h \\ \\ U[-a, a] & if \;\; h \leq u \leq h+t \end{cases}$$

For a better understanding of this sampler notice that when $|x| \leq a$ the horizontal component of the chain, x, is in the basis of the spike. Similarly, when $h \leq u \leq h+t$ the vertical component of the chain, u, is in the spike. This Markovian scheme produces a uniformly ergodic chain since Theorem 3.9.3 applies. The marginal X-chain has $\pi(x)$ as its limiting distribution. The transition kernel of this sub-chain is

$$K(x,y) = \frac{1}{2(a+b)} \qquad \text{if } \; |x| > a$$

and

$$K(x,y) = \frac{1}{h+t}\left\{\frac{h}{2(a+b)} + \frac{t}{2a}\right\} \qquad \text{if } \; |x| \leq a.$$

Consider the partition of the state space $\mathcal{X} = A \cup A^C$ where $A = \{x : |x| < a\}$. Both A and A^C are proper atoms for the X-chain. Relative to this partition, the one step transition matrix of the X-chain is

$$T = \begin{bmatrix} \frac{a(h+t)+tb}{(h+t)(a+b)} & \frac{hb}{(h+t)(a+b)} \\ \\ \frac{a}{a+b} & \frac{b}{a+b} \end{bmatrix}$$

where $T_{11} = P(X_{t+1} \in A,\ X_t \in A)$. The corresponding stationary distribution is

$$\pi_A = [2a(h+t),\ \ 2bh].$$

For a better understanding of what is going on, perform the following reparametrization: $\frac{a}{a+b} = s$ and $\frac{h}{h+t} = r$, thus s and r represents the relative base and hight of the spike respectively, which is all that matters for our purposes. Using the new parameterization the transition matrix relative to the introduced partition of the state space can be rewritten as

$$T = \begin{bmatrix} s + (1-r)(1-s) & r(1-s) \\ \\ s & 1-s \end{bmatrix}$$

The eigenvalues of the T matrix are $\lambda_0 = 1$ and

$$\lambda_1 = \frac{2atb}{(a+b)(1-2bh)} = (1-r)(1-s).$$

and the corresponding eigenvectors are $x_0 = [1, 1]$ and $x_1 = [\frac{r(s-1)}{s}, 1]$. As expected the eigenvalues are positive since this transition matrix represents a positive operator. Thus a small second eigenvalues is both an index of good performance in terms of asymptotic variance of MCMC estimates and of good convergence in total variation distance to stationarity.

The value of λ_1 decreases as $2at$ (area of the spike) or $2bh$ (combined area of the tails) decrease, or as $(a+b)$ (length of the base) increases. In terms of the new parameterization, λ_1 decreases as r or s increase to 1. This has a nice intuitive interpretation: this uniform slice sampler has better convergence behavior as the

target distribution, π, approaches a regular uniform distribution with no spikes.

On the other hand it is interesting to observe that, despite the fact that this Markov chain is uniform ergodic, its second largest eigenvalue can be made as close to 1 as we wish. For example, consider fixing the values of b and h and letting $t \to \infty$. This corresponds to stretching the top of the witch's hat and thus $r \to 0$. In order to keep the area under π equal to one we need to let a go to zero, that is, we need to reduce the base of the spike. The effect of this is that $s \to 0$ and thus $\lambda_1 \to 1$. The n-step transition matrix is

$$\begin{aligned} T^n &= \frac{1}{r(1-s)+s} \begin{pmatrix} s - r(s-1)[(r-1)(s-1)]^n & r(s-1)[[(r-1)(s-1)]^n - 1] \\ s - s[(r-1)(s-1)]^n & r - rs + s[(r-1)(s-1)]^n \end{pmatrix} \\ &= \frac{1}{r(1-s)+s} \begin{pmatrix} s - r(s-1)\lambda_1^n & r(s-1)(\lambda_1^n - 1) \\ s(1-\lambda_1^n) & r(1-s) + s\lambda_1^n \end{pmatrix} \end{aligned}$$

and the eigenvalues are the n-th power of the corresponding eigenvalues of T. As $n \to \infty$ we have that

$$T^n \to \frac{1}{r(1-s)+s} \begin{pmatrix} s & r(1-s) \\ s & r(1-s) \end{pmatrix} = T_\pi$$

thus we have convergence to the stationary distribution that, in the new parametrization, is $\pi_A = \frac{1}{r(1-s)+s}[s, \quad r(1-s)]$. In order to compute the total variation distance to stationarity $||T^n(x,\cdot) - \pi(\cdot)|| = 2\sup_{A \subset \mathcal{X}} |T^n(x,A) - \pi(A)|$ consider the matrix

$$T^n - T_\pi = \frac{\lambda_1^n}{r(1-s)+s} \begin{pmatrix} -r(s-1) & r(s-1) \\ -s & s \end{pmatrix}$$

thus the total variation distance is given by

$$2\lambda_1^n \frac{r(1-s)}{r(1-s)+s}$$

if we start inside the spike and

$$2\lambda_1^n \frac{s}{r(1-s)+s}$$

otherwise. Let

$$\frac{s}{r(1-s)+s} = p$$

so that p represents the area of the witch's hat inside the spike. Rewrite everything in terms of s and p: height of the spike

$$r = \frac{s(1-p)}{p(1-s)},$$

the second largest eigenvalue,

$$\lambda_1 = \frac{p-s}{p},$$

the total variation distance when the starting point is inside the spike

$$\frac{2\lambda_1^n r(1-s)}{r(1-s)+s} = 2\lambda_1^n(1-p) = 2(1-p)\left(\frac{p-s}{p}\right)^n,$$

and outside the spike

$$\frac{2\lambda_1^n s}{r(1-s)+s} = 2\lambda_1^n p = 2p\left(\frac{p-s}{p}\right)^n.$$

Notice that $0 \leq p \leq 1$ and $p \geq s$. By letting s tend to zero and p tend to one, that is by making the base of the spike very small and the area (probability) under the spike

very close to one, if we start outside the spike we are in trouble. The total variation distance to stationarity can be made arbitrarily close to 2 which is the maximum possible value.

The relevant question is, for a fixed shape of the witch's hat, that is for fixed values of $0 \leq s \leq 1$ and $0 \leq p \leq 1$ such that $s+p \leq 1$ and $p \geq s$, what is the minimum number of iterations n required to make the total variation distance to stationarity smaller than ϵ ? Suppose we start our Markov chain outside the spike. Then, by solving

$$2p\left(\frac{p-s}{p}\right)^n = \epsilon \tag{4.7}$$

with respect to n we get

$$n = \frac{log\left(\frac{\epsilon}{2p}\right)}{log\left(\frac{p-s}{p}\right)}. \tag{4.8}$$

Notice that if we take $n = 0$ in (4.7) we get that ϵ must be smaller than $2p$. For such ϵ we are guaranteed that $n \geq 0$. Figure 4.3 plots n as a function of s and p for a fixed value of the precision $\epsilon = 0.001$.

Since $\log(1-x) \approx -x$ for small values of x, from (4.8) it follows that

$$n \approx \frac{log\left(\frac{\epsilon}{2p}\right)}{-\frac{s}{p}} = \frac{p\log\left(\frac{\epsilon}{2p}\right)}{-s}$$

when s is close to zero. Both the plot and the analytic expression of the total variation distance to stationarity when starting outside the spike tell us that, as soon as s becomes sufficiently small, even for large values of p, the number of iterations necessary to reach a certain precision ϵ blows up.

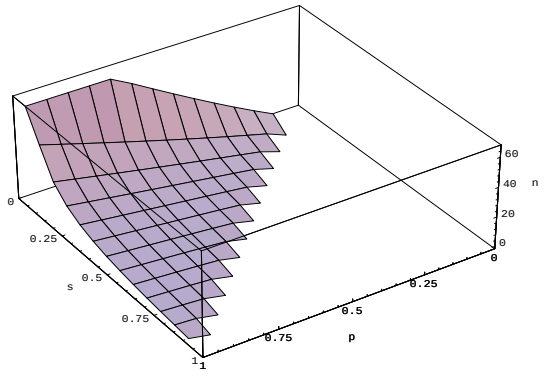

Figure 4.2: How bad can it get?

Chapter 5

Improving the Metropolis-Hastings Algorithm

There is a little village in Sicily where everybody knows everything about everybody else but nothing about themselves (the townsfolk love to gossip). One day the mayor calls everybody into the town square and starts this speech: "Dear citizens, there are werewolves among us. The only way to save the village is that, as soon as one finds, via logical deduction, that he is a werewolf, that same evening he should shoot himself." All agree and goes home. No shooting is heard the first and the second nights. Some shooting is heard the third night. How many werewolves were there?

5.1 Introduction

In Chapter 2 we introduced the Peskun ordering and proposed a weakening of it, the covariance ordering. In Chapter 3 we proved that the slice sampler dominates the independence Metropolis-Hastings algorithm in the Peskun sense and hence also in the covariance sense. In this chapter we propose to modify the Metropolis-Hastings algorithm to obtain a new sampler, the splitting rejection algorithm, which is proved to perform better than the original sampler in terms of asymptotic variance of MCMC estimates.

The key intuition behind the Peskun ordering is that every time $X_{t+1} = X_t = x$ our MCMC estimates become less efficient. For a Metropolis-Hasting algorithm this

happens when a candidate generated from $q(x, dy)$ is rejected. Therefore we will be able to improve the Metropolis-Hasting algorithm in the Peskun sense by reducing the number of rejected proposals. A way to achieve this goal is the following: whenever a candidate is rejected, instead of setting $X_{t+1} = X_t$ as in the regular Metropolis-Hastings algorithm, propose a new candidate. The acceptance-rejection probability of the new candidate has to be adjusted in order to preserve the stationary distribution. An interesting feature of this scheme is that the proposal distribution at the second stage, q_2, is allowed to depended on the rejected value at the first stage as well as on the starting value, so that $q_2 = q_2(x, y_1, dy_2)$.

If the candidate at the second stage is also rejected we could either set $X_{t+1} = X_t = x$ or move on to the third stage and so on. A combination of the two previous strategies can also be considered: every time a candidate is rejected toss a p-coin and if the outcome is heads (that is with probability p), propose a new candidate, otherwise let $X_{t+1} = X_t$. If we want to terminate the splitting process it is sufficient to set $p = 0$.

In order to take full advantage of the new proposed strategy we need to understand that, as the simulation proceeds, new information is acquired, namely the values of the target distribution at the previously rejected points (these values are computed at each iteration when evaluating the acceptance probability). Likewise, to solve the werewolves puzzle it is important to understand the information updating mechanism. As each night with no shooting goes by, we gain new information, namely that the werewolves have not been able to realize that they were such. We know there is at least one werewolf. If there were only one he will be able to realize he is a werewolf the very first day since he sees no other werewolf. Since we do not hear shooting the first night this means that there is more than one werewolf. Suppose now there are two. Each one of them sees the other one and, by the previous reasoning expects the other to kill himself the very first night. The morning of the second

day, not having heard any shooting, the two werewolves update their information and realize there must be more than one werewolf. Therefore, if there were 2 werewolves they would reason they are such the second day. But no shooting is heard the second night either thus there is more than 2 werewolves. Since some shooting is heard the third night we can conclude that there were 3 werewolves. If shooting were heard after a week we would have concluded that there were 7 werewolves.

Also when implementing the regular Metropolis-Hastings algorithm every time we compute the acceptance probability for a new candidate, say y, we evaluate $\pi(y)$. The problem with the Metropolis-Hastings algorithm is that this information cannot be used in later stages of the simulation because this would destroy the Markovian property of the scheme. The splitting rejection algorithm we propose allows us to use this information without losing the Markovian property.

The natural question that arises is how to take full advantage of the newly acquired information as the simulation proceeds by adjusting the proposal distribution in an efficient way. This issue will be discussed in Section 5.5.

5.2 The Splitting Rejection Algorithm

Here is a more detailed description of how the splitting rejection sampler works. Suppose the current position at time t is $X_t = x$. A candidate y_1 is generated from $q_1(x, dy_1)$ and accepted with probability

$$\begin{aligned} \alpha_1(x, y_1) &= 1 \wedge \frac{\pi(y_1) q_1(y_1, x)}{\pi(x) q_1(x, y_1)} \\ &= 1 \wedge \frac{N_1}{D_1} \end{aligned} \tag{5.1}$$

as in the regular Metropolis-Hastings algorithm. The same letter will be used to indicate both the distribution, $\pi(dx)$ and the density $\pi(x)$. If y_1 is rejected we can learn from this and accordingly modify the proposal distribution to $q_2(x, y_1, dy_2)$. The rejection suggests a bad fit of the current proposal and a better one should be constructed in light of this. In order to maintain the same stationary distribution we have to modify the acceptance probability of the new candidate, y_2. A possible (but not necessary) way to reach this goal is to impose the detailed balance condition and derive the acceptance probability that preserves it. The transition kernel for the two stage process for moving from x to $y_2 \neq x$ is

$$q_1(x, y_2)\alpha_1(x, y_2) + \int q_1(x, y_1)q_2(x, y_1, y_2)[1 - \alpha_1(x, y_1)]\alpha_2(x, y_1, y_2)dy_1. \tag{5.2}$$

For detailed balance to hold, we must have

$$\begin{aligned}\pi(x)q_1(x, y_2)\alpha_1(x, y_2) + \pi(x) \int q_1(x, y_1)q_2(x, y_1, y_2)[1 - \alpha_1(x, y_1)]\alpha_2(x, y_1, y_2)dy_1 = \\ \pi(y_2)q_1(y_2, x)\alpha_1(y_2, x) + \pi(y_2) \int q_1(y_2, y_1)q_2(y_2, y_1, x)[1 - \alpha_1(y_2, y_1)]\alpha_2(y, y_1, x)dy_1.\end{aligned} \tag{5.3}$$

Since the first terms on either side are equal, this means that we must have

$$\begin{aligned}\pi(x) \int q_1(x, y_1)q_2(x, y_1, y_2)[1 - \alpha_1(x, y_1)]\alpha_2(x, y_1, y_2)dy_1 = \\ \pi(y_2) \int q_1(y_2, y_1)q_2(y_2, y_1, x)[1 - \alpha_1(y_2, y_1)]\alpha_2(y, y_1, x)dy_1.\end{aligned} \tag{5.4}$$

A sufficient (but by no means necessary) condition is to make the integrands equal,

$$\pi(x)q_1(x,y_1)q_2(x,y_1,y_2)[1-\alpha_1(x,y_1)]\alpha_2(x,y_1,y_2) = \pi(y_2)q_1(y_2,y_1)q_2(y_2,y_1,x)[1-\alpha_1(y_2,y_1)]\alpha_2(y,y_1,x) \quad (5.5)$$

for all x, y_1, and y_2. By the standard Metropolis-Hastings argument this can be achieved by choosing

$$\begin{aligned}\alpha_2(x,y_1,y_2) &= 1 \wedge \frac{\pi(y_2)q_1(y_2,y_1)q_2(y_2,y_1,x)[1-\alpha_1(y_2,y_1)]}{\pi(x)q_1(x,y_1)q_2(x,y_1,y_2)[1-\alpha_1(x,y_1)]} \\ &= 1 \wedge \frac{N_2}{D_2}.\end{aligned} \quad (5.6)$$

If the second stage is reached, it means that $N_1 < D_1$ and we can therefore replace $\alpha_1(x,y_1)$ with $\frac{N_1}{D_1}$ in D_2 and obtain:

$$\begin{aligned}\alpha_2(x,y_1,y_2) &= 1 \wedge \frac{N_2}{q_2(x,y_1,y_2)[\pi(x)q_1(x,y_1) - \pi(y_1)q_1(y_1,x)]} \\ &= 1 \wedge \frac{N_2}{q_2(x,y_1,y_2)[D_1 - N_1]}.\end{aligned} \quad (5.7)$$

The general i-th stage of the splitting rejection algorithm works as follows. If the candidate y_{i-1} proposed at the previous stage is not accepted, generate y_i from $q_i(x,y_1,\cdots,,dy_i)$ and accept it with probability

$$\begin{aligned}\alpha_i(x,y_1,y_2,\cdots,y_i) &= 1 \wedge \left\{ \frac{\pi(y_i)q_1(y_i,y_{i-1})q_2(y_i,y_{i-1},y_{i-2})\cdots,q_i(y_i,y_{i-1},\cdots,x)}{\pi(x)q_1(x,y_1)q_2(x,y_1,y_2)\cdots q_i(x,y_1,\cdots,y_i)} \right. \\ &\qquad \left. \frac{[1-\alpha_1(y_i,y_{i-1})][1-\alpha_2(y_i,y_{i-1},y_{i-2})]\cdots[1-\alpha_{i-1}(y_i,y_{i-1},\cdots,y_1)]}{[1-\alpha_1(x,y_1)][1-\alpha_2(x,y_1,y_2)]\cdots[1-\alpha_{i-1}(x,y_1,\cdots,y_{i-1})]} \right\} \\ &= 1 \wedge \frac{N_i}{D_i}.\end{aligned} \quad (5.8)$$

Again, if the i-th stage is reached, it means that $N_j < D_j$ for $j = 1, \cdots, i-1$, therefore $\alpha_j(x, y_1, \cdots, y_j)$ can be rewritten as $\frac{N_j}{D_j},\ j = 1, \cdots, i-1$ and we obtain the recursive formula

$$D_i = q_i(x, \cdots, y_i)(D_{i-1} - N_{i-1})$$

which leads to

$$\begin{aligned} D_i = q_i(x, \cdots, y_i)[q_{i-1}(x, \cdots, y_{i-1})[q_{i-2}(x, \cdots, y_{i-2}) \cdots \\ [q_2(x, y_1, y_2)[q_1(x, y_1)\pi(x) - N_1] - N_2] - N_3] \cdots - N_{i-1}]. \end{aligned} \tag{5.9}$$

The described procedure gives rise to a Markov chain which is reversible with invariant distribution π.

5.3 The Symmetric Splitting Rejection Algorithm

Consider now the special case of a symmetric proposal distribution that only depends on the last rejected candidate value:

$$q_i(x, y_1, \cdots, y_{i-1}, dy_i) = q(y_{i-1}, dy_i) = q(y_i, dy_{i-1}).$$

In this setting

$$\alpha_1(x, y_1) = 1 \wedge \frac{\pi(y_1)}{\pi(x)}$$

that is: if $\pi(y_1) \geq \pi(x)$ accept y_1 and set $X_{t+1} = y_1$, otherwise accept y_1 with probability $\frac{\pi(y_1)}{\pi(x)}$. Notice that this acceptance probability is the same as in the Metropolis-Hastings algorithm when a symmetric proposal distribution is used.

If y_1 is rejected generate y_2 from $q(y_1, dy_2)$ and accept it with probability

$$\alpha_2(x, y_1, y_2) = 1 \wedge \frac{\pi(y_2)\left[1 - 1 \wedge \frac{\pi(y_1)}{\pi(y_2)}\right]}{\pi(x) - \pi(y_1)}. \tag{5.10}$$

Three cases can occur at this point:

1. if $\pi(y_2) \geq \pi(x)$ then $\alpha_2(x, y_1, y_2) = 1$, thus accept y_2 and set $X_{t+1} = y_2$;
2. if $\pi(y_2) < \pi(y_1)$ then $\alpha_2(x, y_1, y_2) = 0$ thus reject y_2 and move to the next stage;
3. if $\pi(x) > \pi(y_2) \geq \pi(y_1)$ accept y_2 with probability $\alpha_2(x, y_1, y_2) = \frac{\pi(y_2) - \pi(y_1)}{\pi(x) - \pi(y_1)}$.

Rewrite (5.10) as:

$$\begin{aligned} \alpha_2(x, y_1, y_2) &= 1 \wedge \frac{0 \vee [\pi(y_2) - \pi(y_1)]}{\pi(x) - \pi(y_1)} \\ &= F\left(\frac{\pi(y_2) - \pi(y_1)}{\pi(x) - \pi(y_1)}\right) \end{aligned}$$

where F is the cumulative distribution function of a uniform random variable on the interval $(0, 1)$.

If y_2 is rejected, move on to the next stage and draw a new candidate from $q(y_2, dy_3)$. The acceptance probability for this new candidate is:

$$\alpha_3(x, y_1, y_2, y_3) = 1 \wedge \frac{\pi(y_3)}{\pi(x)} \frac{[1 - \alpha_1(y_3, y_2)]\,[1 - \alpha_2(y_3, y_2, y_1)]}{\left[1 - \frac{\pi(y_1)}{\pi(x)}\right]\,[1 - \alpha_2(x, y_1, y_2)]}. \tag{5.11}$$

Again, if the third stage is reached $\pi(y_1) < \pi(x)$ and $\pi(y_2) < \pi(x)$. Equation (5.11) can be rewritten as:

$$\alpha_3(x, y_1, y_2, y_3) = 1 \wedge \frac{0 \vee \{\pi(y_3) - [\pi(y_1) \vee \pi(y_2)]\}}{\pi(x) - [\pi(y_1) \vee \pi(y_2)]}.$$

Three cases can occur at this point:

1. if $\pi(y_3) \geq \pi(x)$ then $\alpha_3(x, y_1, y_2, y_3) = 1$, thus accept y_3 and set $X_{t+1} = y_3$;

2. if $\pi(y_3) < [\pi(y_1) \vee \pi(y_2)]$, then $\alpha_3(x, y_1, y_2, y_3) = 0$, thus reject y_3 and move to the next stage;

3. otherwise accept y_3 with probability

$$\frac{\pi(y_3) - [\pi(y_1) \vee \pi(y_2)]}{\pi(x) - [\pi(y_1) \vee \pi(y_2)]}.$$

The i-th stage works as follows. If y_{i-1} is rejected generate y_i from $q(y_{i-1}, dy_i)$. Note that if the i-th stage is reached then $\pi(y_j) < \pi(x)$ for all $j < i$. Let $y^* = \text{argmax}_{j<i}\pi(y_j)$. The acceptance probability for y_i is given by:

$$\alpha_i(x, y_1, \cdots, y_i) = 1 \wedge \frac{0 \vee [\pi(y_i) - \pi(y^*)]}{\pi(x) - \pi(y^*)} \tag{5.12}$$

Again (5.12) can be interpreted this way:

1. if $\pi(y_i) \geq \pi(x)$ accept y_i;

2. if $\pi(y_i) < \pi(y^*)$ reject y_j and move to the next stage;

3. otherwise accept y_j with probability

$$P_{x,i} = \frac{\pi(y_i) - \pi(y^*)}{\pi(x) - \pi(y^*)}. \tag{5.13}$$

A visualization of the described procedure is given in Figure 5.3.

The acceptance-rejection probability in (5.13) is considerably simpler than (5.8) but is obtained under the strong assumption that the proposal distribution can be improved only on the basis of the last rejected candidate. In some sense this

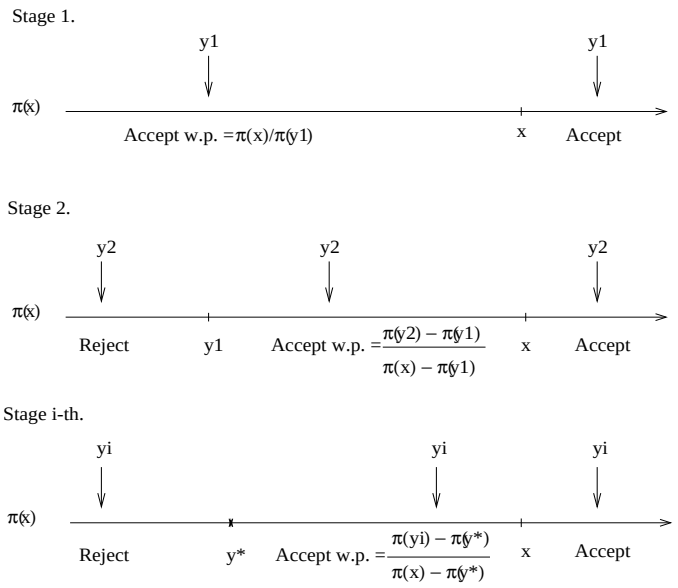

Figure 5.1: Splitting rejection algorithm

assumption weakens the power of the splitting rejection approach which comes from adjusting the proposal in light of all the previously rejected candidates.

An interesting question is the following. Can we show that $P_{x,i}$, or, more generally, that $\alpha_i(x, y_1, \cdots, y_i)$, tends to one as i tends to infinity? That is, can we show that eventually a candidate will be accepted? Not in the general setting presented because nothing guarantees that the proposal distribution will eventually propose candidates y_i such that $\pi(y_i) > \pi(y^*)$, that is, candidates that have some chances of being accepted. This would suggest that every time a candidate is rejected we move to the next stage and propose a new candidate with some probability, say p, while with probability $1 - p$ we set $X_{t+1} = X_t = x$, that is, the current state in retained. This modification of the sampler does not affect the acceptance probability that was computed computed before as the following reasoning shows. The transition kernel of the two stage process for moving from x to $y_2 \neq x$, provided we insert this

additional p-coin step, is given by:

$$q_1(x, y_2)\alpha_1(x, y_2) + p \int q_1(x, y_1)q_2(x, y_1, y_2)[1 - \alpha_1(x, y_1)\alpha_2(x, y_1, y_2)dy_1.$$

A sufficient condition for detailed balance condition to hold is:

$$p\pi(x)q_1(x, y_1)q_2(x, y_1, y_2)[1 - \alpha_1(x, y_1)\alpha_2(x, y_1, y_2) = \\ p\pi(y_2)q_1(y_2, y_1)q_2(y_2, y_1, x)[1 - \alpha_1(y_2, y_1)\alpha_2(y, y_1, x) \quad (5.14)$$

for all x, y_1 and y_2. Since p cancels we get the same acceptance probability as before. Similar reasoning can be repeated for subsequent stages.

5.4 Independence Splitting Rejection Algorithm

Consider now the case of a proposal distribution at stage i that does not depend on the starting value x nor on the previously rejected candidates $y_1, \cdots, y_{i-1}$

$$q_i(x, y_1, \cdots, y_{i-1}, dy_i) = q(dy_i).$$

Again, if we choose such a proposal we do not take advantage of the new information acquired, still, let us consider the acceptance rejection probability of this scheme. The situation is similar to the one for the symmetric proposal with the difference that the quantity $\pi(y_i)$ is replaced by $w(y_i)$ where $w(x) = \frac{\pi(x)}{q(x)}$. Thus, for example, the acceptance rejection probability of the first step is the same as in the independence Metropolis-Hastings sampler [83]:

$$\alpha_1(x, y_1) = 1 \wedge \frac{w(y_1)}{w(x)} \quad (5.15)$$

and in general the acceptance probability for y_i at the i-th stage is

$$\alpha_i(x, y_1, \cdots, y_i) = 1 \wedge \frac{0 \vee [w(y_i) - w(y^*)]}{w(x) - w(y^*)} \tag{5.16}$$

where $y^* = \text{argmax}_{j<i} w(y_j)$.

5.5 Adjusting the Proposal Distribution

We now turn to the question of how to choose a proposal distribution for the candidate values. Say we are currently at stage i. The proposal distribution $q_i(x, y_1, \cdots, dy_i)$ is allowed to depend on all the previously proposed but rejected candidates. In the sequel we will outline a few methods to take advantage of this newly acquired information.

5.5.1 Griddy Proposals

At stage i the values of $\pi(y_j), j = 1, \cdots, (i-1)$ are available, that is we have an evaluation of the target density on a grid. The idea is to use these values to obtain an approximation of $\pi(\cdot)$ (or of its inverse cumulative distribution function) and generate candidates with approximately the right distribution. A similar idea is used in the Griddy Gibbs Sampler [70] and variations on it are suggested in [83].

Let $x = y_0$ and set

$$q_i(x, y_1, \cdots, dy_i) = \frac{\sum_{j=0}^{i-1} \pi(y_j) f(y_i - y_j)}{\sum_{j=0}^{i-1} \pi(y_j)} \tag{5.17}$$

where f is a proper density. Using such a candidate generation function is equivalent to first selecting a point y_k from $y_0, \cdots, y_{i-1}$ according to a distribution that is proportional to the density values $\pi(y_0), \cdots, \pi(y_{i-1})$. Then generate an increment Z

from f and add this increment to y_k to obtain $y_i = y_k + Z$.

The scheme described above is equivalent to using a piecewise constant approximation to the inverse cumulative distribution function of $\pi(\cdot)$ for some choices of f. More sophisticated approximations may be used such as piecewise linear, piecewise quadratic interpolations or higher-order splines.

Using (5.17) we get:

$$\begin{aligned} q_1(x, dy_1) &= f(y_1 - x) \\ q_2(x, y_1, dy_2) &= \frac{\pi(dx) f(y_2 - x) + \pi(y_1) f(y_2 - y_1)}{\pi(dx) + \pi(y_1)}. \end{aligned} \tag{5.18}$$

and so on.

5.5.2 Trust Region Proposals

Another approach is motivated by the model-trust region method of optimization [3]. The initial proposal distribution might be normal based on a local quadratic approximation to the log posterior density. If the initial proposal is rejected this suggests that the approximation is not sufficiently accurate globally but it might work locally therefore, the second stage distribution can be based on the same approximation, but restricted to a neighborhood around the current point. Since differentiability guarantees that a quadratic approximation will work in a sufficiently small neighborhood, third and higher stages could further reduce the trust region supporting the proposal distribution. A variety of strategies for reducing the region are possible. They can be deterministic, based on the rejected proposal itself, or on the unnormalized posterior density at the previously rejected proposals.

5.5.3 Independence plus Random Walk Proposals

As anticipated in Section 5.1 a first stage (early stages) independence proposal can be combined with a second stage (later stages) random walk proposal. If the distribution used for the independence proposal is indeed a good approximation of the target distribution, then the initial proposal will rarely fail and the random walk proposal will rarely be used. But if the independence distribution does not provide a good approximation then the random walk gives protection against the potentially poor behavior of an independence chain with an unfavorable proposal distribution.

5.5.4 Splitting Rejection and Diffusions

In the traditional Metropolis-Hastings algorithm the choice of the candidate is not usually guided by the target distribution. The big exception is given by a new type of algorithm known as self-targeting Metropolis-Hastings algorithm. The idea was introduced in [13] and [7] and since then this area of research has been very active mostly thanks to the contribution of Stramer and Tweedie [79] [80] and Roberts and Tweedie [75].

One of the first constructive attempts to link diffusions processes and Markov chain Monte Carlo was introduced in [75]. The authors consider the candidate distribution

$$q_L(x, \cdot) = N\left(x + \frac{1}{2}h\nabla \log \pi(x), h\right)$$

where $h > 0$, N is the standard normal distribution and ∇ is the gradient operator $\nabla f(x) = \frac{df}{dx}$. The intuition behind this proposal is that q_L are the Euler discretizations of the h-step transition probabilities for a Langevin diffusion process that has π as its stationary distribution.

By exploiting this idea within the splitting rejection framework we could take

the proposal distribution at stage $i+1$ to be

$$q_{i+1}(x, y_1, \cdots, y_i, dy_{i+1}) = N\left(x + \frac{1}{2}h\frac{1}{i+1}\sum_{j=0}^{i}\nabla \log \pi(y_j), h\right)$$

where $y_0 = x$. This means that every time we propose a candidate y we evaluate both $\pi(y)$ and $\nabla \log \pi(y)$ and use this information to construct a better proposal distribution for the next stage.

Unfortunately the Euler discretization does not seem to perform too well due to occasional overshooting, that is, sometimes the value of h is too big and we move from a low probability area to another low probability area under the target distribution and we keep jumping around failing to visit the relevant part of the state space.

This problem is solved by using the discrete approximation scheme proposed in [63, 78, 77, 79] which is proven to be more stable. Stramer and Tweedie [79] also introduce a wider class of diffusions to obtain better rates of convergence to stationarity in total variation distance. For example, if the density is $e^{(-2\sqrt{|x|})}$, then the drift for the Langevin diffusion is $\frac{-1}{\sqrt{x}}$, when $x > 0$, and for large x the jumps are very small. This canse be corrected by choosing a different diffusion within the class considered in [79]. A proposal distribution within this wider class is defined by specifying two functions b and σ which satisfy the equation

$$b(x) = [\frac{1}{2}\nabla \log \pi(x)]\sigma^2(x) + \sigma(x)\nabla\sigma(x)]. \tag{5.19}$$

Given any such functions the self-targeting proposal distribution is

$$q_{ST}(x, \cdot) = N(\mu_{x,h}, \sigma^2_{x,t}) \tag{5.20}$$

where

$$\mu_{x,h} = xe^{\left(\frac{b(x)h}{x}\right)}$$

and

$$\sigma^2_{x,t} = \frac{x\sigma^2(x)}{2b(x)}\left[1 - e^{\frac{2b(x)h}{x}}\right]$$

for some $h > 0$. The choice of this class of proposal distributions is motivated by the fact that a diffusion process with drift b and speed measure σ satisfying (5.19) has stationary measure π.

In light of this new idea we could adjust our proposal distribution as the simulation within the splitting rejection algorithm proceeds as follows. At each stage i evaluate $\sigma(y_i)$ as well as $\nabla \log \pi(y_i)$. Let

$$\sigma_i(x, y_1, \cdots, y_i) = \sigma_i = \frac{1}{i+1}\sum_{j=1}^{i}\sigma(y_j)$$

$$b_i(x, y_1, \cdots, y_i) = \left(\frac{1}{i+1}\sum_{j=0}^{i}\nabla \log \pi(y_j)\right)\frac{\sigma_i^2}{2} + \sigma_i \nabla \sigma_i$$

$$\mu_{((x,y_1,\cdots,y_i),h)} = \mu_{i,h} = xe^{\left(\frac{b_i h}{x}\right)}$$

$$\sigma^2_{((x,y_1,\cdots,y_i),h)} = \sigma_{i,h} = \frac{x\sigma_i^2}{2b_i}\left[1 - e^{\frac{2b_i h}{x}}\right].$$

Take the proposal distribution at the $(i+1)$-th stage to be

$$q_{i+1}(x, y_1, \cdots, y_i, dy_{i+1}) = N(\mu_{i,h}, \sigma^2_{i,h}).$$

Alternatively we could use the rejected values $y_0, y_1, \ldots y_i$ choose a value of h and/or to improve on the functional form of σ or b. Notice that regarding the latter we only have one degree of freedom here since the two funcions should satiefy equation (5.19).

5.5.5 General Guidelines

Here is some general advice on how to choose the proposal distributions. At earlier stages we would suggest to use proposals that are easy to construct (not comuputationallly intensive) such as first and second order approximations. As the simulation proceed we should refine the proposal by using higher order approximations.

Also, it seems a good idea to use proposals centered at the current point but with inflated variance during earlier stages. As the simulation proceeds reduce the variance of the proposal so that the acceptance probability increases.

A way of combining these two ideas is to construct a mixture proposal: $p_1 q_1 + (1 - p_1)q_2$ where $0 \leq p_1 \leq 1$, q_1 in an independence proposal with inflated variance and q_2 is an approximation to π based on first and second order derivatives. At early stages let p_1 be much bigger than $\frac{1}{2}$. If we keep rejecting the proposals decrease the value of p_1, decrease the variance of q_1 and improve the approximation of q_2 to π.

5.6 Comparing Splitting Rejection and Metropolis-Hastings Algorithms

A little bit of care is needed when we say that the splitting rejection algorithm "performs better" than the Metropolis-Hastings algorithm.

Up to now we have only compared the two schemes in terms of the Peskun ordering, that is, the asymptotic variance of MCMC estimates obtained using the splitting rejection algorithm is smaller than the one obtained using the Metropolis-Hastings scheme. But it is important to realize that one iteration in the splitting rejection scheme, that is one move from X_t to X_{t+1} may take considerably more CPU time than one iteration in the Metropolis-Hastings algorithm. This is due to the fact that when implementing splitting rejection algorithms we have to generate the

sequence $y_1, y_2, \ldots$ and evaluate the acceptance probabilities many times.

A more reasonable comparison should be made by taking into account this lack of symmetry between the two schemes. One possibility is to compare the asymptotic variance of the two competing schemes given a fixed number of evaluations of the target distribution.

Let τ_{SR} be the mean CPU time needed for the splitting rejection algorithm to go from X_t to X_{t+1}. Let τ_{MH} be the mean CPU time needed for the Metropolis-Hastings algorithm to move from X_t to the first $X_{t+j} \neq X_t, j = 1, 2 \ldots$. Consider implementing a splitting rejection algorithm where the proposal distribution is not adjusted to take advantage of the newly acquired information. In other words, take the proposal distribution used for the splitting rejection algorithm at any stage to be the same as the one used for the Metropolis-Hastings algorithm. Then τ_{SR} and τ_{MH} should be comparable. But if we construct the proposal distribution in the splitting rejection algorithm in a clever way (following the guidelines given in Sections 5.5.1, 5.5.2 and 5.5.3 for example) we should be able to make $\tau_{SR} < \tau_{MH}$. This means that we should be able to construct splitting rejection algorithms that outperform Metropolis-Hastings algorithms not only because they have smaller asymptotic variance of MCMC estimates but also because they have smaller asymptotic variance for a given fixed amount of CPU time.

5.7 Splitting Rejection for Efficient Slice Samplers

The efficiency of auxiliary variable algorithms relies on the fact that the conditional distribution of the variable of interest, X, given the auxiliary variable, U, can be sampled in a computationally convenient way. In the slice sampling framework $\pi(x|u)$ is $q(x)$ restricted to the set $A_{u,l} = \{x : l(x) > u\}$. In the uniform slice sampler framework $\pi(x|u)$ is uniform on the set $A_{u,\pi} = \{x : \pi(x) > u\}$. For simplicity in the sequel we will refer to the uniform case but the general case can be handled in a similar way.

In order to efficiently sample the uniform conditional distribution we propose an adaptive MCMC method that makes use of the splitting rejection algorithm introduced in Section 5.1. The first step is to find an interval $I = [L_1(x), R_1(x)]$ such that

- I contains the current point x;
- I contains as much of the slice $A_{u,\pi}$ as possible, that is the set $A_{u,\pi} \setminus I$ is as small as possible;
- I contains as little of $A^c_{u,\pi}$ as possible, that is the set $I \setminus A_{u,\pi}$ is as small as possible.

Ideally we would like to have $L_1(x) = \inf_x(A_{u,\pi})$ and $R_1(x) = \sup_x(A_{u,\pi})$, that is, I should be the smallest interval that contains the whole slice. If the range of π is bounded we might set $I = R(\pi)$, that is we let our interval coincide with the range of the target distribution (or of q if we are implementing the more general slice sampler). This may be inefficient since the slice is typically smaller than the range. Alternatively we can exploit some additional information on the typical size of the slice, say w. If this information is not available we could randomly choose an initial interval size w. Given w, an interval around the current position x is

constructed either in a deterministic fashion by setting $I = [x - w; x + w]$, or via a random procedure. Notice that the proposed deterministic construction gives an interval centered around x. Two interesting random procedures are suggested in [59]: the "stepping out" and the "doubling" procedure.

Once an interval I has been found we propose two strategies. A *rejection algorithm* can be implemented having the uniform distribution on I as the proposal distribution to uniformly sample from $A_{u,\pi}$. This means that we randomly draw points from the interval I and we retain the first one that lies in $A_{u,\pi}$. Some care is required because we have to make sure that the detailed balance condition holds. This means that, as discussed in [59], we need to define the set B of acceptable successors states

$$B = \{y : y \in A_{u,\pi} \cap I \text{ and } P(\text{select } I|\text{at state } x) = P(\text{select } I|\text{at state } y)\} \quad (5.21)$$

and accept $y \sim U[I]$ only if the condition $y \in A_{u,\pi} \cap B$ is met. One advantage of exploiting a random construction to define the initial interval I, is that it makes it easier to verify the detailed balance condition and sometimes detailed balance is guaranteed to hold without further checking, that is there is no need to check weather a proposal lies in the set B defined in 5.21. This is the case for the "stepping out" procedure proposed in [59].

Alternatively we can perform an *adaptive rejection algorithm*. Following the guidelines given in Section 5.1 regarding the construction of the splitting rejections algorithm let the first stage proposal distribution be

$$q_1(x, dy_1) \sim U[L_1(x); R_1(x)].$$

The first stage acceptance probability is then

$$\alpha_1(x, y_1) = 1 \wedge \frac{R_1(x) - L_1(x)}{R_1(y_1) - L_1(y_1)} I_{A_{u,\pi}}(y_1)\, I_{[L_1(y_1);R_1(y_1)]}(x). \tag{5.22}$$

This acceptance probability comes from (5.1) since

$$D_1 = \frac{1}{\mu^1(A_{u,\pi})} I_{A_{u,\pi}}(x) \frac{1}{R_1(x) - L_1(x)} I_{[L_1(x);R_1(x)]}(y_1)$$

and

$$N_1 = \frac{1}{\mu^1(A_{u,\pi})} I_{A_{u,\pi}}(y_1) \frac{1}{R_1(y_1) - L_1(y_1)} I_{[L_1(y_1);R_1(y_1)]}(x).$$

Clearly $I_{A_{u,\pi}}(x) = 1$ and $I_{[L_1(x);R_1(x)]}(y_1) = 1$ thus giving (5.22). Let the second stage proposal distribution be

$$q_2(x, y_1, dy_2) \sim U[L_2(x, y_1); R_2(x, y_1)].$$

The second stage acceptance probability is

$$\begin{aligned}
\alpha_2(x, y_1, y_2) =& 1 \wedge I_{A_{u,\pi}}(y_2) I_{[L_1(y_2);R_1(y_2)]}(y_1) I_{[L_2(y_2,y_1);R_2(y_2,y_1)]}(x) && (5.23)\\
&\times \frac{[R_1(x) - L_1(x)][R_2(x, y_1) - L_2(x, y_1)]}{[R_1(y_2) - L_1(y_2)][R_2(y_2, y_1) - L_2(y_2, y_1)]} && (5.24)\\
&\times \frac{1 - 1 \wedge \frac{R_1(y_2) - L_1(y_2)}{R_1(y_1) - L_1(y_1)} I_{[L_1(y_1);R_1(y_1)]}(y_2) I_{A_{u,\pi}}(y_1)}{1 - \frac{R_1(x) - L_1(x)}{R_1(y_1) - L_1(y_1)} I_{A_{u,\pi}}(y_1) I_{[L_1(y_1);R_1(y_1)]}(x)}. && (5.25)
\end{aligned}$$

Notice that if y_1 was not accepted because did not belong to $A_{u,\pi}$ then (5.25) is equal to one. If y_1 was not an acceptable move in the sense that it did not lie in B, then

the denominator of (5.25) will be one. Take the i^{th} stage proposal distribution to be

$$q_i(x, y_1, \cdots, dy_i) \sim U[L_i(x, y_1 \cdots, y_{i-1}); R_i(x, y_1 \cdots, y_{i-1})].$$

Thus the i^{th} stage acceptance probability is

$$\begin{aligned}\alpha_i(x, y_1, \cdots, y_i) =& 1 \wedge I_{A_{u,\pi}}(y_i) I_{[L_1(y_i);R_1(y_i)]}(y_{i-1}) \cdots I_{[L_i(y_i,\cdots y_1);R_i(y_i,y_1)]}(x) \qquad (5.26)\\ &\times \frac{[R_1(x) - L_1(x)] \cdots [R_i(x, y_1, \cdots y_{i-1}) - L_i(x, y_1, \cdots y_{i-1})]}{[R_1(y_i) - L_1(y_i)] \cdots [R_i(y_i, \cdots y_1) - L_i(y_i, \cdots, y_1)]} \qquad (5.27)\\ &\times \frac{[1 - \alpha_1(y_i, y_{i-1})][1 - \alpha_2(y_i, y_{i-1}, y_{i-2})] \cdots [1 - \alpha_{i-1}(y_i, y_{i-1}, \cdots, y_1)]}{[1 - \alpha_1(x, y_1)][1 - \alpha_2(x, y_1, y_2)] \cdots [1 - \alpha_{i-1}(x, y_1, \cdots, y_{i-1})]}. \qquad (5.28)\end{aligned}$$

The above strategy describes a wide class of algorithms. An element in such class is specified by defining the functions $R_i(x, y_1, \cdots, y_{i-1})$ and $L_i(x, y_1, \cdots, y_{i-1})$ for every $i = 1, 2, \cdots$. One possibility is to set

$$R_i(x, y_1, \cdots, y_{i-1}) = x + |x - y_{i-1}|$$

and

$$L_i(x, y_1, \cdots, y_{i-1}) = x - |x - y_{i-1}|$$

so that

$$R_i(x, y_1, \cdots, y_{i-1}) - L_i(x, y_1, \cdots, y_{i-1}) = \frac{1}{2|x - y_{i-1}|}$$

for every $i = 2, 3, \cdots$. Let $R_1(x) = x + w$ and $L_1(x) = x - w$. This results in a

symmetric shrinking of the interval used to define the uniform proposal distribution around the current position x. In this case the condition $y_i \in B$ is always satisfied for every possible value i. An alternative strategy is to set

$$R_i(x, y_1, \cdots, y_{i-1}) = R_{i-1}(x, y_1, \cdots, y_{i-2})$$

and

$$L_i(x, y_1, \cdots, y_{i-1}) = y_{i-1}$$

if $y_{i-1} \leq x$ and

$$R_i(x, y_1, \cdots, y_{i-1}) = y_{i-1}$$

and

$$L_i(x, y_1, \cdots, y_{i-1}) = L_{i-1}(x, y_1, \cdots, y_{i-2})$$

if $y_{i-1} > x$, for $i = 2, 3, \cdots$. $R_1(x)$ and $L_1(x)$ are defined as before. This results in a non-symmetric shrinking of the interval used to define the uniform proposal distribution. In this case the condition $y_i \in B$ can fail.

In both cases the proposal distribution seems to only depend on the last rejected candidate but we are really learning from all the previously rejected proposals thus fully taking advantage of the potentials of the adaptive strategy introduced in Section 5.1. As we move on to successive stages the variance of the proposal distribution decreases and the distribution becomes more and more concentrated around the

last accepted point x: this increases the acceptance probability of points proposed at later stages thus avoiding situations where we keep proposing and rejecting.

Notice that the rejection algorithm proposed at the beginning of this section is nothing but a special case of the adaptive rejection algorithm just described and so are the shrinking procedures proposed in Section 4.2 of [59]. This means that if we recompute the acceptance rejection probability as in (5.26) we do not need to test weather a point belongs to the set B anymore or, rather, this checking is automatically performed when computing the acceptance-rejection ratio. In other words, if a candidate is not a valid acceptable successor then the corresponding acceptance ratio as computed in (5.26) will be zero.

Bibliography

[1] D. J. Aldous and H. Thorisson. Shift-coupling. *Stochastics Processes and Applications*, 44:1–14, 1993.

[2] J. S. Aujla and H.L. Vasudeva. Convex and monotone operator functions. *Annales Polonici mathematici*, 62:1–11, 1995.

[3] D. M. Bates and D. G. Watts. *Nonlinear Regression Analysis and Its Applications*. John Wiley and Sons, New York, 1988.

[4] R. Bellman. *Introduction to Matrix Analysis*. McGraw-Hill, N.Y., 1972.

[5] J. Bendat and S. Sherman. Monotone and convex operator functions. *Transactions Amer. Math. Soc.*, 79:58–71, 1955.

[6] J. Besag and P. J. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Series B.*, 55:25–37, 1993.

[7] J. E. Besag. Comments on "Representations of knowledge in complex systems" by U. Grenander and M.I. Miller. *Journal of the Royal Statistical Society, Series B*, 56, 1994.

[8] K. S. Chan and C. J. Geyer. Discussion of the paper by Tierney. *Ann. Statist.*, 22, 1994.

[9] J. B. Conway. *A Course in Functional Analysis*. Springer-Verlag, 1985.

[10] J. N. Corcoran and Tweedie R. L. Perfect sampling from independent Metropolis-Hastings chains. Preprint at http://www.stats.bris.ac.uk/MCMC.

[11] J. Damien, P. Wakefield and S. Walker. Gibbs sampling for bayesian non-conjugate and hierarchical models using auxiliary variable. *Journal of the Royal Statistical Society, Series B. To Appear.*, 1998.

[12] P. Diaconis, S. Holmes, and R. M. Neal. Analysis of a non-reversible Markov chain sampler. Technical report, Cornell University, 1997. No. BU-1385-M.

[13] J. D. Doll, P. J. Rossky, and H. L. Friedman. Brownian dynamics as smart Monte Carlo simulation. *Journal of Chemical Physics*, 69:4628–4633, 1978.

[14] N. Dunford and J. J. Schwartz. *Linear Operators Part II. First edition.* John Wiley and Sons, New York, 1963.

[15] R. G. Edwards and A. D. Sokal. Generalization of the Fortium-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Phys. Rev. D*, 38:2009–2012, 1988.

[16] J. A. Fill. An interruptible algorithm for perfect sampling via Markov chains. *Annals of Applied Probability*, 7, 1998.

[17] G. S. Fishman. An analysis of Swendsen-Wang and related sampling methods. Technical report, Dept. of Operations research, University of North Carolina, 1992. Tech. Rep. No UNC/OR/TR96-04.

[18] S. Foss and Tweedie R. L. Perfect simulation and backward coupling. Preprint at http://www.stats.bris.ac.uk/MCMC.

[19] S. Foss, R. L. Tweedie, and Corcoran J. Simulating the invariant measures of Markov chains using horizontal backward coupling at regeneration times. *Stochatsic Models*, 1998. To appear.

[20] A. Frigessi, C. Hwang, and L. Younes. Optimal spectral structure of reversible stochastic matrices, Monte carlo methods and the simulation of Markov random fields. *The Annals of Applied Probability*, 2:610–628, 1992.

[21] B. Fristed and L. Gray. *A Modern Approach to Probability Theory*. Birkhäuser, Boston, 1997.

[22] A. Gelman. Inference and monitoring convergence. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[23] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In E. M. Keramidas, editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Fairfax Station: Interface Foundation, 1991.

[24] C. J. Geyer. Practical markov chain monte carlo. *Statistical Science*, 7:473–511, 1992.

[25] C. J. Geyer and E. A. Thomson. Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc. B*, 38:657–699, 1992.

[26] J. C. Geyer. *Markov Chain Monte Carlo Lecture Notes*. Unpublished, 1998.

[27] W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics*, 44:455–472, 1994.

[28] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[29] W R. Gilks, G O. Roberts, and E. George. Adaptive direction sampling. *The Statistician*, 43:179–190, 1994.

[30] M. I. Gordin and B. A. Lifšic. The central limit theorem for stationary Markov processes. *Soviet Math. Dokl.*, 19:392–394, 1978.

[31] P. J. Green. Contribution to the discussion of "the EM algorithm - an old folksong to a fast new tune" by X. L. Meng and D. van Dyk. *J. Roy. Statist. Soc. Ser. B*, 59:511–567, 1997.

[32] J. Größ. Some remarks on the Löwner partial ordering of Hermitian matrices. *Algebra and Stochastic Methods*, 16:191–195, 1996.

[33] O. Häggström, M. N. M. Lieshout, and J. Møller. Characterisation results and markov chain monte carlo algorithms including exact simulation for some spatial point processes. Technical report, Department of Mathematics, Aalborg University, Denmark, 1996.

[34] J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*. Methuen, London, 1964.

[35] W. K. Hastings. Monte Carlo Sampling Methods using Markov Chains and their Applications. *Biometrika*, 57:97–109, 1970.

[36] D. M. Higdon. Auxiliary variable methods for Markov chain Monte Carlo with applications. Technical Report 96-17, Institute of Statistics and Decision Sciences, Duke University, 1996.

[37] A. Hurn. Difficulties in the use of auxiliary variables in Markov chains Monte carlo methods. *Statistics and Computing*, 7:35–44, 1997.

[38] C. Hwang, S. Hwang-Ma, and S. Sheu. Accelerating Gaussian diffusions. *The Annals of Applied Probability*, 3:897–913, 1993.

[39] J. G. Kemeny and J. L. Snell. *Finite Markov chains*. Princeton: Van Nostrand, 1969.

[40] W. S. Kendall. Perfect simulation for the area-interaction point process. Technical report, University of Warwick, 1996.

[41] W. S. Kendall. Perfect simulation for spatial point processes. In *Proceedings of the 51st Session of the ISI, Istambul*, 1997.

[42] W. S. Kendall and E. Thönnes. Perfect simulation in stochastic geometry. Preprint at http://www.stats.bris.ac.uk/MCMC.

[43] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. Providence, R. I.: American Mathematical Society, 1980.

[44] C. Kipnis and S. R. S. Varadhan. Central limit theorem for additive functionals of reversible markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104:1–19, 1986.

[45] A. Korányi. On a theorem by Löwner and its connections with resolvents of selfadjoint trasformations. *Acta Scientiarum Mathematicarum*, 7:63–70, 1956.

[46] T. Lindvall. *Lectures on the Coupling Method*. John Wiley and Sons, New York, 1992.

[47] J. Liu, W. H. Wong, and A. Kong. Correlation structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B*, 57:157–169, 1995.

[48] J. S. Liu. *Correlation Structure and Convergence Rate of the Gibbs Sampler*. PhD thesis, University of Chicago, 1991.

[49] J. S. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. Technical report, Stanford University, 1994.

[50] J. S. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. Technical report, Statistics Department, Stanford, 1996. To appear in Statistics and Computing.

[51] K. Löwner. Über monotone Matrixfunktionen. *Math. Zeitschrift*, 38:177–216, 1934.

[52] E. Marinari and G. Parisi. Simulated tempering: A new monte carlo scheme. *Europhysics letters*, 19:451–458, 1992.

[53] P. Matthews. A slowly mixing Markov chain with implications for the Gibbs sampler. *Stat. Prob. Lett.*, 17:231–236, 1993.

[54] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Annals of Statistics*, 24:101–121, 1996.

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

[56] J. Møller. Perfect simulation of conditionally specified models. *J. Roy. Statist. Soc. Ser. B.* To appear. Preprint at http://www.stats.bris.ac.uk/MCMC.

[57] R. D. Morris. Auxiliary variables for Markov random fields with higher order interactions. Technical report, NASA Ames Research Center, 1997.

[58] P. Mykland, L. Tierney, and B. Yu. Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90:233–241, 1995.

[59] R. M. Neal. Markov chain Monte Carlo methods based on 'slicing' the density function. Preprint at http://www.stats.bris.ac.uk/MCMC.

[60] R. M. Neal. Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation. Technical report, Dept. of Statistics, University of Toronto, 1995. No. 9508.

[61] R. M. Neal. *Bayesian Learning for Neural Networks.* Springer-Verlag, N.Y., 1996.

[62] E. Nummelin. *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge University Press, 1984.

[63] T. Ozaki. A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: A local linearization approach. *Statistica Sinica*, 2:113–135, 1992.

[64] P. H. Peskun. Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60:607–612, 1973.

[65] R. B. Potts. Some generalized order-disorder transformations. *Proc. Camb. Phil. Soc.*, 48:106–109, 1952.

[66] J. Propp and D. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252, 1996.

[67] C. R. Rao and S. K. Mitra. *Generalized Inverses of Matrices and its Applications.* John Wiley and Sons, New York, 1971.

[68] A. Reuter and V. Johnson. General strategies for assessing convergence of MCMC algorithms using coupled sample paths. *Unpublished*, 1995.

[69] B. D. Ripley. *Stochastic Simulation.* John Wiley and Sons, New York, 1987.

[70] C. Ritter and M. A. Tanner. Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87:861–868, 1992.

[71] G. O. Roberts and J. S. Rosenthal. On convergence rates of Gibbs samplers for uniform distributions. Preprint at http://www.stats.bris.ac.uk/MCMC.

[72] G. O. Roberts and J. S. Rosenthal. Shift-coupling and convergence rates of ergodic averages. *Communications in Statistics - Stochastic Models*, 13:147–165, 1994.

[73] G. O. Roberts and J. S. Rosenthal. Geometric ergodicity and hybrid Markov chains. 1997. Preprint at http://www.stats.bris.ac.uk/MCMC.

[74] G. O. Roberts and J. S. Rosenthal. Convergence of slice samper Markov chians. Technical report, University of Cambridge, 1998.

[75] G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83, 1996.

[76] W. Rudin. *Functional Analysis*. New York: McGraw-Hill. (Second ed.), 1991.

[77] I. Shoji. Approximation of continuous time stochastic processes by a local linearisation method. Technical report, The Institute of Statistical Mathematics, Tokyo, 1995.

[78] I. Shoji. Estimation for a continuous time stochastic process: a new local linearization approach. *Stochastic Analysis and Applications*, to appear.

[79] O. Stramer and R. L. Tweedie. Geometric and subgeometric convergence of diffusions with given stationary distributions. (submitted for publication).

[80] O. Stramer and R. L. Tweedie. Self-targeting candidates for Hastings-Metropolis algorithms. (submitted for publication).

[81] R. H. Swensden and J. S. Wang. Non-universal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58:86–88, 1987.

[82] H. M. Taylor and S. Karlin. *An introduction to stochastic modeling. Revised edition.* Academic Press, Inc., Boston, MA, 1994.

[83] L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1762, 1994.

[84] L. Tierney. A Note on Metropolis-Hastings kernels for general state spaces. Technical report, U. of Minnesota, 1995. No. 606.