CONVERGENCE OF MARKOV CHAIN MONTE CARLO ALGORITHMS WITH APPLICATIONS TO IMAGE RESTORATION

by

Alison L. Gibbs

A thesis submitted in conformity with the requirements for the
Degree of Doctor of Philosophy
Graduate Department of Statistics
University of Toronto

Convergence of Markov Chain Monte Carlo Algorithms with Applications to Image Restoration

Alison L. Gibbs
Department of Statistics, University of Toronto
Ph.D. Thesis, 2000

Abstract

Markov chain Monte Carlo algorithms, such as the Gibbs sampler and Metropolis-Hastings algorithm, are widely used in statistics, computer science, chemistry and physics for exploring complicated probability distributions. A critical issue for users of these algorithms is the determination of the number of iterations required so that the result will be approximately a sample from the distribution of interest.

In this thesis, we give precise bounds on the convergence time of the Gibbs sampler used in the Bayesian restoration of a degraded image. We consider convergence as measured by both the usual choice of metric, total variation distance, and the Wasserstein metric. In both cases we exploit the coupling characterisation of the metric to get our results. Our results can also be applied to the coupling-from-the-past algorithm of Propp and Wilson (1996) to get bounds on its running time.

The application of our theoretical results requires the computation of parameters of the algorithm. These computations may be prohibitively difficult in many situations. We discuss how our results can be applied in these situations through the use of auxiliary simulation to estimate these parameters.

We also give a summary of probability metrics and the relationships between them, including several new relationships.

Acknowledgements

I wish to thank Jeffrey Rosenthal, my thesis advisor, for his guidance in this project and for teaching me so much. Jeff's patience, encouragement, and, most importantly, enthusiasm are warmly appreciated.

Many thanks are due to Radford Neal for sharing his ideas, discussing my work with me in great detail, and asking many provocative questions. I would also like to thank Neal Madras for his many contributions to the improvement of this thesis and to acknowledge helpful discussions with Michael Evans and Jeremy Quastel. I wish to thank Professor Francis Su of Harvey Mudd College for sharing his understanding of probability metrics with me, and for the pleasure of working together on what we both wanted to better understand.

Thank you to Laura Kerr, Andrea Carter, Sylvia Williams, and Tom Glinos who have always been available to sort out my administrative and computing problems, and listen sympathetically to my complaints. As department graduate coordinators, Nancy Reid and Keith Knight have provided countless words of advice and encouragement.

My time here has been enlivened and enriched by sharing classes and office space with a wonderful group of fellow graduate students. In particular I thank Brenda Crowe, Nathan Taback, and Ruxandra Spijavca for their friendship and sympathetic ears.

I have had the great privilege of knowing the love and support of two wonderful people, my parents, and of experiencing much patience, support, and understanding from my husband, Stephen. Thank you.

For Isaac

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction to the Problem and Summary of Thesis

Markov chain Monte Carlo (MCMC) algorithms were first used in statistical physics and later in the statistics community for problems in spatial statistics, including image processing. See Besag, Green, Higdon and Mengersen (1995) for some history. They are now widely used, particularly in Bayesian analysis, for exploring complicated probability distributions. See, for example, Gelfand and Smith (1990), Besag and Green (1993), Smith and Roberts (1993), Besag et al. (1995), and Gilks, Richardson and Spiegelhalter (1996).

An important issue in the implementation of MCMC algorithms is whether they actually converge to the distribution of interest, and if so, how quickly. For a discussion of these issues see, for example, Tierney (1994) and Roberts and Rosenthal (1998). Convergence diagnostics do not guarantee convergence, and are known to introduce bias into the results (Cowles, Roberts and Rosenthal 1997). Much work has been done in establishing theoretical results (see Section 2.4.2) but they exist only for special cases and are often difficult to apply in practice. Exact sampling algorithms (see Propp and Wilson (1996), Fill (1998) and Section 2.4.2) which terminate with a sample distributed exactly according to the distribution of interest hold promise, particularly on finite state spaces, and recent extensions are allowing their use in some examples on continuous and unbounded state spaces.

In this thesis we find precise, *a priori* bounds on the convergence time of Markov chain Monte Carlo algorithms used in Bayesian image restoration. Our results can also be applied to bounding the convergence time of the coupling-from-the-past exact sampling algorithm of Propp and Wilson (1996). We consider convergence in both total variation distance and the Wasserstein metric. Both of these metrics have a coupling characterisation, which is used in the development of our results. Our total variation distance result is restricted to discrete state spaces and Markov chains for which a partial order exists on the state space which is preserved by the Markov chain transitions. No such restrictions are necessary for our general result for convergence in the Wasserstein metric.

In the Bayesian approach to image restoration we observe a distorted image (the data) and have a statistical model (the likelihood) for how the true image was randomly distorted to give the observed image. We also have a prior distribution on possible images. The priors we use place greater probability on images in which neighbouring pixels are similar. We wish to explore the posterior distribution for the true image. Our goal may be to use samples from the posterior to estimate the mean *a posteriori* image, or perhaps to find the posterior mode to use as our restored image. Our distributions are very high-dimensional (the dimension being the number of pixels) and the normalising constants are often intractable. However, because of the spatial structure in the prior, they are well-suited to Markov chain Monte Carlo. See Geman and Geman (1984) and Besag (1986) for early descriptions of this approach to image restoration and Green (1996) for a more recent discussion.

For a binary image using an Ising model prior, we obtain bounds on the convergence time of the random scan Gibbs sampler algorithm that are $O(N^2)$, where N is the number of pixels, with a computationally simple constant of proportionality. These bounds hold for small values of the prior parameter. Convergence is measured in total variation distance and at each iteration only one randomly chosen pixel is updated. We provide results for the case when the distribution of interest is the prior, which is of interest in its own right in statistical physics, and for two random distortion mechanisms: additive normal noise, and the incorrect observation of each pixel with a fixed probability. These results are presented in Chapter 3. While it is known in the statistical physics literature that the convergence time of these algorithms is $O(N \log N)$ for appropriate values of the parameters which include those for which our results hold, the proportionality constant for those results is intractable (see, for example, Martinelli (1997)). In Chapter 4 we develop precise $O(N \log N)$ bounds for the convergence time of these algorithms by another method.

In Chapter 4 we introduce a method for bounding the time necessary to achieve convergence in the Wasserstein metric. We apply this method to the restoration of a grey-scale image where each pixel takes on a value in the interval $[0, 1]$ and we achieve computationally simple results that are

$O(N \log N)$. Again we use the random scan Gibbs sampler. Simulations show that our results are reasonably tight. Moreover, because a simple bound exists on total variation distance in terms of the Wasserstein metric on finite state spaces, we are able to use the method developed in this chapter to improve the results of Chapter 3.

The methods described above require the analytic calculation of parameters of the Markov chains relating the distance between two coupled realisations of the chain to the distance at the previous iteration. In most applications, these constants will be difficult or impossible to calculate. In Chapter 5 we demonstrate how auxiliary simulations can be used to get reasonable approximations for these parameters. For comparison, we use auxiliary simulation to approximate the grey-scale result in Chapter 4. This approach is seen as a compromise between the guarantees of our theoretical results, and the uncertainty associated with the use of convergence diagnostics.

Total variation distance is the usual metric used to quantify convergence of MCMC algorithms to their stationary distributions. However, its coupling characterisation requires exact coupling, which may not be practical on continuous state spaces, or may require an algorithm that is more difficult to theoretically analyse. This was our motivation for using the Wasserstein metric which only requires coupling to within a tolerance ϵ; we wanted to extend our results of Chapter 3 to an image of continuous grey-scale pixels.

The literature on probability metrics is vast, and different applications use different metrics to suit the necessary calculations. In researching the choice of metric, we found that no concise, straightforward summary of metrics and their relationships exists. Chapter 6 is an attempt to fill this gap. We have selected nine popular choices of distance between probability measures, which are popular either because of their theoretical properties, or their practical uses. We have summarised all known relationships between them. Proofs are given for relationships and extensions to known relationships that are not known to exist elsewhere. It is hoped that the material in this chapter will be useful to practitioners considering the choice of metric, especially if interest lies in using one metric to get results in another.

Chapter 7 summarises some ideas for future work and extensions of the ideas in this thesis.

In Chapter 2 we outline the construction of MCMC algorithms and some relevant Markov chain theory.

Chapter 2

Markov Chain Monte Carlo Algorithms

2.1 Introduction

Suppose we have a probability distribution $\pi(\cdot)$ on a state space $\mathcal{X}$. In applications in which MCMC is necessary, $\pi(\cdot)$ is typically very high-dimensional or so complicated that standard techniques such as numerical integration with respect to $\pi(\cdot)$ or direct sampling from $\pi(\cdot)$ are not suitable. In many applications in Bayesian statistics and in statistical mechanics, the normalising constant for $\pi(\cdot)$ can not be computed.

MCMC proceeds by constructing a discrete-time Markov chain $\{X_t\}$ such that $\pi(\cdot)$ is the unique limiting distribution. Let $P(x,\cdot)$ represent the transition kernel for this Markov chain, i.e. for each $x \in \mathcal{X}$ and $A \subseteq \mathcal{X}$, $P(x,A)$ represents the probability of jumping from x to somewhere in A. $P^t(x,A)$ represents the probability that the Markov chain in state x is somewhere in A after t iterations. We need to construct this transition kernel such that

$$P^t(x,A) \to \pi(A) \text{ as } t \to \infty, \text{ for all initial states } x.$$

2.2 Some Markov Chain Theory

The following Markov chain results can be found in Billingsley (1986) or Feller (1968) for discrete state spaces, and Meyn and Tweedie (1993) for general state spaces. See Tierney (1994) and Smith and Roberts (1993) for summaries in the particular context of Markov chain Monte Carlo.

The distribution π is stationary with respect to a transition kernel P if for $\mathcal{X}$ discrete:

$$\sum_{x\in\mathcal{X}} \pi(x)P(x,y) = \pi(y) \text{ for all } y \in \mathcal{X}$$

or for $\mathcal{X}$ continuous:

$$\int_{x\in\mathcal{X}} \pi(dx)P(x,\cdot) = \pi(\cdot).$$

In practice, it is often easier to verify the following condition, known as *reversibility* or *detailed balance*:

$$\pi(dx)P(x,dy) = \pi(dy)P(y,dx).$$

If reversibility holds for a distribution π with respect to P then it is easily seen that π is stationary for P by integrating both sides with respect to x.

A Markov chain on a discrete state space is *irreducible* if it is possible to eventually get from every state to every other, i.e. for every pair of states $x, y \in \mathcal{X}$ there exists a positive integer k such that $P^k(x,y) > 0$. It is *aperiodic* if $\gcd\{k > 0 : P^k(x,x) > 0\} = 1$ for every $x \in \mathcal{X}$. A Markov chain on a discrete state space with stationary distribution π will have π as its unique limiting distribution if it is both irreducible and aperiodic (see, for example, Billingsley (1986, Theorem 8.6)).

For general state spaces, we have the following analogues of irreducibility and aperiodicity (see, for example, Rosenthal (1999)). Let τ_A be the time of the first visit to A for any $A \subseteq \mathcal{X}$, i.e.

$$\tau_A = \min\{t : X_t \in A\}.$$

If $\{t : X_t \in A\}$ is empty, set $\tau_A = \infty$. A Markov chain is *ϕ-irreducible* if there exists a non-zero probability measure ϕ on $\mathcal{X}$ such that for any $A \subseteq \mathcal{X}$

with $\phi(A) > 0$, we have $\Pr(\tau_A < \infty | X_0 = x) > 0$ for all $x \in \mathcal{X}$, i.e. any set of positive ϕ measure has positive probability of being hit from any starting point x. Such a measure ϕ is called an *irreducibility measure.* It is *aperiodic* if there does not exist a partition of the state space $\mathcal{X} = \mathcal{X}_1 \dot{\cup} \mathcal{X}_2 \dot{\cup} \ldots \dot{\cup} \mathcal{X}_d$, where $\dot{\cup}$ indicates disjoint union, such that $P(x, \mathcal{X}_{i+1 \bmod d}) = 1$ for all $x \in \mathcal{X}_i$. Such a cyclic partition, if it exists, is unique up to sets of measure 0. We consider convergence of the Markov chain in total variation distance[1]. The total variation distance between two probability measures μ, ν on a space $\mathcal{X}$ is

$$d_{TV}(\mu, \nu) = \sup_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)|.$$

Total variation distance has the following equivalent formulation

$$d_{TV}(\mu, \nu) = \frac{1}{2} \max_{|h| \leq 1} \left| \int h \, d\mu - \int h \, d\nu \right| \tag{2.1}$$

where $h : \mathcal{X} \to \mathbb{R}$ satisfies $|h(x)| \leq 1$. If the state space $\mathcal{X}$ is countable,

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mu(x) - \nu(x)|.$$

For our Markov chains with initial state x, transition matrix P and stationary distribution π, we are interested in how close the distribution of states at time

[1]Note that some authors (for example, Tierney (1996)) define total variation distance as twice our value.

t is to the stationary distribution, i.e.

$$d_{TV}(P^t(x,\cdot),\pi(\cdot)) = \sup_{A\subseteq\mathcal{X}} |P^t(x,A)-\pi(A)|.$$

Proposition 2.1 *$d_{TV}(P^t(x,\cdot),\pi(\cdot))$ is a non-increasing function of t.*

Proof

$$\begin{aligned} d_{TV}(P^{t+1}(x,\cdot),\pi(\cdot)) &= d_{TV}\left(\int P^t(x,dy)P(y,\cdot)\,,\ \pi(\cdot)\right) \\ &= \sup_{A\subseteq\mathcal{X}} \left| \int P^t(x,dy)P(y,A) - \int \pi(dy)P(y,A) \right| \\ &= \sup_{A\subseteq\mathcal{X}} \left| \mathbf{E}_{P^t(x,\cdot)}[P(\cdot,A)] - \mathbf{E}_{\pi(\cdot)}[P(\cdot,A)] \right|. \end{aligned}$$

$h(\cdot) = P(\cdot,A)$ is a function satisfying $0\le h\le 1$ so using the formulation of total variation distance (2.1) gives

$$d_{TV}(P^t(x,\cdot),\pi(\cdot)) \le d_{TV}(P^t(x,\cdot),\pi(\cdot)) \qquad \blacksquare$$

Rosenthal (1999) proves the following theorem, which is also available in Meyn and Tweedie (1993).

Theorem 2.1 *Let $P(x,dy)$ be the transition probabilities for a Markov chain on a general state space $\mathcal{X}$. Suppose there exists a non-zero probability measure ϕ such that the Markov chain is ϕ-irreducible, and also suppose that the Markov chain is aperiodic and has stationary distribution π. Then for*

$\pi-$almost every $x \in \mathcal{X}$, we have

$$\lim_{t\to\infty} d_{TV}(P^t(x,\cdot),\pi(\cdot)) = 0.$$

To have the result in the above theorem hold for all initial states x, we require the stronger condition of Harris recurrence. A Markov chain is *Harris recurrent* if there exists a non-zero measure ϕ on $\mathcal{X}$ such that if $\phi(A) > 0$, then

$$\Pr(\tau_A < \infty | X_0 = x) = 1 \text{ for all } x \in \mathcal{X}.$$

Often, the goal of Markov chain Monte Carlo is to generate samples from π in order to estimate $\mathbf{E}_\pi[g(x)]$ by $\frac{1}{T}\sum_{i=1}^{T} g(x_i)$ where $x_i \sim \pi,\ i = 1,\ldots,T$. The *ergodic theorem* (Meyn and Tweedie 1993, Theorem 17.1.7) confirms that this is an asymptotically consistent estimator, despite the lack of independence in the Markov chain samples.

Theorem 2.2 *If $\{X_t\}$ is a Harris recurrent Markov chain with transition kernel P and stationary distribution π, and g is a real-valued function with $\int |g(x)|\pi(dx) < \infty$, then*

$$\frac{1}{N}\sum_{t=1}^{N} g(X_t) \to \int g(x)\pi(dx)$$

almost surely.

2.3 Constructing Markov Chains with the Required Stationary Distribution

In this section, we will assume that our distribution of interest π has a density with respect to a dominating measure (usually the Lebesgue measure). We will denote this density also by π.

2.3.1 The Metropolis-Hastings Algorithm

Choose a *proposal density* $q(y|x)$. At each step, propose a new state y from q given the current state x. Accept the new state and move to it with probability

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

or reject it and stay at the same state with probability $1-\alpha(x,y)$. If $\pi(x)\,q(y|x) = 0$, set $\alpha(x,y) = 1$. It is easily seen that this Markov chain is reversible with respect to π. For example in the discrete case, if $x \neq y$,

$$\begin{aligned}
\pi(x)\,P(x,y) &= \pi(x)\,q(y|x)\,\min\left(1, \frac{\pi(y)\,q(x|y)}{\pi(x)\,q(y|x)}\right) \\
&= \min\left(\pi(x)\,q(y|x),\ \pi(y)\,q(x|y)\right) \\
&= \pi(y)\,P(y,x).
\end{aligned}$$

So π is a stationary distribution. The following results are in Tierney (1994). Consider $q(y|x)$ as the density for a Markov chain. It is necessary for it to be a (ϕ-)irreducible Markov chain for the resulting Metropolis-Hastings chain also to be (ϕ-)irreducible. Harris recurrence is often achieved for Metropolis-Hastings algorithms because a Markov chain is Harris recurrent if $P(x, \cdot)$ is absolutely continuous with respect to its stationary distribution $\pi(\cdot)$ for all starting points x, or if π is an irreducibility measure (Tierney 1994).

The special case where q is symmetric, i.e. $q(y|x) = q(x|y)$ is called the Metropolis algorithm after the work by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953). It was generalised by Hastings (1970).

2.3.2 The Single-Component Metropolis-Hastings Algorithm

Suppose $\pi(\cdot)$ is the joint distribution of $x = (x_1, x_2, \ldots, x_N)$. Sometimes it is computationally simpler to update only one component of x at each iteration. Suppose at iteration $t+1$ component i is being updated. The proposal distribution is the univariate distribution with density $q(y_i|x_i, x_{-i})$ where $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. The proposed value for component

i is accepted with probability

$$\alpha_{x_{-i}}(x_i, y_i) = \min\left(1,\ \frac{\pi(y_i|x_{-i})\, q(x_i|y_i, x_{-i})}{\pi(x_i|x_{-i})\, q(y_i|x_i, x_{-i})}\right).$$

The remaining components are not changed at iteration $t+1$.

The components can be updated in a systematic or random order.

2.3.3 The Gibbs Sampler

Again suppose $\pi(\cdot)$ is the joint distribution of $(x_1, x_2, \ldots, x_N)$. Each component is updated according to its conditional distribution given the current value of each of the other components

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

These distributions are called the *full conditionals.* The components can either be updated in random order or in a systematic order. The random scan version is easily shown to be reversible. While the systematic scan version is not reversible, π is a stationary distribution for the resulting Markov chain (one iteration comprising one sweep of the components) since it is a stationary distribution for the update of each individual component.

The Gibbs sampler is a special case of the single-component Metropolis-Hastings algorithm where the proposals q are the full conditionals and the acceptance probability is always one.

The Gibbs sampler was given its name in Geman and Geman (1984) where it was used in Bayesian image restoration. Gelfand and Smith (1990) extended its application to continuous state spaces and showed how it can be used in Bayesian inference problems.

2.4 Convergence Issues

2.4.1 Qualitative Convergence

A Markov chain is *geometrically ergodic* if

$$d_{TV}(P^t(x,\cdot),\, \pi(\cdot)) \leq M(x)\rho^t$$

for some finite $M(x)$, and constant $\rho < 1$. It is *uniformly ergodic* if for all x

$$d_{TV}(P^t(x,\cdot),\, \pi(\cdot)) \leq M\rho^t.$$

A set $C \subseteq \mathcal{X}$ is *small* if there exists a probability measure ϕ and constants $\epsilon < 1$ and positive integer k such that

$$P^k(x,\cdot) \geq \epsilon\phi(\cdot) \text{ for all } x \in C.$$

A Markov chain is geometrically ergodic if and only if it satisfies a *geometric drift condition*, i.e. there is a small set C and constants $\lambda < 1$ and $b < \infty$

and a π-almost everywhere finite function $V : \mathcal{X} \to [1, \infty]$ such that

$$\int V(y)P(x, dy) \leq \lambda V(x) + b\mathbf{1}_C(x)$$

(see Meyn and Tweedie (1993, Chapter 15)).

As well as being likely to converge reasonably quickly in practice, geometrically ergodic chains are useful because a Central Limit Theorem exists for averages of functions of their output (see Chan and Geyer in the discussion to Tierney (1994)).

If the entire state space is small, then the Markov chain is uniformly ergodic. This doesn't often occur in statistical models with unbounded state spaces, but is necessary for the coupling-from-the-past algorithm described in the next section to work (Foss and Tweedie 1998).

2.4.2 Quantitative Convergence

We now turn our interest to the critical question of how many iterations of the Markov chain are necessary in order to be able to consider the output to be a sample from the stationary distribution. There exist three approaches to answering this question: convergence diagnostics, theoretical results, and exact simulation.

Convergence Diagnostics

Convergence diagnostics are methods of monitoring the convergence of the algorithm while it is running by considering statistical functions of the output of a single chain or of multiple runs of the same chain. There exist many such procedures (see Cowles and Carlin (1996) and Brooks and Roberts (1997) for reviews) but none are completely satisfactory. All convergence diagnostics are known to sometimes prematurely claim convergence (Cowles and Carlin 1996) and can introduce bias into the results (Cowles et al. 1997, Roberts and Rosenthal 1998).

Theoretical Results

There has been much work on developing rigorous, *a priori*, quantitative bounds on the convergence time (for example, Sinclair and Jerrum (1989), Diaconis and Stroock (1991), Frieze, Kannan and Polson (1994), Ingrassia (1994), Meyn and Tweedie (1993), Rosenthal (1995b), Mengerson and Tweedie (1996), Polson (1996), and Frigessi, Martinelli and Stander (1997)). However, these results exist for specific problems and may not be generalisable, they require extensive and complicated calculations, and the upper bounds they provide on the convergence time are often overly conservative.

Moreover, for some of these results, the order of convergence is known, but the proportionality constant is not available.

Developing theoretically justifiable convergence rates for problems in Bayesian image restoration is the subject of most of this thesis.

Exact Sampling

Recently, the development of algorithms that produce samples distributed exactly according to the distribution of interest (Propp and Wilson 1996, Fill 1998) have generated a great deal of interest. Because we will later describe how our results in Chapters 3 and 4 can be used to bound the running time of the coupling-from-the-past (CFTP) algorithm of Propp and Wilson (1996), we give a brief description of the algorithm here.

CFTP is a method of organising a Markov chain simulation so that it delivers exactly a sample from the distribution of interest. The number of steps necessary is random and determined by the algorithm as it runs. Suppose we could start the Markov chain in every state at time $-\infty$. Then if all realisations of the chain starting at time $-\infty$ have the same state at time 0, we have lost all dependence on the initial state and this common state must be a sample from π. In practice, if there exists an initial time $-T$

such that for all initial states X_{-T}, X_0 is the same, then $X_0 \sim \pi$. And we don't need to find T exactly since coalescence occurs from all initial times less than $-T$ if it occurs from $-T$. Propp and Wilson suggest the doubling strategy of starting at time -2, and if coalescence is not achieved, next start at time -4 and then -8, etc. This is valid as long as the uniform random number used at each time point remains constant.

If the state space is large or infinite it may be difficult or impossible to keep track of Markov chains started in each possible state. This difficulty is overcome for monotone Markov chains such as those considered in the applications in this thesis. In our examples, there exists a partial ordering on the state space with unique maximal and minimal elements. Moreover, the Markov chain transitions preserve this order. Thus it is only necessary to achieve coalescence of the chains started in these maximal and minimal states, as this guarantees coalescence from all initial states.

More formally, suppose we have a sequence of independent Uniform[0,1] random variables, $\xi_i, i = -\infty \ldots \infty$ and we can find a deterministic function $f : \mathcal{X} \times [0,1] \rightarrow \mathcal{X}$ such that the value of the Markov chain at time $t+1$ is determined as

$$X_{t+1} = f(X_t, \xi_t). \tag{2.2}$$

If f is monotone in the first variable, chains begun in higher starting points will stay above chains begun in lower starting points. It follows that

$$x \leq y \implies P(x, [z, \infty)) \leq P(y, [z, \infty))$$

for all $z \in \mathcal{X}$. Suppose X_t^{max} and X_t^{min} are the states at time t for the chains started in the maximal and minimal states, respectively. Updating according to (2.2) ensures that a chain, X_t^0, started in any other initial state will be sandwiched between the maximal and minimal chains, i.e.

$$X_t^{min} \leq X_t^0 \leq X_t^{max}.$$

Thus coalescence of chains started in the maximal and minimal states is sufficient for coalescence of chains started in all possible points.

The extension of CFTP to infinite discrete state spaces and continuous state spaces in special cases has been considered and applied by, for example, Foss and Tweedie (1998), Green and Murdoch (1998), Murdoch and Green (1998), Corcoran and Tweedie (1998), Møller (1999), Murdoch (1999), Møller and Nicholls (1999), and Guglielmi, Holmes and Walker (1999).

2.5 The Coupling Method

The coupling method exploits the construction of a joint distribution with given marginals to prove things about the marginal distributions. For a detailed discussion of the coupling method see Lindvall (1992). In our examples, we consider coupled Markov chains on the state space $\mathcal{X} \times \mathcal{X}$. The marginal distributions are the distributions of Markov chains with different initial states, but both following the transitions of the original Markov chain. Our coupled chains will not proceed independently; we will use the same uniform random number to determine their transitions at each step. This dependence is necessary for our construction of Markov chains which are monotone.

Lindvall (1992) provides many applications of the coupling method, including its application to providing estimates of convergence in total variation distance for Markov chains. Some other applications of coupling that are relevant to Markov chain Monte Carlo include the proof in Rosenthal (1999) of Theorem 2.1, and the convergence results of, for example, Rosenthal (1995b) and Luby, Randall and Sinclair (1995).

Chapter 3

Total Variation Distance Bound for a Binary Image

3.1 Introduction

In this chapter, we show how coupling methodology can be used to give precise, *a priori* bounds on the convergence time in total variation distance of Markov chain Monte Carlo algorithms. Our results hold for monotone Markov chains, for which a partial order exists on the state space which is preserved by the Markov chain transitions. In particular, we develop convergence time bounds for a simplified problem in Bayesian image restoration which involves sampling from a Gibbs distribution using the Gibbs sampler. The case of image synthesis, where there is no observed data, is equivalent to what is referred to in the mathematical physics literature as Glauber dynamics for the stochastic Ising model.

We use coupling and martingale techniques to obtain precise upper bounds on the convergence time in total variation distance for the random scan version of the Gibbs sampler, where each iteration involves the update of only one randomly chosen pixel. For appropriate values of the prior parameter, our bounds are an easily computable constant times N^2, where N is the number of pixels. While we believe that similar arguments will lead to a similar bound on the convergence time for the systematic scan algorithm, the fact that the values of neighbouring pixels may change at each iteration makes analysis of the systematic scan algorithm more difficult. The general methodology outlined in Section 3.3.1 can be applied to any monotone Markov chain Monte Carlo algorithm. In Chapter 4 we show how the calculations of this chapter can be applied to achieve precise bounds that are $O(N \log N)$.

In the mathematical physics literature, it is well known that the convergence rate for the stochastic Ising model is $O(N \log N)$ for appropriate values of the parameters which include those for which our results hold. (See, for example, Frigessi et al. (1997).) However the constant of proportionality is not known.

Our results are presented as follows. The model and Gibbs sampler algorithm are described in Section 3.2. The coupling methodology used to derive our bounds is described in Section 3.3.1. The application of these bounds to the running time of the coupling-from-the-past algorithm of Propp and Wilson (1996) is discussed in Section 3.6. Results for sampling from the Ising model without data and from the posterior distribution with data are presented in Sections 3.4 and 3.5 respectively.

3.2 Image Restoration using the Gibbs Sampler

3.2.1 The Model

We consider the Bayesian restoration of images where the prior consists of a probability model for the true image and the posterior is formed from the prior conditional on the data, which in our cases are the values of the observed image. These observed data are obtained from the true image through a known random distortion process. See Geman and Geman (1984) and Besag (1986) for early descriptions of this approach to image restoration and Green's article in Gilks et al. (1996) for a more recent discussion. The random scan Gibbs sampler is used to produce samples from the posterior distribution. We also consider the use of the Gibbs sampler for simulation of the prior distribution since this is of interest on its own.

Our model of the image is a Markov random field of pixels taking values in $\{+1, -1\}$, with the value of each pixel affected by its nearest neighbours in an attractive manner. Equivalently, it is modelled by the Gibbs distribution

$$\pi(x) = \frac{1}{Z} \exp\{-U(x)\} \tag{3.1}$$

where $x = (x_1, \ldots, x_N)$ is a configuration of the colours at the N pixels, the energy function U reflects the neighbourhood structure within which configurations with pixels having like neighbours are favoured, and Z is the normalising constant, called the partition function in mathematical physics. The particular prior probability model we place on the configuration is the Ising model for which

$$U(x) = -\beta \sum_{<i,j>} x_i x_j \tag{3.2}$$

where the sum is taken over pairs of sites (i, j) which are nearest neighbours, and β is a positive parameter. For a discussion of the physical significance of the Ising model, see Cipra (1987). Conditioning on the value of observed data results in a posterior distribution that is equivalent to a Gibbs distribution with the presence of an external field. With the Ising model prior, the posterior distribution of x given the data, y, is of the form

$$\pi_{posterior}(x|y) = \frac{1}{Z_y} \exp\left\{\beta \sum_{<i,j>} x_i x_j + \sum_i f(x_i, y_i)\right\} \tag{3.3}$$

where Z_y is the normalising constant which is a function of the data, and the function f changes with the random distortion mechanism.

Our data are an observed distortion of the true image. We consider two distortion mechanisms. In Section 3.5.1 we consider y to be obtained from the true image by switching, with a constant probability, the sign of each pixel and in Section 3.5.2 we consider the case where independent normal noise is added to the value of each pixel. Examples of the form of the function f in (3.3) are available in Equations (3.21) and (3.24).

Even for the simple models studied here, examining both the prior and posterior distributions by calculating the probability of each configuration is impractical because of the large configuration space. For example, a grid of 64×64 pixels has 2^{4096} configurations.

The Gibbs sampler is used to produce a sample from the distribution of interest. Pixels are updated according to their conditional distribution given the value of all of the other pixels. At each iteration one randomly chosen pixel is updated. The algorithm is outlined in Section 3.2.2.

3.2.2 The Algorithm

Our goal in the Bayesian image restoration process is to produce samples from the posterior distribution of the image. These samples can be used to explore the posterior distribution, with goals such as finding its mode(s), or calculating expectations. We use the single site random scan Gibbs sampler to obtain these random samples. We also consider the application of the algorithm to the case without data; we are then sampling from the Ising model prior distribution.

For our problem, the full conditional probabilities are easy to calculate and to sample from. They do not require the calculation of the normalising constant and depend only on the current values of the nearest neighbours. In the case of sampling from the posterior conditional on the data, the full conditional for each pixel depends on the value of the observed image at that site, and no other observed pixels. For the case of no data where the distribution of interest is the Ising model (3.1) and (3.2), the full conditionals are

$$\pi_{FC}(x_i|x_{-i}) = \frac{e^{\beta \sum_{j\sim i} x_i x_j}}{e^{\beta \sum_{j\sim i} x_j} + e^{-\beta \sum_{j\sim i} x_j}} \tag{3.4}$$

where the sum is taken over pixels j that are neighbours of pixel i. For the case with data, the full conditionals are

$$\pi_{FC}(x_i|x_{-i}, y) = \frac{e^{\beta\left(\sum_{j\sim i} x_i x_j\right) + f(x_i, y_i)}}{e^{\beta\left(\sum_{j\sim i} x_j\right) + f(1, y_i)} + e^{-\beta\left(\sum_{j\sim i} x_j\right) + f(-1, y_i)}} \tag{3.5}$$

where the function f comes from the random distortion mechanism that creates the observed image.

The iterations continue until the current configuration can be considered to be a sample from the posterior distribution, independent of the initial configuration. We are concerned with the number of iterations required.

The Markov chain whose state space is the space of all possible configurations and whose transition probabilities are $\frac{1}{N}$ times the full conditional probability, with transitions only possible between configurations which differ at only one site, is an irreducible, aperiodic Markov chain with stationary distribution π.

3.3 Bounding the Convergence Time of the Algorithm

Convergence is measured by the total variation distance (tvd), which is the usual metric chosen to assess convergence of MCMC algorithms. For a Markov chain with probability transition matrix P, stationary distribution π, countable state space $\mathcal{X}$, and initial configuration $x^0 \in \mathcal{X}$, the total variation distance at time t is

$$\begin{aligned}
\mathrm{tvd}_{x^0}(t) &= d_{TV}\left(P^t(x^0,\cdot),\pi(\cdot)\right) \\
&= \sup_{A\subset\mathcal{X}} \left|P^t(x^0,A)-\pi(A)\right| \qquad (3.6)\\
&= \frac{1}{2}\sum_{x\in\mathcal{X}} \left|P^t(x^0,x)-\pi(x)\right| \qquad (3.7)
\end{aligned}$$

where $P^t(x^0,x)$ is the probability that the Markov chain with initial state x^0 is in state x at iteration t, and A is any set. As shown in Proposition 2.1, $\mathrm{tvd}_{x^0}(t)$ is non-increasing in t. The convergence time of the Markov chain used by the Gibbs sampler is defined as

$$\tau(\epsilon) = \max_{x^0} \min\{t : \mathrm{tvd}_{x^0}(t') \le \epsilon \text{ for all } t' \ge t\} \qquad (3.8)$$

where ϵ is a pre-specified error tolerance, chosen at the user's discretion. Propp and Wilson (1998) use the arbitrary value $1/e$ as the value of ϵ which gives their mixing time threshold. The first definition of the total variation distance (3.6) leads to perhaps the clearest interpretation of the choice of ϵ: for every possible set A in the state space, convergence to within ϵ in total variation distance guarantees that the difference between the probability that our Markov chain is in A and the probability of A for the stationary distribution is at most ϵ. The relationship between the value of ϵ and the number of iterations required is further explored through simulation of the stochastic Ising model in Section 3.4.

Requiring that the total variation distance is less than ϵ gives an immediate tolerance on the error due to lack of convergence in the estimation of the expectation of bounded functions. This is because of the following equivalent formulation of tvd

$$\mathrm{tvd}_{x^0}(t) \;=\; \frac{1}{2}\max_{|h|\leq 1}\left|\int_{\mathcal{X}} h\, P^t(x^0, dx) - \int_{\mathcal{X}} h\, \pi(dx)\right|$$

where the maximum is taken over functions $h : \mathcal{X} \to \mathbb{R}$ satisfying $\sup_x |h(x)| \leq 1$.

We are concerned with the number of iterations required to achieve convergence for a given algorithm, and not with other important issues such as the variance of estimates of expectations (see, for example, Green and Han (1992)). We recommend that our results be used to determine the number of iterations required to achieve stationarity. The simulation of the Markov chain can then be continued beyond this and these additional values used for purposes such as estimating expectations.

Our results are an application of coupled Markov chains. The coupling methodology is presented in Section 3.3.1. In Section 3.6 we discuss how these results can be applied to exact sampling algorithms involving coupling-from-the-past. Other methods for achieving a bound on (3.8) are discussed in Section 3.3.2.

3.3.1 Using Coupling to Bound the Convergence Time

One method of bounding the convergence time of a Markov chain, $\tau(\epsilon)$, is through monitoring two coupled Markov chains. Suppose X_t^1 and X_t^2 are two Markov chains on the same state space, with the same transition probabilities, and with initial values x^1 and x^2 respectively. At each iteration, the same uniform random number is used to determine the transition for both chains. They are said to be coupled at time T^{x^1,x^2} if

$$T^{x^1,x^2} = \min\{t : X_t^1 = X_t^2 | X_0^1 = x^1, X_0^2 = x^2\}. \qquad (3.9)$$

Our bound on $\tau(\epsilon)$ will be in terms of the maximum mean coupling time

$$T = \max_{x^1,x^2} \mathbf{E}(T^{x^1,x^2}) \tag{3.10}$$

where the maximum is taken over all possible initial states x^1 for X_t^1 and x^2 for X_t^2.

As shown in Aldous (1983), the following relationship exists between the mean coupling time and the convergence time:

$$\tau(\epsilon) \leq 2eT(1 + \log \epsilon^{-1}). \tag{3.11}$$

The method used here was inspired by that of Luby et al. (1995) whose Markov chains were lattice routings in order to generate a random tiling of a planar lattice structure in studying the combinatorics of tiling two-dimensional lattices. They use coupling to get bounds on the convergence time of their Markov chains that are polynomial in the size of the lattice.

For our model for binary images, a partial ordering exists on the set of all configurations. One configuration is greater than another if each pixel of the larger configuration is greater than or equal to the corresponding pixel of the smaller configuration. We set the initial configurations of the two chains to be all $+1$ and all -1. We label these configurations x^{max} and x^{min} respectively. Our process will preserve this order; the chain that starts in the maximal state will always be greater than or equal to the chain that starts in the minimal state. This is because at each iteration our algorithm will use the same random number to determine the transition for both chains, and the update function is a deterministic function of this random number, and a monotone function of the current state. In particular, suppose at iteration t site i has been chosen for updating and ξ_t is the Uniform[0,1] random number to be used used for updating at that iteration. If the value of the full conditionals (Equations (3.4) or (3.5)) evaluated at $x_i = +1$ are less than or equal to ξ_t, we set the value of pixel i at time $t+1$ as +1. The full conditionals place greater probability on configurations in which pixels are like their neighbours, and the chain started in the maximal state will have at least as many neighbours of pixel i that are +1 as the chain started in the minimal state. Thus the value of pixel i in the chain started in the maximal state will always be greater than or equal to its value in the chain started in the minimal state.

As argued in Propp and Wilson (1996) for monotone Markov chains such as this, it suffices to consider the case where the initial configurations are the extreme states. Chains started in any other initial states x^1, x^2 ($x^{min} \leq x^1 \leq x^2 \leq x^{max}$) must couple in a time less than or equal to the coupling time for x^{min} and x^{max} for the same set of random numbers determining the transitions.

Let $\Phi(t)$ be a function that assigns a positive integer to the difference between the configurations at time t of the Markov chains started in the maximal and minimal states. Φ should be defined such that $\Phi(0)$ is N, the number of sites, and $0 \leq \Phi(t) \leq N$ for all t. Two chains will have coupled at time t if $\Phi(t) = 0$. Once coupled, they will remain so. Define the coupling time

$$T^{x^{max},x^{min}} = \inf\{t : \Phi(t) = 0\}.$$

Then

$$T = \mathbf{E}(T^{x^{max},x^{min}}). \tag{3.12}$$

Let $\Delta\Phi(t) = \Phi(t+1) - \Phi(t)$ denote the change in the value of Φ after one iteration of the random scan Gibbs sampler. Suppose a region of the parameter space for β can be found such that $\mathbf{E}\{\Delta\Phi(t)|X_t^1, X_t^2\} < 0$ for all t for which $X_t^1 \neq X_t^2$, say $\mathbf{E}\{\Delta\Phi(t)|X_t^1, X_t^2\} \leq -a_\beta$ where $-1 < -a_\beta < 0$. Then for these values of β, as shown in the proof to Theorem 3.1, the quantity $\mathbf{E}(T^{x^1,x^2})$ can be bounded above by Na_β^{-1}.

Theorem 3.1 *Suppose there exist two coupled realisations, X_t^1, X_t^2 of a Markov chain where $X_0^1 = x^1$ and $X_0^2 = x^2$. And suppose a constant $a > 0$*

can be found such that $\mathbf{E}\{\Delta\Phi(t)|X_t^1, X_t^2\} < -a$ *for all* t *for which* $X_t^1 \neq X_t^2$*, where* $\Delta\Phi(t)$ *is the change in distance between the two Markov chains from iteration* t *to* $t+1$ *and the distance between the initial states is* $\Phi(0) = N$*. Then the following bound exists on the mean coupling time (3.9)*

$$\mathbf{E}(T^{x^1,x^2}) \leq \frac{N}{a}. \tag{3.13}$$

Proof Define the stochastic process $Z_t = \Phi(t) + at$. Z_t is a supermartingale up to time T^{x^1,x^2} since

$$\begin{aligned} &\mathbf{E}(Z_{t+1}|X_t^1, X_t^2) - Z_t \\ &= \mathbf{E}(\Delta\Phi|X_t^1, X_t^2) + a \\ &\leq 0. \end{aligned}$$

T^{x^1,x^2} is a stopping time. Since Z_t is nonnegative we can apply the Optional Stopping Theorem (see, for example, Durrett (1996, Theorem 7.6, p. 274)) giving

$$\mathbf{E}(Z_{T^{x^1,x^2}}) \leq \mathbf{E}(Z_0)$$

and since

$$Z_{T^{x^1,x^2}} = \Phi(T^{x^1,x^2}) + aT^{x^1,x^2} = aT^{x^1,x^2}$$

we have

$$a\,\mathbf{E}(T^{x^1,x^2}) \leq N. \quad \blacksquare$$

In our examples, a is of the form $\frac{f(\beta)}{N}$ where, as will be seen, $f(\beta)$ is straightforward to compute, as is the range of possible values of β which guarantees that the distance function is decreasing on average. Combining (3.13) with (3.11) gives the form of our results

$$\tau_\beta(\epsilon) \leq \frac{2eN^2}{f(\beta)}(1 + \log \epsilon^{-1}).$$

We have introduced the subscript β to make explicit τ's dependence on the value of the model parameter.

3.3.2 Other Convergence Results for this and Related Models

In mathematical physics, the case without data is known as the stochastic Ising model with Glauber dynamics and it is well-known that its convergence rate, asymptotically in N, is $O(N \log N)$. In dimensions higher than one, this result holds for values of β below a critical value at which a phase transition occurs. Madras and Piccioni (1999) use Dobrushin's criterion to bound the spectral gap of the Markov chain transition matrix, and show that for small values of β the chain converges at a different rate than for larger values of β for which they show it is slowly mixing. For the Ising model with an external field, the convergence rate is known to be $O(N \log N)$ for all β in two dimensions, and for small enough β and large enough external field in higher dimensions. (See, for example, Martinelli (1997).) These results use the log Sobolev inequality and it may be impossible to calculate a precise upper bound using this method, so these results are difficult to apply in practice. While our results are $O(N^2)$, we are able to give the proportionality constant. Frigessi et al. (1997) present the $O(N \log N)$ results in the context of Bayesian image restoration. In Chapter 4 we describe how the calculations of this chapter can be used to get a bound that is $O(N \log N)$.

The total variation distance can also be bounded above by a simple function of the eigenvalue of the Markov chain transition matrix which is second largest in absolute value. Poincaré and Cheeger inequalities can be used to get simple bounds on this eigenvalue in terms of a set of canonical paths on a graph associated with the Markov chain. The vertices of the graph are the states of the Markov chain and an edge set is chosen between states such that an edge exists between states x^1 and x^2 only if there is a positive probability of moving from state x^1 to x^2 in one iteration. See, for example, Diaconis and Stroock (1991) and Sinclair (1992). While this approach seems promising in providing precise bounds, for our image restoration problem we were only able to find canonical paths that gave convergence $O(e^N)$, even for the one-dimensional model.

Using a path bounds approach, Jerrum and Sinclair (1993) develop a Markov chain algorithm for estimating the partition function of the Ising model that they show is polynomial time.

For the case with no data, corresponding to the stochastic Ising model with no external field, our results apply for small values of β in dimensions higher than one, corresponding to large temperature when the model is considered in thermodynamic terms. Frigessi, di Stefano, Hwang and Sheu (1993) consider the question of which Markov chain Monte Carlo algorithm provides fastest convergence for this problem, comparing them via their eigenvalues. They show that, for high temperature, the single-site Metropolis algorithm gives the slowest convergence of any random scan updating dynamic. While the Gibbs sampler is better, they also show that convergence can be improved by considering dynamics which include the current value of the site being updated.

3.4 The Case of No Data: the Stochastic Ising Model

We will first apply our result to the case where we have no observed image, so we are sampling from our prior distribution, the Ising model without a external field.

3.4.1 One Dimension

We begin by considering the one-dimensional case, with each interior site equally influenced by its two nearest neighbours. So the prior density is

$$\pi_\beta(x) = \frac{1}{Z} \exp\left(\beta \sum_{i=1}^{N-1} x_i x_{i+1}\right) \tag{3.14}$$

and the full conditionals (3.4) for interior sites are

$$\pi_{FC}(x_i|x_{-i}) = \frac{\exp\{\beta(x_{i-1}x_i + x_i x_{i+1})\}}{\exp\{\beta(x_{i-1} + x_{i+1})\} + \exp\{\beta(-x_{i-1} - x_{i+1})\}} \tag{3.15}$$

and for end sites are

$$\pi_{FC}(x_i|x_{-i}) = \frac{\exp\{\beta x_i x_j\}}{\exp\{\beta x_j\} + \exp\{-\beta x_j\}}, \quad (i,j) = (1,2) \text{ or } (i,j) = (n, n-1). \tag{3.16}$$

The Left-to-Right Sweep Distance Function

Recall that our bound on the convergence time requires a bound on the mean time to couple for Markov chains started in the maximal state, where each pixel is $+1$, and the minimal state, where each pixel is -1. Define the distance function between the current two states of these Markov chains to be $\Phi_s \stackrel{\text{def}}{=} N - c$ where N is the total number of pixels and c is the number of sites at the right end that have coupled. For example, at some time t, suppose the configurations of the Markov chains started in the maximal and minimal states are

$$\begin{array}{llllllll} X_t^{max}: & + & + & \ldots & + & - & + & + \\ X_t^{min}: & + & - & \ldots & - & - & + & + \end{array} ;$$

then $\Phi_s(t) = N - 3$. Note that $\Phi_s(0) = N$ and $\Phi_s(T) = 0$. We call this distance function the "Sweep Distance Function". The following upper bound on the convergence time exists for sampling from the one-dimensional Ising model at all values of β.

Theorem 3.2 *For sampling via the random scan Gibbs sampler from the one-dimensional Ising model with N sites given by (3.14), the convergence time (3.8) can be bounded above by*

$$\tau_\beta(\epsilon) \leq 2eN^2 \left(\frac{e^{2\beta} + e^{-2\beta}}{2e^{-2\beta}} \right) (1 + \log \epsilon^{-1})$$

for all values of the Ising model parameter β, where ϵ is the specified tolerance for convergence in total variation distance.

Proof Consider all possible configurations of a site and its neighbours for which a change in Φ_s may occur. Since we are considering the random scan Gibbs sampler, one pixel is updated at each iteration. At each step in the algorithm, Φ_s will change by $+1$ if the $(N-c+1)^{th}$ site changes and decrease by 1 or more if the $(N-c)^{th}$ site changes. We call sites which can contribute to an increase in Φ_s "bad" sites, and sites which can contribute to a decrease in Φ_s "good" sites. Updating a good site may result in a change in c of more than one if sites to the left of the site being updated have already coupled. However, we will create our bound by considering worst case scenarios, so we will consider good updates which only decrease Φ_s by 1. Note that, because sites to the right of the $(N-c+1)^{th}$ site have the same neighbours in both configurations, they will change in the same manner, so they cannot affect the value of the distance function. If a site to the left of the $(N-c)^{th}$ site is chosen for updating, the value of the distance function cannot change.

The configurations of three interior sites illustrated in Figure 3.1 will possibly result in a change to Φ_s. The site being updated is to the left of the boundary for good sites (the $(N-c)^{th}$ site) and to the right of the boundary for bad sites (the $(N-c+1)^{th}$ site). The top row indicates the current configuration of X_t^{max}, the chain started in the maximal configuration, and the bottom row indicates the current configuration of X_t^{min}, the chain started in the minimal configuration. The update probabilities are calculated from (3.15). A site is updated to $+1$ if the uniform random number used at the current iteration for updating is less than or equal to the value of (3.15) when $x_i = +1$, and is otherwise updated to -1. Recall that both coupled chains are updated with the same uniform random number. As an example, suppose the site to the left of the boundary in the first configuration shown has been selected for updating. Then

$$
\begin{aligned}
&\Pr(\Delta\Phi_s \text{ occurring}) \\
&\quad = \Pr\left(\text{configuration becomes} \quad \begin{array}{lccccccc} X_t^{max}: & + & + & + & & + & - & + \\ & & & & \text{or} & & & \\ X_t^{min}: & - & + & + & & - & - & + \end{array}\right) \\
&\quad = \min\left\{\frac{e^{2\beta}}{e^{2\beta}+e^{-2\beta}}, \frac{1}{2}\right\} + \min\left\{\frac{e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}, \frac{1}{2}\right\}
\end{aligned}
$$

where the minimums are over the probabilities of X_{t+1}^{max} and X_{t+1}^{min} respectively being updated as shown given their configurations at iteration t.

If the boundary is at the end, the end site is a good site and there are no bad sites. This site couples with probability 1 if its neighbouring pixel is the same in both configurations, and with probability $2e^{-\beta}/(e^{\beta}+e^{-\beta})$ if its

Configuration	**Probability of $\Delta\Phi_s$ occurring if site chosen**

GOOD SITES ($\Delta\Phi_s = -1$):

$$\begin{array}{l} X_t^{max}: \\ X_t^{min}: \end{array} \quad \begin{array}{cc|c} + & + & + \\ - & - & + \end{array} \; , \; \begin{array}{cc|c} + & + & - \\ - & - & - \end{array} \qquad \frac{1}{2} + \frac{e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}$$

GOOD SITES ($\Delta\Phi_s < -1$):

$$\begin{array}{l} X_t^{max}: \\ X_t^{min}: \end{array} \quad \begin{array}{cc|c} + & + & + \\ + & - & + \end{array} \; , \; \begin{array}{cc|c} + & + & - \\ + & - & - \end{array} \qquad 1$$

$$\begin{array}{l} X_t^{max}: \\ X_t^{min}: \end{array} \quad \begin{array}{cc|c} - & + & + \\ - & - & + \end{array} \; , \; \begin{array}{cc|c} - & + & - \\ - & - & - \end{array} \qquad 1$$

BAD SITES ($\Delta\Phi_s = +1$):

$$\begin{array}{l} X_t^{max}: \\ X_t^{min}: \end{array} \quad \begin{array}{c|cc} + & + & - \\ - & + & - \end{array} \; , \; \begin{array}{c|cc} + & + & + \\ - & + & + \end{array} \qquad \frac{1}{2} - \frac{e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}$$

$$\begin{array}{l} X_t^{max}: \\ X_t^{min}: \end{array} \quad \begin{array}{c|cc} + & - & - \\ - & - & - \end{array} \; , \; \begin{array}{c|cc} + & - & + \\ - & - & + \end{array} \qquad \frac{1}{2} - \frac{e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}$$

Figure 3.1: Possible configurations that may lead to a change in the sweep distance function in one dimension.

neighbouring pixels differ and the expected change in the distance is at most

$$-\frac{1}{N}\left(\frac{2e^{-\beta}}{e^{\beta}+e^{-\beta}}\right).$$

If the boundary is in the interior, we obtain a bound on the expected change in the distance function as follows.

In order to obtain an upper bound on $\mathbf{E}[\Delta\Phi_s]$ that holds for all configurations, we assume that the site to the left of the boundary is a good site with the smallest probability of changing Φ_s and that results in a change of Φ_s of only 1. At each iteration, the site to be updated is chosen uniformly from the N pixels. Thus

$$\begin{aligned}
&\mathbf{E}[\Delta\Phi_s|X_t^{max}, X_t^{min}] \\
&\leq \frac{1}{N}\left\{\left(\frac{1}{2}-\frac{e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}\right)-\left(\frac{1}{2}+\frac{e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}\right)\right\} \\
&\quad \text{for all } t \text{ where } X_t^{max} \neq X_t^{min} \\
&< 0 \text{ for all } \beta.
\end{aligned}$$

Applying Theorem 3.1 and Equation (3.12) with $a = \frac{1}{N}\frac{2e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}$, the mean coupling time can be bounded above by

$$T \leq N^2\left(\frac{2e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}\right)^{-1}$$

where T has been defined in (3.10), and applying (3.11) gives the result. ∎

For example, if $\epsilon = 0.01$ and $\beta = 0.5$, $\tau \leq 128N^2$. Using a value such as $\beta = 1.5$ gives more influence to the smoothing inherent in the prior distribution and gives the convergence bound $\tau \leq 6162N^2$.

Note that by considering the distance function as the total number of sites less the number of sites coupled at *both* ends, the mean coupling time can be reduced by a factor of 2.

There is no phase transition in the one-dimensional Ising model (see, for example, Cipra (1987)), so a result such as this that holds for all β should exist. However, in higher dimensions, convergence is known to change at the critical value of β at which phase transition occurs. Convergence is known to be slow for β above this value. Our results for higher dimensions hold for small β, below this critical value.

3.4.2 Extension to Higher Dimensions and Larger Neighbourhood Systems

The Sweep Distance Function

In two and higher dimensions, there is no simple distance function analogous to the sweep distance function of one dimension. The immediately obvious analogue, where the number of sites coupled at an endpoint is replaced by the size of a corner that is coupled, is not appropriate since, on any future step of the algorithm, any of the sites along the coupled boundary may change, destroying the structure. An irregular boundary around the coupled sites can change in many ways, including losing contact with any corner or edge sites, making it very complex to keep track of the size of the coupled corner. Considering a cluster of coupled sites seems to be too complex to be useful.

For a systematic scan Gibbs sampler it may be possible to define a distance function like this, since at each iteration all pixels are updated and the number coupled in a corner structure can be maintained.

We address this problem by defining a different distance function, which will lead to restrictions on the values of β.

The Number-of-Sites-Different Distance Function

Define the distance function Φ_d as the number of sites where the two chains differ. Then $\Phi_d(0) = N$ and $\Phi_d(T) = 0$. This distance function can be used in any dimension. We call Φ_d the "Number-of-Sites-Different Distance Function". A change in Φ_d may now occur for any site chosen for updating, unless it is in the middle of a string of at least three coupled sites.

Our result is stated in terms of n, the number of nearest neighbours that are equally influential; n is typically 2 in one dimension, either 4 or 8 in two dimensions, etc. Our upper bound on the convergence time is still a simple function of the model parameter β times N^2 where N is the total number of sites; however it now holds only for a restricted range of β. As the number of influential neighbours increases, the range of admissible values of β decreases.

Theorem 3.3 *For sampling via the random scan Gibbs sampler from the Ising model (3.1) and (3.2) in arbitrary dimension with N sites where each site is influenced by its n nearest neighbours, the convergence time (3.8) can be bounded above by*

$$\tau_\beta(\epsilon) \leq 2eN^2 \left\{ \frac{e^{n\beta} + e^{-n\beta}}{(n+2)e^{-n\beta} - ne^{n\beta}} \right\} \left(1 + \log \epsilon^{-1}\right)$$

for

$$0 \leq \beta \leq \frac{1}{2n} \log \left(\frac{n+2}{n} \right),$$

where β is the Ising model parameter and ϵ is the specified tolerance for convergence in total variation distance.

In two dimensions, the critical value for the two-dimensional Ising model where each pixel has four influential neighbours is known to be $\log(1+\sqrt{2})/2$ (Liggett 1985, p. 204). When β is above this value convergence is known to be slow. Our upper bound on β is well below the critical value, but to our knowledge these are the first precise bounds for any value of β.

Proof of theorem Consider all possible configurations of a site and its n neighbours in which a change in Φ_d, the number-of-sites-different distance function, may occur. Since we are using the random scan Gibbs sampler, at each iteration Φ_d can change by at most 1. A site which can lead to a change in Φ_d of -1 is considered a "good" site and $+1$ a "bad" site.

For ease of presentation, the possible configurations are illustrated in Figure 3.2 in one dimension with two influential neighbours. The argument in higher dimensions and with more influential neighbours is completely analogous. Figure 3.2 shows the configurations in one dimension of interior sites (mirror images not repeated), where the middle site is the one randomly chosen for updating, and end sites where the right-most pixel is being updated, which will possibly result in a change in Φ_d. The top row indicates the current configuration of X_t^{max}, the chain started in the maximal configuration, and the bottom row indicates the current configuration of X_t^{min}, the chain started in the minimal configuration.

Note that each bad site has $\begin{array}{c} + \\ - \end{array}$ as at least one of its neighbours, which would be a good site were it chosen. So there are **at most** 2 bad sites for

Configurations of interior sites	**Probability of $\Delta\Phi_d$ occurring if site chosen**
GOOD SITES ($\Delta\Phi_d = -1$):	
$\begin{array}{lccc} X_t^{max}: & + & + & + \\ X_t^{min}: & - & - & - \end{array}$	$\frac{2e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}$
$\begin{array}{lccc} X_t^{max}: & + & + & + \\ X_t^{min}: & + & - & - \end{array} , \begin{array}{ccc} + & + & - \\ - & - & - \end{array}$	$\frac{1}{2} + \frac{e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}$
$\begin{array}{lccc} X_t^{max}: & + & + & + \\ X_t^{min}: & + & - & + \end{array} , \begin{array}{ccc} + & + & - \\ + & - & - \end{array}$	1
BAD SITES ($\Delta\Phi_d = +1$):	
$\begin{array}{lccc} X_t^{max}: & + & - & + \\ X_t^{min}: & - & - & + \end{array} , \begin{array}{ccc} + & - & - \\ - & - & - \end{array}$	$\frac{1}{2} - \frac{e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}$
$\begin{array}{lccc} X_t^{max}: & + & - & + \\ X_t^{min}: & - & - & - \end{array}$	$\frac{e^{2\beta}-e^{-2\beta}}{e^{2\beta}+e^{-2\beta}}$

Configurations of end sites	**Probability of $\Delta\Phi_d$ occurring if site chosen**
GOOD SITES ($\Delta\Phi_d = -1$):	
$\begin{array}{lcc} X_t^{max}: & + & + \\ X_t^{min}: & - & - \end{array}$	$\frac{2e^{-\beta}}{e^{\beta}+e^{-\beta}}$
$\begin{array}{lcc} X_t^{max}: & + & + \\ X_t^{min}: & + & - \end{array} , \begin{array}{lcc} X_t^{max}: & - & + \\ X_t^{min}: & - & - \end{array}$	1
BAD SITES ($\Delta\Phi_d = +1$):	
$\begin{array}{lcc} X_t^{max}: & + & + \\ X_t^{min}: & - & + \end{array} , \begin{array}{lcc} X_t^{max}: & + & - \\ X_t^{min}: & - & - \end{array}$	$\frac{e^{\beta}-e^{-\beta}}{e^{\beta}+e^{-\beta}}$

Figure 3.2: Possible configurations that may lead to a change in the number-of-sites-different distance function in one dimension.

each good site. If a bad site is chosen, a change of $+1$ in Φ_d occurs with probability at most

$$\frac{e^{2\beta} - e^{-2\beta}}{e^{2\beta} + e^{-2\beta}}.$$

If a good site is chosen, a change of -1 in Φ_d occurs with probability at least

$$\frac{2e^{-2\beta}}{e^{2\beta} + e^{-2\beta}}.$$

In the general case, where each site is influenced by its n nearest neighbours, there are at most n bad sites for each good site. If a good site is chosen, a change of -1 in Φ_d occurs with probability at least

$$\frac{2e^{-n\beta}}{e^{n\beta} + e^{-n\beta}}.$$

To see that this is the good change whose configuration has the highest probability of occurring, we consider the full conditionals (3.4). Suppose site i has been chosen for updating.

$$\Pr(x_i = \iota \,|\, n \text{ neighbours of } i \text{ are } -\iota) = \frac{e^{-n\beta}}{e^{n\beta} + e^{-n\beta}} \tag{3.17}$$

$$\Pr(x_i = \iota \,|\, n-1 \text{ neighbours of } i \text{ are } -\iota\ ) = \frac{e^{(-n+2)\beta}}{e^{-(-n+2)\beta} + e^{(-n+2)\beta}} \tag{3.18}$$

$$\vdots$$

$$\Pr(x_i = \iota \,|\, n \text{ neighbours of } i \text{ are } \iota) = \frac{e^{n\beta}}{e^{n\beta} + e^{-n\beta}} \tag{3.19}$$

where $\iota \in \{+1, -1\}$. Both coupled chains are updated using the same uniform random number, ξ. A site is updated to $+1$ if $\xi \leq \Pr(x_i = +1 | x_{-i})$. A good change occurs if the x_i is updated to $+1$ or -1 in both chains. The configurations where this has the least probability of occurring are those in which all neighbours of the site are the opposite value and from (3.17) this has probability $e^{-n\beta}/(e^{n\beta} + e^{-n\beta})$. Thus we have two times this value for the smallest probability of a good change among all possible configurations.

If a bad site is chosen, a change of $+1$ in Φ_d occurs with probability at most

$$\frac{e^{n\beta} - e^{-n\beta}}{e^{n\beta} + e^{-n\beta}}.$$

This is one minus the probability of a good update for the configuration which has least probability of coupling at the updating site.

At each iteration, a particular site is chosen with probability $\frac{1}{N}$ for updating. Thus, for all t where $X_t^{max} \neq X_t^{min}$

$$\begin{aligned}
&\mathbf{E}(\Delta\Phi_d | X_t^{max}, X_t^{min}) \\
&\quad = \frac{1}{N}\left\{\sum_{\text{bad sites}} \Pr(\text{this change occurs}) - \sum_{\text{good sites}} \Pr(\text{this change occurs})\right\} \\
&\quad \leq \frac{1}{N}\left\{(\text{Number of bad sites}) \cdot \frac{e^{n\beta} - e^{-n\beta}}{e^{n\beta} + e^{-n\beta}}\right. \\
&\qquad\qquad \left. -(\text{Number of good sites}) \cdot \frac{2e^{-n\beta}}{e^{n\beta} + e^{-n\beta}}\right\}
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{N}\left\{n \cdot (\text{Number of good sites}) \cdot \frac{e^{n\beta} - e^{-n\beta}}{e^{n\beta} + e^{-n\beta}}\right. \\
&\qquad \left. - (\text{Number of good sites}) \cdot \frac{2e^{-n\beta}}{e^{n\beta} + e^{-n\beta}}\right\} \\
&= \frac{(\text{Number of good sites})}{N}\left\{n \cdot \frac{e^{n\beta} - e^{-n\beta}}{e^{n\beta} + e^{-n\beta}} - \frac{2e^{-n\beta}}{e^{n\beta} + e^{-n\beta}}\right\}.
\end{aligned}$$

For

$$\beta < \frac{\log(\frac{n+2}{n})}{2n},$$

this is negative. The number of good sites is Φ_d. The chain has coupled when Φ_d reaches 0, so at each iteration, the number of good sites is at least 1. Thus, the mean coupling time T can be bounded above by

$$T \leq N^2 \left(\frac{(n+2)e^{-n\beta} - ne^{n\beta}}{e^{n\beta} + e^{-n\beta}}\right)^{-1}$$

and applying (3.11) gives the result. ∎

For example, if $n = 4$ as for two dimensions with neighbours above, below, and beside, $\epsilon = 0.01$, and $\beta = 0.05$, then $\tau \leq 2322N^2$. Reducing β to 0.01 gives an improvement in our upper bound on τ to $38N^2$. In the case of $n = 8$, as would occur in two dimensions including adjacent diagonals as neighbours, $\beta = 0.01$ gives $\tau \leq 108N^2$.

In addition to introducing restrictions on the values of β, the results for this distance function give larger bounds on the convergence time than those obtained with the sweep distance function in one dimension. However, the result using Φ_d is applicable in any dimension.

Note: The result of Theorem 3.3 is not sharp. Our result is $O(N^2)$, rather than the known rate of $O(N \log N)$ (the upper limit for β in our results is well below the critical value). Moreover, the limiting configuration (n bad sites for each good site, with these sites in the configurations that the bad sites are those most likely to uncouple and the good sites are those least likely to couple) cannot occur in isolation. However, we have obtained a precise bound.

As an indication of the role of the error tolerance, ϵ, we simulated 1000 coupled pairs of Markov chains, started in the maximal and minimal states. We used the random scan Gibbs sampler with full conditionals (3.4). The image was a square grid of pixels of size 32×32, with neighbours being the pixels directly above, below, and beside. We use the following characterisation of the total variation distance between two probability measures μ and ν

$$d_{TV}(\mu, \nu) = \inf \Pr(X \neq Y)$$

where the infimum is over all random variables X and Y where $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$ (see, for example, Lindvall (1992, p. 19)). In Figure 3.3

we have plotted the number of iterations versus the probability the Markov chains have not coupled which is our lower bound for the total variation distance. Tight requirements on ϵ require increasing numbers of iterations, while fewer than 7000 iterations do not give a randomised chain. Note that, while this forward coupling time gives an indication of the time required for convergence to stationarity, we cannot use the resulting state as a sample from the stationary distribution. Doing so would bias our results in favour of states at which the probability of coupling is greater (Propp and Wilson 1996).

3.5 The Case with Observed Data

Suppose $x = (x_1, x_2, \ldots, x_N)$ is the true configuration and $y = (y_1, y_2, \ldots, y_N)$ is the observed configuration. To model the true configuration, we seek a sample from the posterior distribution of X given Y, $\pi_{posterior}(x|y)$.

To calculate the posterior, use Bayes' Theorem

$$\pi_{posterior}(x|y) \propto \rho(y|x)\, \pi_\beta(x)$$

where $\rho(y|x)$ is the likelihood model for the distorted data given the true image and π_β is the Ising model prior.

We will only consider the number-of-sites-different distance function, since

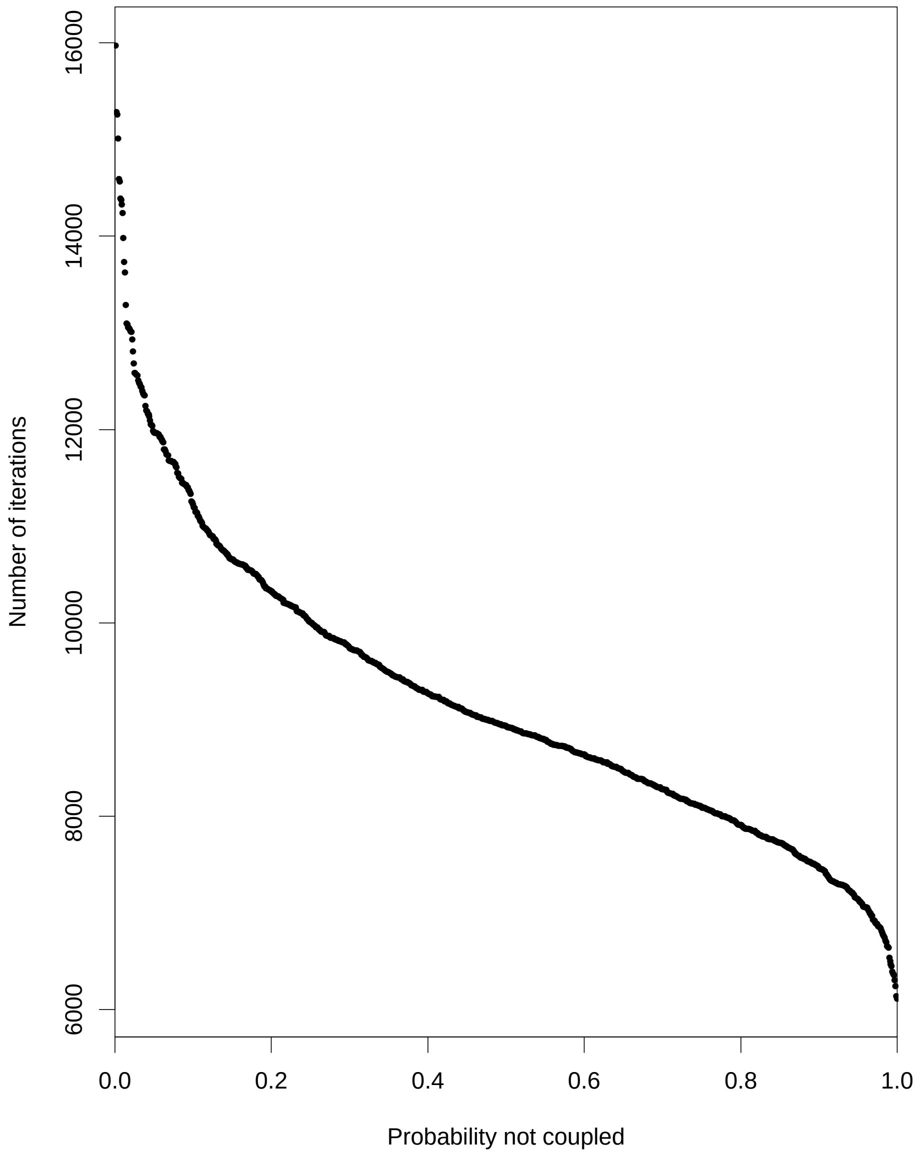


Figure 3.3: The number of iterations required for various error tolerances in total variation distance (indicated as the probability not coupled) based on 1000 simulations.

it applies to all dimensions.

3.5.1 True Image with Random Flips

Suppose the observed configuration y consists of the true configuration x, $x_i \in \{+1, -1\}$, with each spin site flipped with probability α, i.e.

$$\Pr(Y_i = y_i | X_i = x_i) = \begin{cases} \alpha & \text{if } y_i \neq x_i \\ 1 - \alpha & \text{if } y_i = x_i \end{cases} .$$

Then

$$\rho(y|x) = \alpha^{\sum_{j=1}^{N} \mathbf{1}_{y_j \neq x_j}} (1-\alpha)^{N - \sum_{j=1}^{N} \mathbf{1}_{y_j \neq x_j}}. \tag{3.20}$$

As an example of the results of our calculations, we give the posterior distribution and the full conditionals in one dimension since it is notationally simplest. Higher dimensional calculations, with more influential neighbours, are completely analogous. Combining (3.20) with the prior (3.1) and (3.2), the posterior distribution for the true configuration given the observed configuration is

$$\pi_{posterior}(x|y) = \tag{3.21}$$
$$\frac{1}{Z_y} \exp\left[\beta \sum_{i=1}^{N-1} x_i x_{i+1} + \sum_{j=1}^{N} \mathbf{1}_{y_j \neq x_j} \log(\alpha) + \left\{N - \sum_{j=1}^{N} \mathbf{1}_{y_j \neq x_j}\right\} \log(1-\alpha)\right]$$

where Z_y is the required posterior normalising constant. This is an example of a Gibbs distribution with an external field.

The posterior full conditionals for interior sites can then be calculated to be

$$\begin{aligned}
&\pi_{FC}(x_i|x_{-i}, y) \\
&= \exp\left\{\beta(x_{i-1}x_i + x_i x_{i+1}) + \mathbf{1}_{y_i \neq x_i}\log\left(\frac{\alpha}{1-\alpha}\right)\right\} \\
&\quad \times \left[\exp\left\{\beta(x_{i-1} + x_{i+1}) + \mathbf{1}_{y_i \neq 1}\log\left(\frac{\alpha}{1-\alpha}\right)\right\}\right. \\
&\quad\quad \left. + \exp\left\{\beta(-x_{i-1} - x_{i+1}) + \mathbf{1}_{y_i \neq -1}\log\left(\frac{\alpha}{1-\alpha}\right)\right\}\right]^{-1}
\end{aligned} \tag{3.22}$$

We now state our convergence bound for arbitrary dimension.

Theorem 3.4 *Suppose we have observed, in arbitrary dimension, an image of N pixels taking the values $+1$ or -1 where it is known that each pixel is incorrectly observed with probability α. For sampling from the posterior Gibbs distribution with our prior distribution the Ising model, (3.1) and (3.2), via the random scan Gibbs sampler, the convergence time (3.8) can be bounded as*

$$\tau_\beta(\epsilon) \leq 2eN^2\left\{\frac{k_\alpha + e^{2n\beta} + e^{-2n\beta}}{k_\alpha - ne^{2n\beta} + (n+2)e^{-2n\beta}}\right\}\left(1 + \log\epsilon^{-1}\right)$$

for

$$0 \leq \beta \leq \frac{1}{2n}\log\left[\frac{k_\alpha + \{k_\alpha^2 + 4n(n+2)\}^{\frac{1}{2}}}{2n}\right]$$

where $k_\alpha = (1-\alpha)/\alpha + \alpha/(1-\alpha)$, β is the Ising model parameter, n is the *number of nearest neighbours of interior pixels, and ϵ is the specified tolerance for convergence in total variation distance.*

Proof As in the case of no data, there are most n good sites (contributing to a decrease in Φ_d) for every bad site (contributing to an increase in Φ_d). Update probabilities are calculated from the full conditionals (3.22). The ordering of the update probabilities given the neighbours of the pixel being updated is unaffected by the additional terms in the exponents involving $\log[\alpha/(1-\alpha)]$, i.e. it is the same as given in equations (3.17)–(3.19) in the case of no data. Thus, regardless of the observed value at the site being updated, the good configuration with least probability of coupling is all $+1$, all -1, with update probability

$$1 - \frac{e^{n\beta}}{e^{n\beta} + e^{-n\beta + \log\left(\frac{\alpha}{1-\alpha}\right)}} + \frac{e^{-n\beta}}{e^{-n\beta} + e^{n\beta + \log\left(\frac{\alpha}{1-\alpha}\right)}}. \tag{3.23}$$

And the bad configuration with greatest probability of uncoupling has update probability

$$\frac{e^{n\beta}}{e^{n\beta} + e^{-n\beta + \log\left(\frac{\alpha}{1-\alpha}\right)}} - \frac{e^{-n\beta}}{e^{-n\beta} + e^{n\beta + \log\left(\frac{\alpha}{1-\alpha}\right)}}.$$

Thus,

$$\begin{aligned}
&\mathbf{E}[\Delta\Phi | X_t^{max}, X_t^{min}] \\
&\quad \leq \frac{(\text{Number of good sites})}{N} \times
\end{aligned}$$

$$\Bigg\{ n \cdot \frac{e^{-\log\left(\frac{\alpha}{1-\alpha}\right)}\left(e^{2n\beta}-e^{-2n\beta}\right)}{\left(e^{n\beta}+e^{-n\beta-\log\left(\frac{\alpha}{1-\alpha}\right)}\right)\left(e^{-n\beta}+e^{n\beta-\log\left(\frac{\alpha}{1-\alpha}\right)}\right)}$$

$$- \left(1-\frac{e^{-\log\left(\frac{\alpha}{1-\alpha}\right)}\left(e^{2n\beta}-e^{-2n\beta}\right)}{\left(e^{n\beta}+e^{-n\beta-\log\left(\frac{\alpha}{1-\alpha}\right)}\right)\left(e^{-n\beta}+e^{n\beta-\log\left(\frac{\alpha}{1-\alpha}\right)}\right)}\right)\Bigg\}$$

$$= \frac{(\text{Number of good sites})}{N} \cdot \frac{ne^{2n\beta}-(n+2)e^{-2n\beta}-k_\alpha}{e^{2n\beta}+e^{-2n\beta}+k_\alpha}$$

where $k_\alpha = (1-\alpha)/\alpha + \alpha/(1-\alpha)$. For this to be negative

$$ne^{4n\beta} - k_\alpha e^{2n\beta} - (n+2) < 0.$$

Applying Theorem 3.1 and (3.11) gives our result. ■

Note that the result is the same when the flip rate is $1-\alpha$ as when it is α, so the values of β that guarantee convergence in $O(N^2)$ time are the same for a flip rate of, for example, .05 as for .95.

If $\alpha = 0$ the observed image is correct and when $\alpha = 1$ the observed image is completely incorrect. In these cases, our result holds for all β. At each iteration the randomly chosen pixel will become the correct value and the expected change in the distance function is the negative probability that a good site is chosen.

If $\alpha = 1/2$ the observed image gives no information. Our result then coincides with the no data case of Theorem 3.3. To see this for the bound on the convergence time, divide numerator and denominator by $e^{n\beta}+e^{-n\beta}$.

As an example of the results our theorem gives, if $\alpha = 0.05$, $n = 4$, $\beta = 0.05$ and $\epsilon = 0.01$, $\tau \leq 38N^2$, improving the bound from the case with no data by a factor greater than 60. Moreover, for $n = 4$ and $\alpha = 0.05$, the range of admissible values of β is 4 times as great as that in the case with no data. Smaller values of α increase the range of β and decrease the convergence time bound, reflecting the increased reliability of the observed image.

3.5.2 True Image with Additive Normal Noise

Geman and Geman (1984) consider the observed data to be obtained form the true image by a deterministic blurring mechanism and distortion due to the sensing equipment, in combination with normal noise. We will consider the simple case without blurring or sensor distortion and where the normal noise is additive at each pixel, *i.e.* $y = x + \mathcal{N}$, where $\mathcal{N}$ is a vector with each entry an independent sample from a $N(\mu, \sigma^2)$ distribution. We will only consider the case where $\mu = 0$. Since the value of $\mathcal{N} = y - x$ is independent of x, the likelihood model for the data given the true image is

$$\rho(y|x) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}||x - y||^2\right).$$

In one dimension, this gives the posterior density

$$\pi_{posterior}(x|y) = \frac{1}{Z_y} \exp\left\{\beta \sum_{i=1}^{N-1} x_i x_{i+1} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - y_i)^2\right\}. \tag{3.24}$$

For interior points ($i = 2, \ldots, N-1$) the full conditionals are

$$\begin{aligned}
&\pi_{FC}(x_i|x_{-i}, y) \\
&= \exp\left\{\beta(x_{i-1}x_i + x_i x_{i+1}) - \frac{1}{2\sigma^2}(y_i - x_i)^2\right\} \\
&\quad \times \left[\exp\left\{\beta(x_{i-1} + x_{i+1}) - \frac{1}{2\sigma^2}(y_i - 1)^2\right\}\right. \\
&\qquad \left. + \exp\left\{\beta(-x_{i-1} - x_{i+1}) - \frac{1}{2\sigma^2}(y_i + 1)^2\right\}\right]^{-1}.
\end{aligned}$$

Theorem 3.5 *Suppose we have observed, in arbitrary dimension, an image of N pixels where it is known that pixel i, $i = 1, \ldots, N$ should have a value of $+1$ or -1 but has been observed as a random sample from a $N(x_i, \sigma^2)$ distribution where x_i is the true value of the i^{th} pixel. For sampling from the posterior Gibbs distribution with our prior distribution the Ising model, (3.1) and (3.2), via the random scan Gibbs sampler, the convergence time (3.8) can be bounded as*

$$\tau_\beta(\epsilon) \leq 2eN^2 \left\{\frac{k_{\sigma,y_{min}} + e^{2n\beta} + e^{-2n\beta}}{k_{\sigma,y_{min}} + (n+2)e^{-2n\beta} - ne^{2n\beta}}\right\} \left(1 + \log \epsilon^{-1}\right)$$

for

$$0 \leq \beta \leq \frac{1}{2n} \log \left[\frac{k_{\sigma,y_{min}} + \left\{k^2_{\sigma,y_{min}} + 4n(n+2)\right\}^{\frac{1}{2}}}{2n}\right]$$

where $y_{min} = \min_i\{|y_i|\}$*, the smallest of the observed pixels in absolute value,* $k_{\sigma,y_{min}} = e^{2y_{min}/\sigma^2} + e^{-2y_{min}/\sigma^2}$*,* β *is the Ising model parameter,* n *is the number of nearest neighbours of interior pixels, and* ϵ *is the specified tolerance for convergence in total variation distance.*

Proof Suppose site i is being updated. It can be shown that, regardless of the value of y_i, the good configuration which has the smallest probability of becoming coupled is all $+1$, all -1. The probability of the middle site becoming the same in the two chains, given the data value is

$$\frac{2e^{-2n\beta} + e^{2y_i/\sigma^2} + e^{-2y_i/\sigma^2}}{e^{2n\beta} + e^{-2n\beta} + e^{2y_i/\sigma^2} + e^{-2y_i/\sigma^2}}. \tag{3.25}$$

Similarly, regardless of the value of y_i, the bad configuration which has the greatest probability of becoming uncoupled is has probability

$$1 - \frac{2e^{-2n\beta} + e^{2y_i/\sigma^2} + e^{-2y_i/\sigma^2}}{e^{2n\beta} + e^{-2n\beta} + e^{2y_i/\sigma^2} + e^{-2y_i/\sigma^2}}$$

of the middle site becoming different. The data value that minimises the least probable good probability and maximises the most probable bad is $\min_i\{|y_i|\}$. Substituting this for y_i and using the same argument as in the proof of Theorem 3.4 gives our result. ∎

For the case where $\sigma = 0.3$ and $n = 8$, a value of y_{min} such as 0.65 gives $\beta \leq 0.773$. As a guide to what is an appropriate value of β, we consider the work of Besag (1986). For $n = 8$, he found that a parameter value which is equivalent to $\beta = 0.75$ in our model worked well in practice.

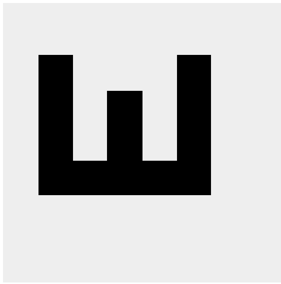
Note that smaller values of the variance of the normal noise increase the range of possible values of β for which our results hold and decrease the upper bound on the convergence time, reflecting the increased reliability of the observed image. In the limit as $\sigma \to \infty$, our observed image gives no information. In this case, our result coincides with the no data case of Theorem 3.3. To see this for the bound on the convergence time, divide top and bottom by $e^{2n\beta} + e^{-2n\beta}$. The largest bound on the convergence time and the smallest range for β, minimised over values of y_{min}, occur when $y_{min} = 0$; in this case, the result again coincides with the no data case.

Figure 3.4 gives an example of the image restoration process. Figure 3.4(a) shows the original image, drawn on a 32×32 grid. It was randomly degraded with $N(0, 0.4^2)$ noise, added independently to each pixel. The degraded image is shown in Figure 3.4(b). Our prior parameter, β, was set at 0.05, and each pixel's neighbours were the pixels to the left and right and directly above and below. The specified error tolerance for randomisation in total variation distance was 0.01. The algorithm was run for the number of iterations our theory specifies, taking y_{min} to be 0, from the initial state with every pixel black. Our approximate sample from the posterior distribution is shown in Figure 3.4(c).

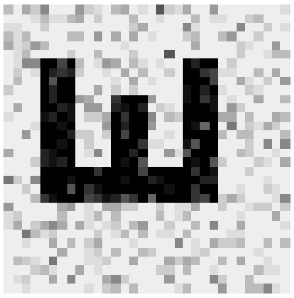
3.6 The Expected Number of Steps Required for Exact Sampling

Recently there has been a great deal of interest in exact sampling algorithms such as the coupling-from-the-past algorithm of Propp and Wilson (1996). This algorithm can be used in our examples by running two coupled realisations of the Markov chain, starting in the maximal and minimal states at some time $-t$ in the past. If the two chains have coupled at time zero, the state at time zero is an exact sample from the Markov chain's stationary distribution.

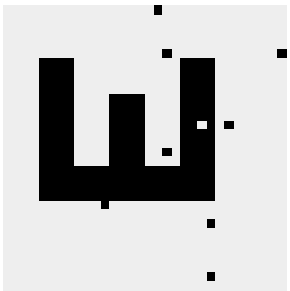
Our results give an indication of the expected value of $-t$ required for coalescence at time zero in Propp and Wilson's algorithm. As Propp and Wilson note, the random variables T^*, the smallest t such that chains started in the maximal and minimal state have coupled at time t, and T_*, the smallest t such that chains started in the maximal and minimal state at time $-t$ will be in the same state at time zero, have the same probability distribution.



(a)



(b)



(c)

Figure 3.4: A simulated restoration of a 32×32 image. (a) true image; (b) observed image; (c) sample from the posterior distribution.

By Theorem 3.1

$$\mathbf{E}(T_*) = \mathbf{E}(T^*) \leq \frac{N}{a}$$

where a is the positive constant with $\mathbf{E}(\Delta\Phi|X_t, Y_t) < -a$ for all t. For example, in the case of no data

$$\mathbf{E}(T^*) \leq N^2 \left\{ \frac{e^{n\beta} + e^{-n\beta}}{(n+2)e^{-n\beta} - ne^{n\beta}} \right\}.$$

Application of Markov's inequality to our result gives an upper bound on the probability that the coupling-from-the-past algorithm will take a very large number of runs as follows

$$\Pr(T^* > B) < \frac{\mathbf{E}(T^*)}{B} \leq \frac{N}{aB}.$$

Chapter 4

Convergence in the Wasserstein Metric

4.1 Introduction

In this chapter we introduce the use of the Wasserstein metric to the study of the theoretical rates of convergence of MCMC algorithms. Like total variation distance, which is the usual metric chosen to quantify an MCMC algorithm's distance from its stationary distribution, the Wasserstein metric has a coupling characterisation. However, the Wasserstein metric may be more useful on continuous state spaces than the total variation distance. The coupling characterisation for total variation distance is the probability that two random variables with the relevant distributions become equal, while the Wasserstein coupling characterisation is the expected distance between the random variables. Thus it is possible to consider convergence in the Wasserstein metric by considering coupling chains which may never coalesce exactly.

Our results hold for bounded state spaces. Convergence time of the Markov chain started in any state to the stationary distribution can be bounded as a function of the diameter of the space. Approximate convergence to the stationary distribution is bounded by the time for coupled chains to have an expected distance within ϵ, where ϵ is user-specified. Convergence to within ϵ precision has been considered by Møller (1999) in the context of applying exact sampling algorithms, which require coalescence of coupled Markov chains, to continuous state spaces. A discussion of the application of the results of this chapter to exact sampling is given in Section 4.6.

The particular application we address is the Bayesian restoration of a noisy image. We consider an image composed of pixels taking on values in a $[0, 1]$ grey-scale and a binary, black-white image. Our results hold for an image of any size or shape. For grey-scale images, we employ a pairwise difference prior distribution for the true image. For binary images, we use an Ising model prior. Both of these prior distributions give higher probability to images in which pixels tend to be like their nearest neighbours. Our results can accommodate any neighbourhood structure.

The theory that leads to the convergence bound is given in Section 4.2 and more discussion of the choice of probability metric is provided in Section 4.3 and Chapter 6. In Section 4.4, our method for creating a precise, *a priori* bound on the required number of iterations is applied to a Gibbs sampler used in Bayesian image restoration where the individual pixels are values from a $[0, 1]$ grey scale. Our method also allows an improvement to the bound on the convergence time found in Chapter 3 for the problem where the pixels are binary. This improvement is given in Section 4.5.

4.2 Convergence in the Wasserstein Metric

We now present a general method for bounding the convergence time of a discrete time Markov chain $\{X_t\}$ with bounded state space $\mathcal{X}$ in terms of the Wasserstein metric.

If μ, ν are two probability measures on the same space $\mathcal{X}$, the Wasserstein metric is

$$d_W(\mu, \nu) = \inf \mathbf{E}[d(X, Y)] \tag{4.1}$$

where d is any given metric on $\mathcal{X}$ and the infimum is taken over all random variables X, Y with $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$. By the Kantorovich-Rubinstein Theorem (see, for example, Dudley (1989, Theorem 11.8.2)), for $\mathcal{X}$ a separable metric space,

$$d_W(\mu, \nu) = \sup \left\{ \left| \int_{\mathcal{X}} f \, d\mu - \int_{\mathcal{X}} f \, d\nu \right| \right\}$$

where the supremum is over all functions f satisfying the Lipschitz condition

$$|f(x) - f(y)| \leq d(x, y).$$

The Wasserstein metric is sometimes referred to as the Kantorovich metric. It has been applied in the solution of Monge's 18^{th} century optimisation problem of finding the most efficient way of transporting soil (see, for example, Rachev (1984)). If μ is the distribution of soil particles and ν is the distribution of points where it is consumed, then $d_W(\mu, \nu)$ is the smallest cost at which all of the soil can be transported to its consumers.

Consider a Markov chain on a bounded state space $\mathcal{X}$ with $\text{diam}(\mathcal{X}) = \sup_{x,y \in \mathcal{X}} d(x, y)$ where $d(\cdot, \cdot)$ is any metric on $\mathcal{X}$. Assume that the Markov chain converges to a unique stationary distribution π. Let $P^T(x^0, \cdot)$ denote the distribution of the chain with initial state x^0 after T iterations. Theorem 4.1, stated and proven below, is a new, general result for Markov chains on bounded state spaces that can be used to determine the number of iterations that are necessary to achieve convergence in the Wasserstein metric. We consider two coupled realisations of the chain. If, on average, these realisations are getting closer together at each iteration, we can bound the time until the Wasserstein metric, $d_W(P^T(x^0,\cdot),\pi(\cdot))$, is small as follows.

Theorem 4.1 *Consider two coupled realisations of a Markov chain $\{X_t\}$ and $\{Y_t\}$ on a bounded state space $\mathcal{X}$ with stationary distribution π. Let $P^T(x^0,\cdot)$ denote the distribution of the chain with initial state x^0 after T iterations. Suppose we can find a constant $c \in (0,1)$ such that*

$$\mathbf{E}[d(X_{t+1},Y_{t+1})|X_t,Y_t] \leq c\, d(X_t,Y_t) \tag{4.2}$$

for all t. Then, $d_W(P^T(x^0,\cdot),\pi(\cdot)) < \epsilon$ for

$$T > \frac{\ln\left(\frac{\epsilon}{\operatorname{diam}(\mathcal{X})}\right)}{\ln c}$$

for any initial state x^0 where $\operatorname{diam}(\mathcal{X}) = \sup_{x,y\in\mathcal{X}} d(x,y)$.

The proof uses the following lemma.

Lemma 4.1 *Suppose $\{X_t\}$, $\{Y_t\}$ are two coupled Markov chains for which* *there exists a positive constant c such that*

$$\mathbf{E}[d(X_{t+1}, Y_{t+1})|X_t, Y_t] \leq c\, d(X_t, Y_t)$$

for all t. Then for any fixed T and any X_0, Y_0,

$$\mathbf{E}[d(X_T, Y_T)|X_0, Y_0] \leq c^T d(X_0, Y_0).$$

Proof of lemma The proof follows by induction. Suppose for some k

$$\mathbf{E}[d(X_k, Y_k)|X_0, Y_0] \leq c^k d(X_0, Y_0).$$

Then

$$\begin{aligned} \mathbf{E}[d(X_{k+1}, Y_{k+1})|X_0, Y_0] &= \mathbf{E}\left[\, \mathbf{E}[d(X_{k+1}, Y_{k+1})|X_k, Y_k]\,|X_0, Y_0\right] \\ &\leq \mathbf{E}[c\, d(X_k, Y_k)|X_0, Y_0] \\ &\leq c^{k+1}\, d(X_0, Y_0). \quad \blacksquare \end{aligned}$$

Proof of theorem We wish to bound the number of iterations, T, which guarantee $d_W(P^T(x^0, \cdot), \pi(\cdot)) \leq \epsilon$ where x^0 is any initial state. Consider another realisation of the Markov chain started in x which is a sample from π. Applying the lemma,

$$\begin{aligned} d_W(P^t(x^0, \cdot), \pi(\cdot)) &= \mathbf{E}[d(X_T, Y_T)|X_0 = x^0, Y_0 = x] \\ &\leq c^T d(x^0, x). \end{aligned}$$

$d_W(P^t(x^0,\cdot),\pi(\cdot))$ is less than or equal to ϵ for

$$T \geq \frac{\ln\left(\frac{\epsilon}{d(x^0,x)}\right)}{\ln c}.$$

Substituting $\mathrm{diam}(\mathcal{X})$ for $d(x^0,x)$ gives an upper bound on T. ■

Thus, if we can find a value of c satisfying (4.2), we can find a bound on the convergence time guaranteeing the Wasserstein metric is less than a specified tolerance ϵ.

4.3 Probability Metrics

There exist dozens of distance measures to quantify closeness between two probability measures (see, for example, Rachev (1991)). However, most considerations of the convergence of Markov chains have used total variation distance (for example Diaconis and Stroock (1991), Jerrum and Sinclair (1993), Tierney (1994), Tierney (1996)). Recall that the total variation distance between two probability measures μ and ν defined on $\mathcal{X}$ is

$$d_{TV}(\mu,\nu) = \sup_{A\subset\mathcal{X}} |\mu(A)-\nu(A)|.$$

If $\mathcal{X}$ is finite,

$$d_{TV}(\mu,\nu) = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x)-\nu(x)|.$$

This representation is half the L^1 metric used by Geman and Geman (1984). Some authors (for example, Tierney (1996)) define total variation distance as twice our definition.

Some of the popularity of total variation distance can be attributed to its ability to exhibit a threshold phenomenon (see, for example, Aldous and Diaconis (1987)). For many regular examples, the total variation distance drops suddenly from near 1 to near 0. Moreover, the equivalent formulation

$$d_{TV}(\mu, \nu) = \frac{1}{2} \max_{|h| \leq 1} \left| \int h \, d\mu - \int h \, d\nu \right|$$

where the maximum is taken over functions $h : \mathcal{X} \to \mathbb{R}$ satisfying $|h(x)| \leq 1$, allows a bound on total variation distance in terms of the expected value of some functions. Much of the success in bounding convergence in total variation distance arises from its coupling characterisation

$$d_{TV}(\mu, \nu) = \inf \Pr(X \neq Y)$$

where the infimum is taken over random variables X and Y whose distributions are μ and ν respectively. For examples of this, see Aldous and Diaconis (1987), Rosenthal (1995b), Luby et al. (1995), and Chapter 3.

To the author's knowledge, this is the first application of convergence in the Wasserstein metric to Markov chain Monte Carlo algorithms. The

Wasserstein metric was chosen for this application because of its coupling characterisation, Equation (4.1). Total variation distance is not practical in many applications on continuous state spaces as the total variation distance between any continuous distribution and any discrete distribution is always 1, irrespective of how well the continuous distribution is approximated by the discrete distribution. The Wasserstein metric metrizes convergence in distribution (see, for example, Dudley (1989)) while convergence in total variation distance is stronger than convergence in distribution.

The Prokhorov metric

$$d_P(\mu, \nu) =$$

$$\inf\{\alpha > 0 : \mu(B) \leq \nu(B^\alpha) + \alpha \text{ and } \nu(B) \leq \mu(B^\alpha) + \alpha \text{ for all Borel sets } B\}$$

where $B^\alpha = \{x : \inf_{y \in B} d(x, y) \leq \alpha\}$ also metrizes convergence in distribution. Convergence in the Wasserstein metric implies convergence in the Prokhorov metric because of the following relationship (see, for example, Huber (1981, p. 33))

$$d_P^2 \leq d_W.$$

The following relationship exists between total variation distance and the

Wasserstein metric:

$$d_W \leq \text{diam}(\mathcal{X})\, d_{TV}$$

where $\text{diam}(\mathcal{X}) = \sup_{x,y}\{d(x,y) : x, y \in \mathcal{X}\}$. If $\mathcal{X}$ is a finite set there is a bound the other way. If $d_{min} = \min_{x,y} d(x,y)$ for distinct points x, y in $\mathcal{X}$, then

$$d_{min}\, d_{TV} \ \leq \ d_W. \tag{4.3}$$

On an infinite set no such relation can occur as it is possible for d_W to go to 0 while d_{TV} remains fixed at 1. The relationship (4.3) will be proven in Section 4.5 and applied to give an improved precise bound for a result in Chapter 3. For a summary of some other commonly used probability metrics and the relationships that exist among them see Chapter 6.

4.4 Restoring a Grey-Scale Image

4.4.1 The Model and Algorithm

We now apply the convergence bound in the Wasserstein metric of Theorem 4.1 to an algorithm used to restore a distorted image.

Consider a grid of N pixels, each of which takes on a value in $[0, 1]$. For example a white pixel is 0, a black pixel 1 with values in between representing the various shades of grey. Our belief about the image represented in this manner is that pixels tend to be like their nearest neighbours. This belief is modelled by the pairwise-difference prior distribution on the value of the image $x = \{x_i\}_{i=1}^N$ which has density

$$\pi_\gamma(x) \propto \exp\left\{-\sum_{<i,j>} \frac{1}{2}\left[\gamma(x_i - x_j)\right]^2\right\} \tag{4.4}$$

on $[0,1]^N$ and 0 elsewhere. In the density, the sum is taken over pairs of pixels (i, j) which are nearest neighbours. The value of the parameter γ reflects the strength of the attractive force between neighbouring pixels. For a discussion of suitable priors, including this one, see Besag et al. (1995) and the references therein.

Our results hold for γ less than an upper limit, which depends on the neighbourhood structure and the variance of the normal noise. This upper limit on γ is an artifact of our proof which does not hold in the case where there are no corner or edge pixels. Simulations (see Section 4.4.3) indicate that our bounds are tight, except in the case where γ is close to its upper limit; in this case the number of iterations our theory predicts is overly conservative.

Note that we are not restricted by the size or shape of the grid, nor by its neighbourhood structure.

Rather than observing the true image x, we observe a distorted image where this distortion is due to random variation in our sensing mechanisms. We model this distortion as the addition of normal noise, added independently to the value of each pixel. We represent the observed image by $y = \{y_i\}_{i=1}^N$ and assume the noise has mean 0 and variance σ^2. Then the likelihood function is

$$\rho(y|x) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - y_i)^2\right\}.$$

Applying Bayes' Theorem gives our posterior density function for the distribution of the true image

$$\pi_{posterior}(x|y) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - y_i)^2 - \sum_{<i,j>}\frac{1}{2}\left[\gamma(x_i - x_j)\right]^2\right\} \tag{4.5}$$

on $[0,1]^N$ and 0 elsewhere. Our goal is to generate random samples from this posterior distribution which we can use to estimate moments and probabilities for the value of the true image.

We will generate these random samples by using the Gibbs sampler. We will use a randomly chosen, single-site updating scheme, considering one iteration as the update of one randomly selected pixel.

Let n_i be the number of neighbours that are influential for pixel i, $n_{max} = \max_i n_i$ and $n_{min} = \min_i n_i$. For example, in a two-dimensional image, n_i

is usually 4 or 8 for interior pixels, and, correspondingly, 3 or 5 for pixels on the boundaries, and 2 or 3 for pixels on the corners. From (4.5) the full conditional densities for pixels x_i, given the values of all other pixels, x_{-i}, and the observed image y are

$$\pi_{FC}(x_i|x_{-i}, y) \propto \exp\left\{-\frac{(\sigma^{-2}+n_i\gamma^2)}{2}\left[x_i - \left(\sigma^{-2}+n_i\gamma^2\right)^{-1}\left(\sigma^{-2}y_i + \gamma^2\sum_{j\sim i} x_j\right)\right]^2\right\} \tag{4.6}$$

on $[0,1]$ and 0 elsewhere. The pixels $\{x_j,\ j \sim i\}$ are those which are neighbours of the i^{th} pixel. Note that this is the restriction to $[0,1]$ of the normal distribution with mean $\left(\sigma^{-2}+n_i\gamma^2\right)^{-1}\left(\sigma^{-2}y_i + \gamma^2\sum_{j\sim i} x_j\right)$ and variance $\left(\sigma^{-2}+n_i\gamma^2\right)^{-1}$. The algorithm proceeds by choosing an initial state, and at each iteration, randomly selecting a pixel for updating according to (4.6). To sample from the normal distribution restricted to $[0,1]$, we use an inverse normal cumulative distribution function approximation, transforming the uniform random number as described in Fishman (1996, p. 152) for restricted sampling. For an accurate approximation to the inverse normal distribution function, see, for example, Thisted (1988, p. 332). This Markov chain simulation continues until the result can be assumed to be approximately a sample from the stationary distribution (4.5). Note that other methods, such as rejection, are possible for sampling from the normal distribution restricted to $[0,1]$. However, in obtaining our convergence result given in Section 4.4.2, we require that our Markov chain is monotone. For this reason it is necessary to use a sampling method such as the inverse transform method which preserves the order of the uniform random variables in the corresponding samples from the distribution of interest.

4.4.2 The Convergence Result

For x^1, x^2 two configurations of our image space, we use the following distance function

$$d(x^1, x^2) = \sum_{i=1}^{N} |x_i^1 - x_i^2|.$$

The following bound on the convergence time holds for small values of the prior parameter when there are edge effects. If each pixel has the same number of neighbours, for example an image that has been wrapped around, there are no restrictions.

Theorem 4.2 *Consider an image of N pixels, each taking a value in $[0,1]$, and randomly distorted by the addition of $N(0,\sigma^2)$ noise independently at each pixel and restored using the random scan Gibbs sampler algorithm. The distance between the distribution of the Markov chain and its stationary dis-*

tribution will be less than ϵ in the Wasserstein metric at iteration T for

$$T > \frac{\ln\left(\frac{\epsilon}{N}\right)}{\ln\left(\frac{N-1}{N} + \frac{n_{max}}{N}\gamma^2\left(\frac{1}{\sigma^2} + n_{min}\gamma^2\right)^{-1}\right)}$$

for

$$0 < \gamma < \left[\frac{1}{(n_{max} - n_{min})\sigma^2}\right]^{1/2}$$

where n_{max} is the maximum over all pixels of the number of influential neighbouring pixels, n_{min} is the minimum over all pixels of the number of influential neighbouring pixels, and γ is the value of the smoothing parameter from the prior distribution (4.4).

Before we prove this theorem, we will show that coupled realisations of the Markov chain are monotone as defined in Section 2.4.2 and state and prove a lemma about the means of the normal distribution restricted to $[0, 1]$. This allows us to simplify our calculation of the distance between the current states of the chains started in the maximal and in the minimal states.

A partial ordering exists on the state space where one configuration is greater than or equal to another when each corresponding pixel has this ordering, i.e.

$$x \geq y \iff x_i \geq y_i,\ i = 1, \ldots, N.$$

Under this partial order there exists a unique minimal state, $x^{min} = \vec{0}$, and a unique maximal state, $x^{max} = \vec{1}$, all pixels white and all pixels black, respectively.

Lemma 4.2 *Consider two coupled realisations of the Markov chain described in Section 4.4.1. The partial order is preserved after transitions, i.e. if* $x^{t-1,high} \geq x^{t-1,low}$ *then* $x^{t,high} \geq x^{t,low}$*, where* $x^{t,high}$ *is the state of the chain with initial state* x^{high} *at time* t*.*

To prove this lemma we require the following lemma, which we quote in the form given in Roberts and Rosenthal (1999, Lemma 5).

Lemma 4.3 *Suppose that* μ_1 *and* μ_2 *are two probability measures on* $I\!R$*, such that there is a version of the Radon-Nikodym derivative* $R(x) = \mu_2(dx)/\mu_1(dx)$*, which is a non-decreasing function. Suppose also that* f *is a non-decreasing function from* $I\!R$ *into* $I\!R^+$*. Let* $\mathbf{E}_i$*,* $i = 1, 2$ *denote expectations with respect to the two measures* μ_i*,* $i = 1, 2$*. Then for any set* A *for which the following conditional expectations exist,*

$$\mathbf{E}_1[f(X)|X \in A] \leq \mathbf{E}_2[f(X)|X \in A].$$

Proof of Lemma 4.2 Consider two coupled realisations of the Markov chain started in intial states x^{high} and x^{low}. At each iteration one pixel, say the i^{th}, is randomly chosen for updating as a sample from the full conditionals (4.6). For each coupled realisation, we use the same uniform random number and generate the sample using an inverse distribution function approximation. The full conditionals at iteration t are $N(\alpha_i^t, \beta_i^2)$ distributions restricted to $[0, 1]$ where

$$\beta_i^2 = \left(\frac{1}{\sigma^2} + n_i \gamma^2\right)^{-1}$$

is the same for the chains regardless of initial state and

$$\alpha_i^t = \beta^2 \left(\frac{y_i}{\sigma^2} + \gamma^2 \sum_{j \sim i} x_j^{t-1}\right)$$

is different for the two coupled chains. We label the two values $\alpha_i^{t,high}$ and $\alpha_i^{t,low}$. We let $TN(\alpha_i^t, \beta_i^2)$ represent the $N(\alpha_i^t, \beta_i^2)$ distribution restricted to $[0, 1]$.

We will first show that our Markov chain is stochastically monotone. Then updating using the inverse normal distribution function as described in Section 4.4.1 preserves the order in our coupled chains after each update. Our Markov chain is stochastically monotone if, whenever $x^{t-1,high} \geq x^{t-1,low}$,

$$\Pr(x_i^{t,high} > a | x^{t-1,high}) \geq \Pr(x_i^{t,low} > a | x^{t-1,low}) \tag{4.7}$$

for any $a \in \mathbb{R}$ where

$$x_i^{t,high} \sim TN(\alpha_i^{t,high}, \beta_i^2) \quad \text{and} \quad x_i^{t,low} \sim TN(\alpha_i^{t,low}, \beta_i^2).$$

Equivalently, we require that

$$\mathbf{E}\{\mathbf{1}_{\{z_i^{t,high}>a\}}|z_i^{t,high} \in [0,1]\} \geq \mathbf{E}\{\mathbf{1}_{\{z_i^{t,low}>a\}}|z_i^{t,low} \in [0,1]\} \tag{4.8}$$

where

$$z_i^{t,high} \sim N(\alpha_i^{t,high}, \beta_i^2) \quad \text{and} \quad z_i^{t,low} \sim N(\alpha_i^{t,low}, \beta_i^2).$$

If $x^{t-1,high} \geq x^{t-1,low}$, then $\alpha_i^{t,high} \geq \alpha_i^{t,low}$ and

$$R(x) = \exp\left\{-[(\alpha_i^{t,high})^2 - (\alpha_i^{t,low})^2]/(2\beta_i^2)\right\} \exp\left\{(\alpha_i^{t,high} - \alpha_i^{t,low})x/\beta_i^2\right\}$$

is an increasing function. Application of Lemma 4.3 gives (4.8), giving stochastic monotonicity (4.7).

Using the inverse normal distribution function for updating as described in Section 4.4.1, stochastic monotonicity ensures that the partial order is maintained after transitions of the Markov chain. To see this, let ξ be the uniform random number used for updating both chains and suppose site i is the site being updated at iteration t. The new values of the i^{th} pixels are $x^{*,high}$, $x^{*,low}$, chosen to satisfy

$$\Pr(x_i^{t,high} \leq x^{*,high}|x^{t-1,high}) = \xi$$

and

$$\Pr(x_i^{t,low} \leq x^{*,low}|x^{t-1,low}) = \xi.$$

Then by (4.7), $x^{t-1,high} \geq x^{t-1,low}$ results in $x^{t,high} \geq x^{t,low}$. ∎

The following lemma shows that the difference in the mean of normal distributions with the same variance is at least as great as the difference in the mean of the corresponding normal distributions restricted to $[0,1]$. It will be used in the proof of Theorem 4.2.

Lemma 4.4 *Let $e_\beta(\alpha)$ be the mean of the $TN(\alpha, \beta^2)$ distribution, which is the $N(\alpha, \beta^2)$ distribution restricted to $[0,1]$. If $\alpha^{high} \geq \alpha^{low}$, then*

$$e_\beta(\alpha^{high}) - e_\beta(\alpha^{low}) \leq \alpha^{high} - \alpha^{low}.$$

Proof of Lemma 4.4 We will first define the following quantities:

- Let $f_\beta(\alpha) = e_\beta(\alpha) - \alpha$.

- Let $m_\beta(s,t)$ be the mean of the $N(0, \beta^2)$ distribution restricted to $[s,t]$.

- Let $p_\beta(s,t) = \Pr(Z \in [s,t])$ where $Z \sim N(0, \beta^2)$.

Now

$$\begin{aligned} f_\beta(\alpha) &= \int_0^1 x\, e^{-(x-\alpha)^2/(2\beta^2)} dx \Big/ \int_0^1 e^{-(x-\alpha)^2/(2\beta^2)} dx \;-\; \alpha \\ &= \int_{-\alpha}^{1-\alpha} u\, e^{-u^2/(2\beta^2)} du \Big/ \int_{-\alpha}^{1-\alpha} e^{-u^2/(2\beta^2)} du \\ &= m_\beta(-\alpha, 1-\alpha). \end{aligned} \tag{4.9}$$

And if $s < t < u$

$m_\beta(s,u)$

$$
\begin{aligned}
&= \int_s^u x\, e^{-x^2/(2\beta^2)} dx \Big/ \int_s^u e^{-x^2/(2\beta^2)} dx \\
&= \left\{ \int_s^t x\, e^{-x^2/(2\beta^2)} dx + \int_t^u x\, e^{-x^2/(2\beta^2)} dx \right\} \Big/ \int_s^u e^{-x^2/(2\beta^2)} dx \\
&= \left\{ \frac{\int_s^t \frac{1}{\sqrt{2\pi}\beta} e^{-x^2/(2\beta^2)} dx \int_s^t x e^{-x^2/(2\beta^2)} dx}{\int_s^t e^{-x^2/(2\beta^2)} dx} \right. \\
&\qquad \left. + \frac{\int_t^u \frac{1}{\sqrt{2\pi}\beta} e^{-x^2/(2\beta^2)} dx \int_t^u x e^{-x^2/(2\beta^2)} dx}{\int_t^u e^{-x^2/(2\beta^2)} dx} \right\} \Big/ \left\{ \int_s^u \frac{1}{\sqrt{2\pi}\beta} e^{-x^2/(2\beta^2)} dx \right\} \\
&= \frac{p_\beta(s,t) m_\beta(s,t) + p_\beta(t,u) m_\beta(t,u)}{p_\beta(s,t) + p_\beta(t,u)}. \qquad (4.10)
\end{aligned}
$$

Now for $s < t < u$,

$$m_\beta(s,t) \le m_\beta(t,u)$$

and from (4.10)

$$m_\beta(s,t) \le m_\beta(s,u) \le m_\beta(t,u). \qquad (4.11)$$

Now let $s < t < s+1$. Then $-t < -s < -t+1 < -s+1$ and using (4.9) and (4.11) we have

$$f_\beta(t) = m_\beta(-t,-t+1) \le m_\beta(-s,-t+1) \le m_\beta(-s,-s+1) = f_\beta(s).$$

Thus

$$e_\beta(t) - t \le e_\beta(s) - s$$

i.e.

$$e_\beta(t) - e_\beta(s) \le t - s.$$

Since this holds whenever $s < t < s + 1$, it also holds for any $s < t$ by breaking the interval $[s, t]$ up into sub-intervals whose width is smaller than one. ■

Proof of Theorem 4.2 It suffices to consider the number of iterations until coupled chains started in the maximal and minimal states have converged to within ϵ tolerance in the Wasserstein metric. This ensures convergence of a chain started in any other state to the stationary distribution. To see this, suppose the state x is a sample from the stationary distribution π and x^0 is any other state. Then

$$\begin{aligned}
&d_W(P^t\left(x^0, \cdot\right), \pi(\cdot)) \\
&\quad = d_W\left(P^t(x^0, \cdot), \int_{\mathcal{X}} P^t(x, \cdot)\pi(dx)\right) \\
&\quad \leq \int_{\mathcal{X}} d_W\left(P^t(x^0, \cdot),\ P^t(x, \cdot)\right)\pi(dx) \\
&\quad \leq \int_{\mathcal{X}} d_W\left(P^t(x^{max}, \cdot),\ P^t(x^{min}, \cdot)\right)\pi(dx) \\
&\quad = d_W(\mathcal{L}(x^{t,max}, \mathcal{L}(x^{t,min}))
\end{aligned}$$

where the first inequality is the triangle inequality and the second follows from monotonicity of the coupled chains and the definition of the Wasserstein metric (4.1). Note also that $d(x^{max}, x^{min}) = \mathrm{diam}\boldsymbol{\mathcal{X}}$.

Then to get our result, we need to find a constant $c \in (0,1)$ so that

$$\mathbf{E}\left[d(x^{t+1,max}, x^{t+1,min})|x^{t,max}, x^{t,min}\right] \leq c\, d(x^{t,max}, x^{t,min}).$$

Now

$$\begin{aligned}
&\mathbf{E}\left[d(x^{t+1,max}, x^{t+1,min})|x^{t,max}, x^{t,min}\right] \\
&= \sum_{i=1}^{N} \mathbf{E}\left[x_i^{t+1,max} - x_i^{t+1,min}|x^{t,max}, x^{t,min}\right] \\
&= \sum_{i=1}^{N} \left\{ \frac{N-1}{N}(x_i^{t,max} - x_i^{t,min}) + \frac{1}{N}\left[e_{\beta_i}(\alpha_i^{t+1,max}) - e_{\beta_i}(\alpha_i^{t+1,min})\right] \right\} \\
&\leq \sum_{i=1}^{N} \left\{ \frac{N-1}{N}(x_i^{t,max} - x_i^{t,min}) \right. \\
&\qquad + \frac{1}{N}\left[\left(\frac{1}{2\sigma^2} + n_i\frac{\gamma^2}{2}\right)^{-1} \left(\frac{y_i}{2\sigma^2} + \frac{\gamma^2}{2}\sum_{j\sim i} x_j^{t,max}\right) \right. \\
&\qquad\qquad \left.\left. - \left(\frac{1}{2\sigma^2} + n_i\frac{\gamma^2}{2}\right)^{-1} \left(\frac{y_i}{2\sigma^2} + \frac{\gamma^2}{2}\sum_{j\sim i} x_j^{t,min}\right) \right] \right\} \\
&\leq \sum_{i=1}^{N} \left\{ \frac{N-1}{N}(x_i^{t,max} - x_i^{t,min}) \right. \\
&\qquad \left. + \frac{1}{N}\gamma^2 \left(\frac{1}{\sigma^2} + n_{min}\gamma^2\right)^{-1} \sum_{j\sim i}(x_j^{t,max} - x_j^{t,min}) \right\}
\end{aligned}$$

where the first inequality uses Lemma 4.4. In the expression

$$\sum_{i=1}^{n}\sum_{j\sim i}(x_j^{t,max} - x_j^{t,min})$$

the difference in the values of the k^{th} pixel appears n_k times, once for each pixel i that it neighbours. Thus

$$
\begin{aligned}
&\mathbf{E}\left[d(x^{t+1,max}, x^{t+1,min})|x^{t,max}, x^{t,min}\right] \\
&= \sum_{i=1}^{N}\left\{\frac{N-1}{N}\left(x_i^{t,max} - x_i^{t,min}\right)\right\} \\
&\qquad + \frac{\gamma^2}{N}\left(\frac{1}{\sigma^2} + n_{min}\gamma^2\right)^{-1}\sum_{k=1}^{N}\left\{n_k(x_k^{t,max} - x_k^{t,min})\right\} \\
&\leq \sum_{i=1}^{N}\left\{\frac{N-1}{N}\left(x_i^{t,max} - x_i^{t,min}\right)\right\} \\
&\qquad + \frac{\gamma^2}{N}\left(\frac{1}{\sigma^2} + n_{min}\gamma^2\right)^{-1} n_{max}\sum_{k=1}^{N}(x_k^{t,max} - x_k^{t,min}) \\
&= \left[\frac{N-1}{N} + \frac{n_{max}}{N}\gamma^2\left(\frac{1}{\sigma^2} + n_{min}\gamma^2\right)^{-1}\right] d(x^{t,max}, x^{t,min}).
\end{aligned}
$$

The coefficient of the right side is less than 1 for

$$n_{max}\gamma^2\left(\frac{1}{\sigma^2} + n_{min}\gamma^2\right)^{-1} < 1 \quad \blacksquare$$

If $n_{max} = n_{min}$, i.e. each pixel has the same number of neighbours so there are no edge effects, the result in the theorem holds for all values of γ.

In the limit $\sigma \to \infty$, we have no information from the observed image and no result holds (the upper limit for the range of γ is 0). If $\sigma = 0$, the result holds for all γ.

The fourth column of Table 4.1 shows the theoretical results for various values of N, ϵ, n_{max}, n_{min}, γ and σ. Values are compared to a 32×32 grid with $\epsilon = 0.1$, $n_{max} = 4$, $n_{min} = 2$, $\gamma = 1$, and $\sigma = 0.2$. The required number of iterations until convergence, T, can be seen to vary with the size of the image, N, and the value of the prior smoothing parameter, γ, but varies little with the other parameters.

4.4.3 Results from Simulations

The final column of Table 4.1 gives the number of iterations that were required for the Markov chains started in the maximal and minimal states to come within ϵ for several different simulations. Simulations were written in C. Samples from the truncated normal distribution were obtained using the inverse normal distribution function approximation of Thisted (1988, p. 332) with the adjustment for restricted sampling given by Fishman (1996, p. 152). In each case, the original image consisted of four overlapping rectangles, in black, white, and two shades of grey. While these simulations to approximate coalescence do not measure the same quantity as our metric, the closeness of the simulated coalescence times to the theoretical number of iterations required, with the exception of the case where the value of γ is close to our upper limit, is reassuring. Note that one simulation (the first listed) required more iterations than the theoretical value; our bound is on the number of iterations required until the mean distance is less than ϵ, and does not guarantee the distance on any one realisation will be that small. The simulation of 1000 restorations with $\gamma = 1$ described in the next paragraph gives a better indication that our theory is giving a tight bound on the actual convergence time of the Wasserstein metric.

We simulated 1000 restorations of our 32×32 pixels image, distorted by normal noise with standard deviation 0.2. Each simulation consisted of two realisations of our Markov chain, started in the maximal and minimal states. We set the prior smoothing parameter γ at 1.0 and the neighbourhood structure such that interior pixels were influenced by their four nearest neighbours. These simulations were each run for 11096 iterations, the number our theory requires for convergence to within $\epsilon = 0.1$ precision in the Wasserstein metric. For each simulation, the actual distance between the states of two Markov chains after iteration 11096 was recorded. The distribution of these distances was right-skewed, with a mean of 0.07769, as compared to the precision of 0.1 that we requested in our determination of the number of runs required. While most of the distances were below this, there were many high values, including 12 values greater than 1.0. Approximately one-fifth of the simulations (216 of 1000) had not coupled to within 0.1 precision.

N	ϵ	n_{max}, n_{min}	γ	σ	Theoretical value of T	Number of iterations required in simulation
32×32	0.1	4, 2	1	0.2	11096	11159
10×10					808	681
256×128					488513	461421
1024×768					14658270	14088943
32×32	0.1	4, 2	1	0.2	11096	10423
	0.01				13863	12354
	1				8329	7641
32×32	0.1	4, 2	1	0.2	11096	10599
		2, 1			10240	9054
		8, 3			13234	12134
32×32	0.1	4, 2	1	0.2	11096	9531
			0.1		9467	7054
			3		58081	23019
32×32	0.1	4, 2	1	0.2	11096	10928
				0.1	9838	7717
				0.3	13603	12084

Table 4.1: Convergence times for the restoration of a grey-scale image.

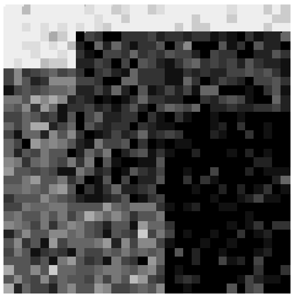
Figure 4.1 gives an example of the image restoration process. Figure 4.1(a) shows the true image, drawn on a 32×32 grid. It was randomly degraded with $N(0, 0.15^2)$ noise, added independently to each pixel. The degraded image is shown in Figure 4.1(b). In the prior distribution, γ was set at 4 and neighbours were considered to be the pixels above, below, and beside ($n_{max} = 4$ and $n_{min} = 2$). The specified accuracy was $\epsilon = 0.1$. Note that this error is distributed across 1024 pixels. The algorithm was run for 58081 iterations, the number that our theory requires for convergence to within ϵ accuracy, and our initial state was every pixel black. Figure 4.1(c) shows the mean of 10 independent samples from the posterior distribution.

4.5 Results for the Restoration of a Binary Image

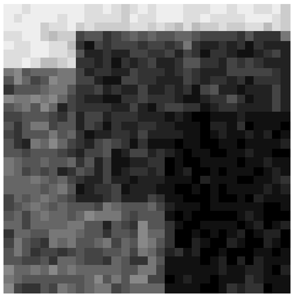
The case where the image is a grid of binary pixels was considered in Chapter 3. Consider a configuration $\{x_i\}$ of N pixels which take on values $+1$ or -1. The following prior distribution, the Ising model, assigns greater probability to configurations where neighbouring pixels are alike

$$\pi_\beta(x) = \frac{1}{Z} \exp\left\{ -\beta \sum_{<i,j>} x_i x_j \right\} \tag{4.12}$$

(a)



(b)



(c)

Figure 4.1: A simulated restoration of a 32×32 image. (a) true image; (b) observed image; (c) the mean of 10 independent samples from the posterior distribution.

where the sum is taken over pairs of sites (i, j) which are nearest neighbours, β is a positive parameter, and Z is the normalising constant.

In Chapter 3, precise $O(N^2)$ bounds were found on the convergence time of the Gibbs sampler for sampling from this prior, and for sampling from the posterior model obtained by combining this prior with observed data. The observed data are distorted images obtained from the true image by random distortion mechanisms. Results were found for two distortion mechanisms: additive normal noise and random flips. Using coupling, upper bounds were obtained on the necessary number of iterations until total variation distance is less than ϵ of the form

$$\frac{2eN^2}{f(\beta)}(1 - \log \epsilon)$$

where $f(\beta)$ is an easily evaluated, known positive function of the parameter of the prior distribution. These bounds hold for values of β below a given threshold. The function $f(\beta)$ differs with the distortion mechanism.

The distance function considered in Chapter 3 was the number of pixels which differ in the two chains. The relationship of the function $f(\beta)$ to the change in the distance function allows us to apply the calculated values of $f(\beta)$ from Chapter 3 to get a result for convergence in the Wasserstein metric. The theorem that follows gives these results for the three cases considered in

Chapter 3.

Theorem 4.3 *Consider the use of the random scan Gibbs sampler for the restoration of an image of N pixels in arbitrary dimension where each pixel can take on the values $\{-1, +1\}$ and is influenced by its n nearest neighbours. The algorithm will have converged in the sense that the Wasserstein metric will be less than ϵ at time T for*

$$T > \frac{\ln\left(\frac{\epsilon}{N}\right)}{\ln\left(1 - \frac{f(\beta)}{N}\right)} \tag{4.13}$$

where $f(\beta)$ and the possible values of β, the parameter from the Ising model prior, are given as follows for three cases:

1. *The distribution of interest is the Ising model, Equation (4.12). Then*

$$f(\beta) = \frac{(n+2)e^{-n\beta} - ne^{n\beta}}{e^{n\beta} + e^{-n\beta}}$$

for

$$0 \leq \beta \leq \frac{1}{2n} \log\left(\frac{n+2}{n}\right).$$

2. *The distribution of interest is the posterior distribution with Ising model prior and the observed image is modelled as the true image with each pixel incorrectly observed with probability α. Then*

$$f(\beta) = \frac{k_\alpha - ne^{2n\beta} + (n+2)e^{-2n\beta}}{k_\alpha + e^{2n\beta} + e^{-2n\beta}}$$

for

$$0 \leq \beta \leq \frac{1}{2n} \log \left[\frac{k_\alpha + \{k_\alpha^2 + 4n(n+2)\}^{1/2}}{2n} \right]$$

where $k_\alpha = (1-\alpha)/\alpha + \alpha/(1-\alpha)$.

3. *The distribution of interest is the posterior distribution with Ising model prior and the observed image is modelled as the true image with* $N(0, \sigma^2)$ *noise added independently to each pixel. Then*

$$f(\beta) = \frac{k_{\sigma, y_{min}} + (n+2)e^{-2n\beta} - ne^{2n\beta}}{k_{\sigma, y_{min}} + e^{2n\beta} + e^{-2n\beta}}$$

for

$$0 \leq \beta \leq \frac{1}{2n} \log \left(\frac{k_{\sigma, y_{min}} + \sqrt{k_{\sigma, y_{min}}^2 + 4n(n+2)}}{2n} \right)$$

where y_{min} *is the value of the smallest (in absolute value) observed pixel and* $k_{\sigma, y_{min}} = e^{2y_{min}/\sigma^2} + e^{-2y_{min}/\sigma^2}$.

Proof In the proofs of Theorems 3.3, 3.4, and 3.5, it was shown how the expected change in the distance function in one iteration can be expressed as the product of $-\frac{f(\beta)}{N}$ and the current distance. Thus,

$$\begin{aligned} &\mathbf{E}[d(X_{t+1}^{max}, X_{t+1}^{min}) | X_t^{max}, X_t^{min}] \\ &= d(X_t^{max}, X_t^{min}) \\ &\quad + \mathbf{E}[d(X_{t+1}^{max}, X_{t+1}^{min}) - d(X_t^{max}, X_t^{min}) | X_t^{max}, X_t^{min}] \end{aligned} \tag{4.14}$$

$$= \quad d(X_t^{max}, X_t^{min}) - \frac{f(\beta)}{N} \cdot d(X_t^{max}, X_t^{min}).$$

This expression can then be applied with Theorem 4.1 to give (4.13). The expressions for $f(\beta)$ and the values of β which ensure that $f(\beta) > 0$ for the three cases are calculated in the proofs of Theorems 3.3, 3.4, and 3.5, respectively. ■

By Taylor series expansion of the above results, it is easily seen that the bounds on the convergence times are $O(N \ln N)$.

This result also gives an upper bound on the time until total variation distance is less than ϵ because of the following relationship.

Proposition 4.2 *On a finite set $\mathcal{X}$, the following relationship exists between total variation distance and the Wasserstein metric:*

$$d_{min}\, d_{TV} \;\leq\; d_W$$

where $d_{min} = \min d(x, y)$ where the minimum is taken over all possible pairs of distinct points x, y in $\mathcal{X}$.

Proof For μ, ν two measures on $\mathcal{X}$ and X, Y random variables with $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$,

$$d_W(\mu, \nu) = \inf_{X,Y}\{\mathbf{E}[d(X, Y)]\}$$

and

$$d_{TV}(\mu, \nu) = \inf_{X,Y}\{\Pr(X \neq Y)\} = \inf_{X,Y}\{\mathbf{E}[\mathbf{1}_{\{X\neq Y\}}]\}.$$

But

$$d(X, Y) \geq d_{min}\, \mathbf{1}_{\{X\neq Y\}}. \quad \blacksquare$$

We can now get the following bounds for convergence in total variation distance.

Corollary 4.1 *The random scan Gibbs sampler algorithms for sampling from the models described in Theorem 4.3 will have converged in the sense that the total variation distance will be less than ϵ at time T for T greater than the values given in each of the three cases below:*

1. *The distribution of interest is the Ising model, Equation (4.12). Then*

$$T > \ln\left(\frac{\epsilon}{N}\right) / \ln\left(1 - \frac{1}{N}\frac{(n+2)e^{-n\beta} - ne^{n\beta}}{e^{n\beta} + e^{-n\beta}}\right)$$

   for

$$0 \leq \beta \leq \frac{1}{2n}\log\left(\frac{n+2}{n}\right).$$

2. *The distribution of interest is the posterior distribution with Ising model prior and the observed image is modelled as the true image with each pixel incorrectly observed with probability α. Then*

$$T > \ln\left(\frac{\epsilon}{N}\right) / \ln\left(1 - \frac{1}{N}\frac{k_\alpha - ne^{2n\beta} + (n+2)e^{-2n\beta}}{k_\alpha + e^{2n\beta} + e^{-2n\beta}}\right)$$

for

$$0 \leq \beta \leq \frac{1}{2n} \log \left[\frac{k_\alpha + \{k_\alpha^2 + 4n(n+2)\}^{1/2}}{2n} \right]$$

where $k_\alpha = (1-\alpha)/\alpha + \alpha/(1-\alpha)$.

3. *The distribution of interest is the posterior distribution with Ising model prior and the observed image is modelled as the true image with* $N(0, \sigma^2)$ *noise added independently to each pixel. Then*

$$T > \ln\left(\frac{\epsilon}{N}\right) / \ln\left(1 - \frac{1}{N} \frac{k_{\sigma,y_{min}} + (n+2)e^{-2n\beta} - ne^{2n\beta}}{k_{\sigma,y_{min}} + e^{2n\beta} + e^{-2n\beta}}\right)$$

for

$$0 \leq \beta \leq \frac{1}{2n} \log \left(\frac{k_{\sigma,y_{min}} + \sqrt{k_{\sigma,y_{min}}^2 + 4n(n+2)}}{2n} \right)$$

where y_{min} *is the value of the smallest (in absolute value) observed pixel and* $k_{\sigma,y_{min}} = e^{2y_{min}/\sigma^2} + e^{-2y_{min}/\sigma^2}$.

Proof Our distance function $d(X, Y)$ is the number of sites where X, Y differ. Thus $d_{min} = 1$ and the Wasserstein results of Theorem 4.3 give immediate upper bounds on convergence in total variation distance. ■

As an example, in the case of normal noise, for a 32×32 grid with pixels influenced by their 4 nearest neighbours, if $\sigma = 0.3$, $y_{min} = 0.1$, $\beta = 0.1$, and $\epsilon = 0.1$, the number of iterations required for convergence in both the

Wasserstein metric and total variation distance is is 36281. In Chapter 3 our upper bound on the convergence time was 72246394.

4.6 Application to Exact Sampling

Exact sampling algorithms (Propp and Wilson 1996, Fill 1998) have recently generated a great deal of interest in the Markov chain Monte Carlo literature. In particular, the algorithm of Propp and Wilson (1996) involving the concept of coupling-from-the-past has been applied and extended to a number of different applications. For a monotone Markov chain for which there exist unique maximal and minimal elements, the algorithm involves running two realisations of the Markov chain started in each of these states from time $-t$ forward. If the chains have coupled at time 0 the resulting state is exactly a sample from the distribution of interest.

If the algorithm is monotone as defined in Section 2.4.2 and the distance between maximal and minimal states is equivalent to the diameter of the space, our theory for convergence in the Wasserstein metric can be applied to give a bound on the expected running time of the coupling-from-the-past algorithm. As discussed in Chapter 3 and Section 4.4, the image processing examples in this chapter are examples of monotone Markov chains.

Our theory for convergence in the Wasserstein metric assumes that coupling of the maximal and minimal states of our monotone chain to within ϵ tolerance is adequate. This idea was considered by Møller (1999) in the context of applying exact sampling algorithms using coupling-from-the-past to continuous state spaces. Møller considered chains which may not have a maximal state, but for which a dominating chain can be constructed. Møller's coupling-from-the-past algorithm requires the dominating chain to come within ϵ of the Markov chain started in the minimal state, where ϵ is the accuracy specified by the user. He showed that this algorithm gives an exact sample from the stationary distribution to within ϵ accuracy in a finite time.

The distribution of the time to couple into the future is the same as the distribution of the smallest t such that chains started at time $-t$ will have coupled at time 0 (Propp and Wilson 1996). Thus our results can be used to give an indication of how far in the past it is necessary to go back to achieve approximate coalescence at time 0 in Møller's algorithm.

The binary image restoration algorithm of Section 4.5 is an example of a finite state space, monotone chain, so the coupling from the past algorithm of Propp and Wilson (1996) can be immediately applied. The bounds of Theorem 4.3 give an indication of the required starting time in the past necessary to achieve coalescence at time 0.

For uniformly ergodic Markov chains, it is possible to achieve exact coalescence on continuous state spaces by the application of the multigamma coupler of Murdoch and Green (1998), or, as indicated by Møller (1999) following a suggestion by Duncan Murdoch, a hybrid of the multigamma coupler and Møller's algorithm. Another approach to achieving exact coalescence on continuous state spaces involves the insertion every k^{th} (for example, $k = 10$) iteration of a Metropolis step with updates of the form suggested by Neal (1999). In this step, a hyper-grid is placed on the state space with its origin determined randomly. The length of the sides of the hyper-boxes of the grid is fixed. The proposal state for the Markov chain is the centre of the box containing the current state. If the current states of two chains are in the same box, they will have the same proposed new state, and a positive probability of exact coalescence.

Chapter 5

Using Auxiliary Simulation to Approximate Theoretical Convergence Rates

5.1 Introduction

In this chapter, we examine how auxiliary simulation can be used to find approximate values for the parameters of the Markov chain Monte Carlo algorithms that must be calculated in order to apply our theoretical convergence results. In particular, we use auxiliary simulation to calculate the parameter c from our result for convergence in the Wasserstein metric as described in Chapter 4. The same approach can be used to calculate the parameter a used in Chapter 3 for our convergence results in total variation distance.

Recall Theorem 4.1. Suppose $\{X_t\}$, $\{Y_t\}$ are two coupled realisations of a Markov chain on a bounded state space $\mathcal{X}$. If we can find a constant $c \in (0, 1)$ such that

$$\mathbf{E}[d(X_{t+1}, Y_{t+1})|X_t, Y_t] \leq c\, d(X_t, Y_t)$$

for all t, then a Markov chain started in any initial state will have converged in the sense that the Wasserstein metric between the distribution of its state at time T and the stationary distribution will be less than ϵ for

$$T > \frac{\ln\left(\frac{\epsilon}{\operatorname{diam}(\mathcal{X})}\right)}{\ln c}$$

where $\operatorname{diam}(\mathcal{X}) = \sup_{x,y \in \mathcal{X}} d(x, y)$.

Calculating the value of c is the difficult part of applying this theorem. In Chapter 4 we calculated it for an example where the Gibbs sampler was used and the full conditional distributions were normal distributions truncated to $[0, 1]$. We also showed how calculations in Chapter 3 can be applied to calculate c for an example using the Gibbs sampler where each component can take only two values. For more complicated algorithms, it may be prohibitively difficult or perhaps impossible to calculate c analytically.

As will be demonstrated, our auxiliary simulation approach gives a reasonable estimate for c for an example for which we have an analytic value. However, it can not provide the guarantees that a calculated value would. Our aim is to bridge the gap between the theoretical results of Chapter 4, which may be difficult to apply to complex models, and what is reasonable to carry out in practice.

A similar approach to bridging the gap between theory and practice is given in Cowles and Rosenthal (1998). In that paper, Cowles and Rosenthal describe how auxilary simulation can be used to verify the conditions and estimate the parameters for the theoretical results in Rosenthal (1995b). Their work is extended and refined in Cowles (1999) in the context of hierarchical normal linear models.

It should be noted that care must be taken in applying the results of auxiliary simulation to calculate the convergence time. In order to find the true value of c, we must find the supremum of the value described in the next section over all pairs of initial states. It must be recognised that some part of the state space may have been missed in the choice of initial states for which auxiliary simulations are carried out. However, our example suggests that our approach works reasonably well. Another limitation of our method is that it is very computer intensive.

An advantage of this method is that, unlike convergence diagnostics, the auxiliary simulations we are performing to estimate the convergence time do not bias the final results of the Markov chain Monte Carlo algorithm which is run independently using the number of iterations suggested by the auxiliary simulation.

In Section 5.2 we outline the method we recommend for carrying out simulations to estimate c. In Section 5.3 we estimate c for the grey-scale model from Section 4.4 which we compare with our calculated upper bound. Other examples including models with other prior distributions for which we have no analytic results are being investigated and will appear in future work.

5.2 Suggested Approach to Obtaining an Estimate of c by Auxiliary Simulation

We want to estimate the value of c, the maximum over pairs of states x, y of

$$c_{x,y} = \frac{\mathbf{E}[d(X_{t+1}, Y_{t+1})|X_t = x, Y_t = y)]}{d(x,y)}.$$

Simulations to estimate $c_{x,y}$ should be carried out for a variety of states x and y. For example, we can generate x and y randomly pixel by pixel. We should also consider pairs of states that are very close together, very far apart, and states that are highly probable *a priori*.

We recommend a two-stage approach. In the first stage, a variety of pairs of initial states is explored, with the goal of identifying which initial states lead to the largest value of $c_{x,y}$. In the second stage, we generate more values from the identified states in order to get an error estimate.

Exploratory step: From each pair of initial states, simulate a number, N, of one-step iterations. The estimate of $c_{x,y}$ is the mean of the N ratios of the distance apart after the iteration to the distance between x and y. N should be chosen so that the standard error in $c_{x,y}$ is appropriately small. Note that for random scan algorithms, a better estimate of $c_{x,y}$ can be obtained by updating each pixel N_1 times, with N_1 chosen so that the variance of the mean distance ratio for each pixel is as small as desired. The estimate of $c_{x,y}$ is then the average over pixels of the average distance ratio per pixel. Any one estimate for c may over-estimate its true value because it is possible for pixels to get further apart as well as closer together. However, it is not desirable to over-estimate c by a large amount as an overly conservative estimate will indicate that an unreasonably large number of iterations is necessary to achieve convergence.

Error estimate: An estimate of the error in c can be achieved as follows. Focus on the initial states which gave the largest estimates for $c_{x,y}$ in our exploratory step. Randomly generate N_2 pairs of initial states and take the maximum of the corresponding values of $c_{x,y}$; this maximum is an estimate for c. Do this N_3 times. The N_3 estimated values of c generated in this manner are independent and identically distributed. Take as our estimate of c the upper limit of the 95% confidence interval of the mean of these estimates. When rounding it is appropriate to be conservative and always round up.

5.3 Example

5.3.1 The Grey-Scale Image Restoration Problem with Quadratic Difference Prior

We now use auxiliary simulation to estimate the value of c for which we calculated analytically an upper bound in Section 4.4. We set the values of the parameters as $\gamma = 1$, $\sigma = 0.2$, and $N = 1024$. Our image is two-dimensional with a neighbourhood structure where each pixel has as neighbours the pixels above, below and beside. The analytically obtained upper bound for c in this case is 0.99917368. The observed image used in this simulation is the same set of overlapping blocks used in Section 4.4.

For the exploratory stage, we calculated $c_{x,y}$ for the following pairs of states:

- all black and all white (the maximal and minimal states)
- a solid-coloured grey square surrounded by a lighter shade of grey and a lighter version of the same image; these states have high prior probability
- randomly generated independent pixels

- states that are close together generated by:
  - starting from all black and all white, running two coupled realisations of the Markov chain for 8500 iterations and using the states at that point as the states x, y
  - randomly generating an initial state for x and then changing the value of one randomly selected pixel for y
  - using a high prior probability state for x and changing the value of one randomly selected pixel for y

In each of these cases, it was necessary to simulate $N_1 = 100$ or 1000 iterations per pixel in order to estimate the mean ratio for that pixel with standard error at most 10^{-5}.

The largest values of $c_{x,y}$ were obtained when the states x, y were close together, so we use these for our estimate. For each of the three categories of states that are close together as described above, we generated the maximum of the values of $c_{x,y}$ for $N_2 = 100$ pairs of initial states $N_3 = 100$ times. The confidence intervals for the mean of these i.i.d. maxima are $(0.999158032748, 0.999158037252)$, $(0.9991479733, 0.9991485927)$, and $(0.9991595904, 0.9991598076)$. Thus we take as our estimate of c 0.99916.

Based on this estimate, and setting our tolerance for convergence in the Wasserstein metric to be $\epsilon = 0.1$, we conclude that 10989 iterations are required for convergence. For comparison, our theoretically obtained upper bound for c leads to a requirement of 11171 iterations, or 1.7% more iterations. It should be noted that our theoretical result is an upper bound on the true convergence time that may be overly conservative.

Chapter 6

Probability Metrics

6.1 Introduction

Studying the convergence of Markov chain Monte Carlo algorithms to their stationary distributions requires a choice of a *probability metric* to measure that convergence. There are a host of metrics available to quantify the distance between probability measures, each with particular properties which make them theoretically interesting, or useful in some applications. In this chapter, we collect in one place some of the most widely used metrics and summarise the known relationships between them in a handy reference table. We also provide some new bounds between several of the metrics.

An encyclopedic and dense account of probability metrics is given by Rachev (1991), and we do not intend to duplicate his account here. By contrast, this chapter is limited to nine chosen metrics. Eight appear often in accounts of probability metrics; the ninth, the discrepancy metric, is less well-known but is included because of its applicability to problems in which other metrics are not suitable.

We limit ourselves to metrics between probability measures (*simple metrics*) rather than the broader context of metrics between random variables (*compound metrics*).

This chapter is organized as follows. Section 6.2 lists metrics in wide use among probabilists and statisticians. Section 6.3 discusses bounds between them. Some examples of their applications are described in Section 6.4.

6.2 Probability Metrics

Throughout this chapter, let Ω be a complete separable metric space, and let $\mathcal{B}$ be the Borel σ-algebra on Ω. In Markov chain Monte Carlo applications, Ω is the state space of the Markov chain. Let $\mathcal{M}$ be the space of all probability measures on $(\Omega, \mathcal{B})$. We consider convergence in $\mathcal{M}$ under various notions of distance. Some of these are not metrics, but are non-negative notions of "distance" between probability distributions on Ω that are often encountered in practice.

In what follows, let μ, ν be two probability measures on Ω. Let f and g be their corresponding density functions (when they exist) with respect to an arbitrary dominating measure λ. If $\Omega = \mathbb{R}$, let F, G be the corresponding distribution functions. When needed, X, Y will denote random variables on Ω such that $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$.

Total variation distance

1. State space: Ω any measurable space.

2. Definition:

$$\begin{aligned} d_{TV}(\mu,\nu) &:= \sup_{A\subset\Omega} |\ \mu(A) - \nu(A)\ | \\ &= \frac{1}{2}\max_{|h|\leq 1}\left|\int h\ d\mu - \int h\ d\nu\right| \end{aligned}$$

   where $h : \Omega \to \mathbb{R}$ satisfies $|h(x)| \leq 1$. For a countable state space Ω, the definition above becomes

$$d_{TV} = \frac{1}{2}\sum_{x\in\Omega} |\ \mu(x) - \nu(x)\ |$$

which is half the L^1 norm between the two measures. Some authors (for example, Tierney (1996)) define total variation distance as twice our definition.

3. Note that when Ω is a continuous state space, the total variation distance is often not suitable, since the distance between a discrete and a continuous probability measure is 1.

4. Total variation distance does not metrize weak convergence.

5. Total variation distance has a coupling characterisation:

$$\begin{aligned} d_{TV}(\mu,\nu) &= \inf\{\Pr(X \neq Y) : \text{ r.v. } X,\ Y \text{ s.t. } \mathcal{L}(X)=\mu,\ \mathcal{L}(Y)=\nu\} \\ &= \inf\{\mathbf{E}[\mathbf{1}_{X\neq Y}]\} \end{aligned}$$

(Lindvall 1992, p.19).

Uniform (or Kolgomorov) metric

1. State space: $\Omega = \mathbb{R}$.

2. Definition:

$$d_U(F,G) = \sup_x |F(x) - G(x)|, \quad x \in \mathbb{R}$$

3. The Uniform metric does not metrize weak convergence.

Lévy metric

1. State space: $\Omega = \mathbb{R}$.

2. Definition:

$$d_L(F, G) = \inf\{\epsilon > 0 : G(x - \epsilon) - \epsilon \leq F(x) \leq G(x + \epsilon) + \epsilon, x \in \mathbb{R}\}.$$

3. The Lévy metric metrizes weak convergence:

4. We can now define an ϵ-neighbourhood of F: $N_\epsilon(F) = \{G : d_L(F, G) \leq \epsilon\}$.

5. This metric depends on the metric on $\mathbb{R}$ and is not scale-invariant.

Prokhorov metric

1. State space: Ω any measurable *metric* space. (This is the analogue of the Lévy metric for arbitrary spaces.)

2. Definition:

$$d_P(\mu, \nu) = \inf\{\epsilon > 0 : \mu(B) \leq \nu(B^\epsilon) + \epsilon \text{ for all Borel sets B }\}$$

   where $B^\epsilon = \{x : \inf_{y \in B} d(x, y) \leq \epsilon\}$.

3. This metric is not scale-invariant and depends on the metric of Ω. It is possible to show that this metric is symmetric in μ, ν (Huber 1981).

Hellinger metric

1. State space: Ω any measurable space.

2. Definition: the measures μ, ν must have densities f, g with respect to λ:

$$d_H^2(\mu, \nu) = \int_\Omega (\sqrt{f} - \sqrt{g})^2 \, d\lambda = 2\left(1 - \int_\Omega \sqrt{fg} \, d\lambda\right)$$

   Note: different texts refer to different versions of this metric. We follow Zolotarev (1983).

   If Ω is a countable space this reduces to

$$d_H^2(\mu, \nu) = \sum_{\omega \in \Omega} \left(\sqrt{\mu(\omega)} - \sqrt{\nu(\omega)}\right)^2$$

   (Diaconis and Zabell 1982).

3. Can be "factored" in terms of marginals (see Zolotarev (1983, p.279)). This makes it possible to express the distance between distributions of vectors with independent components in terms of the distances between the distributions of the corresponding components.

4. This metric does not depend on any metric of Ω.

Wasserstein and Kantorovich metrics

1. State space: $\mathbb{R}$ or any measurable metric space.

2. Definition: For $\Omega = \mathbb{R}$, the Kantorovich metric is defined by

$$\begin{aligned} d_K(\mu, \nu) &= \int_{-\infty}^{\infty} |F(x) - G(x)| dx \\ &= \int_0^1 |F^{-1}(t) - G^{-1}(t)| dt. \end{aligned}$$

   For any separable metric space, this is equivalent to

$$d_K(\mu, \nu) = \sup \left\{ \left| \int h \, d\mu - \int h \, d\nu \right| : \|h\|_L \leq 1 \right\}, \tag{6.1}$$

   the supremum being taken over all h satisfying the Lipschitz condition

$$|h(x) - h(y)| \leq d(x, y).$$

3. This metric metrizes weak convergence.

4. This metric is not scale-invariant and it depends on the metric of Ω through the Lipschitz condition.

5. By the Kantorovich-Rubinstein theorem, the Kantorovich metric is equal to the Wasserstein metric:

$$d_W(\mu, \nu) = \inf_J \{\mathbf{E}[d(X, Y)] : \mathcal{L}(X) = \mu, \mathcal{L}(Y) = \nu\},$$

where the infimum is taken over all joint distributions J with marginals μ, ν. See Szulga (1982, Theorem 2).

Relative entropy (or Kullback-Leibler separation or divergence)

1. State space: Ω any measurable space.

2. Definition: if the measures μ, ν have densities f, g with respect to λ:

$$d_I(\mu, \nu) = \int_\Omega f \log(f/g)\ d\lambda.$$

   For Ω a countable space:

$$d_I(\mu, \nu) = \sum_{\omega \in \Omega} \mu(\omega) \log\left(\frac{\mu(\omega)}{\nu(\omega)}\right).$$

3. This is not a metric, since it is not symmetric and does not satisfy the triangle inequality. However, it has many useful properties, such as being additive for independent processes (useful for product spaces).

χ^2 distance

1. State space: Ω any measurable space.

2. Definition: if the measures μ, ν have densities f, g with respect to λ:

$$d_{\chi^2}(\mu, \nu) = \int_\Omega \frac{(f-g)^2}{g}\, d\lambda.$$

For a countable space Ω this reduces to:

$$d_{\chi^2}(\mu, \nu) = \sum_{\omega \in \Omega} \frac{(\mu(\omega) - \nu(\omega))^2}{\mu(\omega)}.$$

Note: Reiss (1989, p.98) defines the χ^2 distance as the square root of the above expression.

Discrepancy

1. State space: Ω any measurable metric space.

2. Definition:

$$d_D(\mu, \nu) = \sup_{\text{all closed balls } B} |\mu(B) - \nu(B)|.$$

3. Although it depends on the metric of Ω, this definition is scale-invariant and does not depend on the "size" of the space.

6.3 Some Relationships Between Probability Metrics

Figure 6.1 is a diagram of the relationships between various probability metrics considered in this chapter. Some of the relationships only hold on some state spaces, or for a restricted class of metrics. An arrow from metric A to metric B indicates that an upper bound exists for A in terms of B. The annotations involving functions of x indicate the nature of the relationship between the two metrics. An x by itself indicates the relationship is direct, i.e. $d_A \leq d_B$. An expression involving x is the function of the bounding metric, B, that gives an upper bound on A. For example, if the arrow from A to B is annotated with $\sqrt{x/2}$, then

$$d_A \leq \sqrt{d_B/2}.$$

The diameter of the space is given by $\operatorname{diam}\Omega = \sup_{x,y\in\Omega} d(x,y)$; the results involving $\operatorname{diam}\Omega$ are only useful if Ω is bounded. For Ω finite, $d_{min} = \inf_{x,y\in\Omega} d(x,y)$. The function ϕ is described in the result relating the Prokhorov metric to Discrepancy. Table 6.1 is a key to the abbreviations for metrics used in the diagram.

The relationships illustrated in Figure 6.1 and any restrictions on them

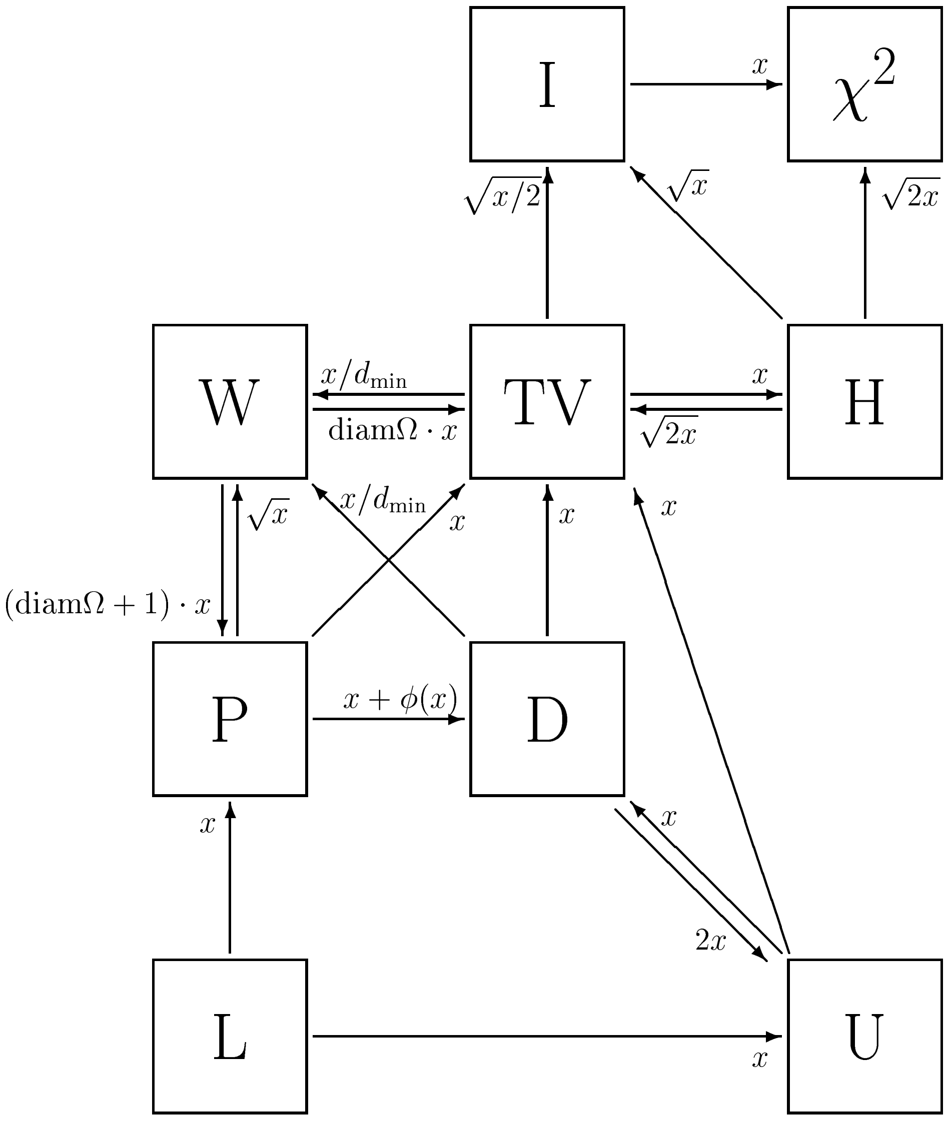


Figure 6.1: Relationships among probability metrics.

Abbreviation	Metric
D	Discrepancy
H	Hellinger metric
I	Relative entropy
L	Lévy metric
P	Prokhorov metric
TV	Total variation distance
U	Uniform (or Kolmogorov) metric
W	Wasserstein (or Kantorovich) metric
χ^2	χ^2 distance

Table 6.1: Abbreviations for metrics used in Figure 6.1.

are summarised below. References are given where proofs of these results are known to appear. Proofs are given for new results.

Uniform bounds Lévy

$$d_L(F,G) \le d_U(F,G).$$

Proof

$$\begin{aligned} d_L(F,G) &= \inf\{\epsilon > 0 : G(x-\epsilon) - \epsilon \le F(x) \le G(x+\epsilon) + \epsilon, x \in \mathbb{R}\} \\ &\le \inf\{\epsilon > 0 : G(x) - \epsilon \le F(x) \le G(x) + \epsilon, \forall x \in \mathbb{R}\} \\ &= d_U(F,G). \quad \blacksquare \end{aligned}$$

Bounds relating the Hellinger and Total variation distances

$$\frac{d_H^2}{2} \le d_{TV} \le d_H.$$

See LeCam (1969, p.36). It follows that when d_H is small, the variation distance is small.

Relative entropy bounds Total variation distance

For countable state spaces Ω,

$$\sqrt{2}\, d_{TV} \leq \sqrt{d_I}.$$

This inequality is due to Kullback (1967). It follows that when d_I is small, the total variation distance is small.

Discrepancy bounds the Prokhorov metric in special cases

The following theorem shows how discrepancy may be bounded by the Prokhorov metric by finding a suitable right-continuous function ϕ. For bounded Ω, $\phi(\epsilon)$ gives an upper bound on the additional ν-measure of the extended ball B^ϵ over the ball B, where $B^\epsilon = \{x : \inf_{y \in B} d(x, y) \leq \epsilon\}$.

Theorem 6.1 *Let Ω be any measurable metric space, and let ν be any probability measure satisfying*

$$\nu(B^\epsilon) \leq \nu(B) + \phi(\epsilon)$$

for all balls B and complements of balls B and some right-continuous function ϕ. Then for any other probability measure μ, if $d_P(\mu, \nu) = x$, then $d_D(\mu, \nu) \leq x + \phi(x)$.

For example, if ν is the uniform distribution on the circle or line, then $\phi(x) = 2x$.

Proof For μ, ν as above,

$$\mu(B) - \nu(B^x) \geq \mu(B) - \nu(B) - \phi(x).$$

And if $d_P(\mu, \nu) = x$, then $\mu(B) - \nu(B^{\tilde{x}}) \leq \tilde{x}$ for all $\tilde{x} > x$ and all Borel sets B. Combining with the above inequality, we see that

$$\mu(B) - \nu(B) - \phi(\tilde{x}) \leq \tilde{x}.$$

By taking the supremum over B which are balls or complements of balls, obtain

$$\sup_B(\mu(B) - \nu(B)) \leq \tilde{x} + \phi(\tilde{x}).$$

The same result may be obtained for $\nu(B) - \mu(B)$ by noting that $\nu(B) - \mu(B) = \mu(B^c) - \nu(B^c)$ which, after taking the supremum over B which are balls or complements of balls, obtain

$$\sup_B(\nu(B) - \mu(B)) = \sup_{B^c}(\mu(B^c) - \nu(B^c)) \leq \tilde{x} + \phi(\tilde{x})$$

as before. Since the supremum over balls and complements of balls will be larger than the supremum over balls, if $d_P(\mu, \nu) = x$, then $d_D(\mu, \nu) \leq \tilde{x} + \phi(\tilde{x})$ for all $\tilde{x} > x$. For right-continuous ϕ, the theorem follows by taking the limit as $\tilde{x}$ decreases to x. ∎

Using this result, one sees that for $\nu = U$, the uniform distribution on the circle or line,

$$d_D(\mu, U) \leq 3\, d_P(\mu, U).$$

Prokhorov and Wasserstein metrics

Huber (1981, p.33) shows that

$$d_P^2(\mu, \nu) \leq d_W(\mu, \nu) \leq 2\ d_P(\mu, \nu)$$

for any μ, ν probability measures on a complete separable metric space whose metric d is bounded by 1. In general, we show that

Theorem 6.2 *The Wasserstein and Prokhorov metrics satisfy*

$$d_P^2 \leq d_W \leq (\mathit{diam}(\Omega) + 1)\ d_P.$$

In particular, d_P and d_W define the same topology.

Proof For any joint distribution J on random variables X, Y,

$$\begin{aligned}
\mathbf{E}_J[d(X,Y)] &\leq \varepsilon \cdot \Pr(d(X,Y) \leq \varepsilon) + \mathrm{diam}(\Omega) \cdot \Pr(d(X,Y) > \varepsilon) \\
&= \varepsilon + (\mathrm{diam}(\Omega) - \varepsilon) \cdot \Pr(d(X,Y) > \varepsilon)
\end{aligned}$$

If $d_P(\mu,\nu) \leq \varepsilon$, we can choose a coupling so that $\Pr(d(X,Y) > \varepsilon)$ is bounded by ε (Huber 1981, p.27):

$$\leq \varepsilon + (\mathrm{diam}(\Omega) - \varepsilon)\varepsilon \leq (\mathrm{diam}(\Omega) + 1)\varepsilon.$$

Taking infimum of both sides over all couplings, we obtain

$$d_W \leq (\mathrm{diam}(\Omega) + 1) d_P.$$

To bound Prokhorov by Wasserstein, use Markov's inequality and choose ε such that $d_W(\mu,\nu) = \varepsilon^2$. Then

$$\Pr(d(X,Y) > \varepsilon) \leq \frac{1}{\varepsilon}\mathbf{E}_J[d(X,Y)] \leq \varepsilon$$

where J is any joint distribution on X, Y. By Strassen's theorem (see, for example, Huber (1981, Theorem 3.7, p.27)), $\Pr(d(X,Y) > \varepsilon) \leq \varepsilon$ is equivalent to $\mu(B) \leq \nu(B^\varepsilon) + \varepsilon$ for all Borel sets B, giving $d_P^2 \leq d_W$. ■

Prokhorov and Uniform are bounded above by Total variation and below by Lévy

For measures on $\mathbb{R}$ we have the following relations (see Huber (1981, p.34)):

$$d_L \leq d_P \leq d_{TV}$$

$$d_L \leq d_U \leq d_{TV}.$$

Relative entropy bounds χ^2

Theorem 6.3 *The relative entropy d_I and χ^2 distance d_{χ^2} satisfy*

$$d_I(\mu,\nu) \leq d_{\chi^2}(\mu,\nu).$$

Proof Since log is a concave function, Jensen's inequality yields

$$d_I(\mu,\nu) \leq \log\left(\int_\Omega (f/g)\ f\ d\lambda\right) \leq \log\left(1 + d_{\chi^2}(\mu,\nu)\right) \leq d_{\chi^2}(\mu,\nu),$$

where the second inequality is obtained by noting that

$$\int_\Omega \frac{(f-g)^2}{g}\ d\lambda = \int_\Omega \left(\frac{f^2}{g} - 2f + g\right)\ d\lambda = \int_\Omega \frac{f^2}{g}\ d\lambda - 1. \quad \blacksquare$$

The Wasserstein metric and total variation distance

Theorem 6.4 *The Wasserstein metric and the total variation distance satisfy the following relation:*

$$d_W \leq diam(\Omega) \cdot d_{TV}$$

where $diam(\Omega) = \sup\{d(x,y) : x, y \in \Omega\}$.

If Ω is a finite set, there is a bound the other way. If $d_{min} = \min_{ij} d(x_i, x_j)$ for points x_i in Ω, then

$$d_{min} \cdot d_{TV} \leq d_W.$$

Note that on an infinite set no such relation of the second type can occur because d_W may go to 0 while d_{TV} remains fixed at 1. ($\min_{a \neq b} d(a,b)$ could be 0 on an infinite set.)

Proof The first inequality follows from the coupling characterisations of Wasserstein and total variation by taking the infimum of the expected value of both sides over all possible joint distributions:

$$d(X,Y) \leq \mathbf{1}_{X \neq Y} \cdot \operatorname{diam}(\Omega).$$

The reverse inequality follows similarly from:

$$d(X,Y) \geq \mathbf{1}_{X \neq Y} \cdot \min_{a \neq b} d(a,b). \quad \blacksquare$$

Total variation bounds Discrepancy

It is clear that

$$d_D \leq d_{TV}$$

since total variation is the supremum over a larger class of sets than discrepancy.

No expression of the reverse type can hold since d_D may go to 0 while d_{TV} remains at 1. An elementary example is the convergence of a standardised Binomial (n,p) random variable with distribution μ_n which converges to the standard normal distribution, ν, as $n \to \infty$. For all $n < \infty$, $d_{TV}(\mu_n, \nu) = 1$, while $d_D(\mu_n, \nu) \to 0$ as $n \to \infty$. Another example is a random walk on the circle generated by irrational rotations (Su 1998).

Diaconis (1988, pp. 30-34) describes an interesting example that converges both in total variation distance and in discrepancy, but is known to converge at different rates. The example is a simple random walk with a randomness multiplier on the integers mod p where p is an odd number. The process is given by $X_0 = 0$ and $X_n = 2X_{n-1} + \epsilon_n \pmod p$ where the ϵ_i are independent and identically distributed taking values 0, ± 1 each with probability 1/3. The stationary distribution for this process is uniform. Using Fourier analysis, Chung, Diaconis and Graham (1987) show that $O(\log_2 p \log\log_2 p)$ steps are sufficient to achieve convergence in total variation distance, and are necessary when $p = 2^t - 1$, for t a positive integer. However, as proven in Su (1995, pp. 29-31), $O(\log p)$ steps are sufficient for convergence in discrepancy. In these results, the proportionality constants are known. Moreover, the convergence is qualitatively different in the two metrics; there is a cutoff in total variation distance where its value drops quickly from near 1 to near 0, but not in discrepancy.

Discrepancy equivalent to Uniform on $\mathbb{R}$

Theorem 6.5 *When the state space is $\mathbb{R}$, we have that*

$$d_U \leq d_D \leq 2d_U.$$

This shows that the topologies generated by d_D and d_U are equivalent on $\mathbb{R}$.

Proof A closed ball is an interval of the form $[a, b]$. By continuity of probabilities

$$|\mu((-\infty, x]) - \nu((-\infty, x])| = \lim_{a \searrow -\infty} |\mu([a, x]) - \nu([a, x])|.$$

Now

$$\begin{aligned} d_U(\mu, \nu) &= \sup_x |\mu((-\infty, x]) - \nu((-\infty, x])| \\ &= \lim_{a \searrow -\infty} \sup_x |\mu([a, x]) - \nu([a, x])| \\ &\leq d_D(\mu, \nu) \end{aligned}$$

since we are restricting the class of balls. For the other inequality, consider any closed ball B on $\mathbb{R}$. B is the set difference of $C = (-\infty, b]$ and $D = (-\infty, a)$. Then

$$\begin{aligned} |\mu(B) - \nu(B)| &= |\mu(C) - \mu(D) - \nu(C) + \nu(D)| \\ &\leq |\mu(C) - \nu(C)| + |\nu(D) - \mu(D)| \\ &= |\mu(C) - \nu(C)| + |\mu(D^c) - \nu(D^c)| \leq 2d_U(\mu, \nu). \end{aligned}$$

Taking the supremum of both sides over all balls B, we see that $d_D \leq 2d_U$.

■

Hellinger and χ^2

$$d_H^2 \leq 2d_{\chi^2}.$$

See Reiss (1989, p.99). It follows that when d_{χ^2} is small, d_H is small.

Hellinger and Relative entropy

$$d_H^2 \leq d_I.$$

See Reiss (1989, p.99). It follows that when d_I is small, d_H is small.

Wasserstein metric and Discrepancy

If Ω is a finite set,

$$d_{min} \cdot d_D \leq d_W$$

where $d_{min} = \min_{i,j} d(x_i, x_j)$ for points x_i, x_j in Ω.

Proof In the equivalent form of the Wasserstein metric, Equation (6.1), take

$$h(x) = \begin{cases} d_{min} & \text{for } x \text{ in } B \\ 0 & \text{otherwise} \end{cases}$$

for B any closed ball. $h(x)$ satisfies the Lipschitz condition. Then

$$\begin{aligned} d_{min} \, |\mu(B) - \nu(B)| &= \left| \int_\Omega h \, d\mu - \int_\Omega h \, d\nu \right| \\ &\leq d_W \end{aligned}$$

and taking B to be the ball that maximises $|\mu(B) - \nu(B)|$ gives the result.

■

On continuous spaces, it is possible for d_W to go to 0 while d_D remains at 1. For example, take delta measures δ_ϵ converging on δ_0.

6.4 Some Applications of Metrics

Of all the probability metrics in wide use, the total variation distance appears to be the most common. Applications include bounding rates of convergence of random walks (for example Diaconis (1988), Su (1995), Rosenthal (1995a), Diaconis and Stroock (1991)), and Markov chain Monte Carlo algorithms (Tierney 1994, Gilks et al. 1996). Much of the success in achieving rates of convergence in total variation distance has resulted from its useful coupling characterisation.

However, other metrics can be useful because of their special properties. For instance, the Hellinger metric is useful when working with convergence of product measures because it factors nicely in terms of the convergence of the components. Reiss (1989) uses this fact and the relation between the Hellinger metric and total variation distance to obtain total variation bounds. The Hellinger metric is also used in the theory of asymptotic efficiency (see, for example, LeCam (1986)) and minimum Hellinger distance estimation (see, for example, Lindsay (1994)).

In Section 4.5 we obtained a bound on the rate of convergence of a Markov chain Monte Carlo algorithm in total variation distance via its relationship with the Wasserstein metric. Use of the coupling characterisation for the Wasserstein metric yielded a bound on the rate of convergence that is *better* than what was obtained directly from bounding total variation in Section 3.5.2. The fact that the Wasserstein metric is a minimal distance of two random variables with fixed distributions has led to its use in the study of distributions with fixed marginals (see, for example, Rüschendorf, Schweizer and Taylor (1996)).

For continuous state spaces, total variation distance is not always suitable. Su (1998) examines a random walk on the circle generated by a single irrational rotation, which proceeds as follows: fix an irrational α and at each step rotate the current position by $\pm\alpha$ with probability $1/2$. This walk does not converge in total variation because the k-th step probability distribution is finitely supported. However this walk does converge in the weak* topology. The Prokhorov metric, which metrizes weak* convergence, is not easy to bound. The discrepancy metric bounds weak* convergence when the limiting measure is uniform, so Su (1998) obtains a rate of convergence in discrepancy. For an elementary example which converges in the weak* topology, but for which total variation distance is not suitable, consider a standardised binomial random variable whose distribution converges to the standard normal. Because the binomial distribution is discrete, the total variation distance stays at 1, while probability metrics that metrize weak* convergence go to 0.

In Section 4.4, our Markov chain Monte Carlo algorithm does converge in total variation distance, but coupling bounds are difficult to apply since the state space is continuous and one must wait for random variables to couple exactly. On the other hand, the Wasserstein metric has a coupling bound which depends on the *distance* between two random variables; in this example it is enough to wait for the random variables only to couple to within ϵ.

Chapter 7

Conclusions

In this thesis we have developed new precise upper bounds on the convergence time of Gibbs sampler algorithms used in Bayesian image restoration. We have considered measuring convergence in total variation distance, which is the usual choice, and have achieved additional success by considering convergence in the Wasserstein metric. The computation of parameters required by our methods may be intractable for more complex models and Markov chain Monte Carlo dynamics, but we discuss how auxiliary simulation can be used to provide useful approximate values. Also, our results can be applied to the exact sampling algorithm of Propp and Wilson (1996) to achieve bounds on the running time of coupling-from-the-past.

The following list contains some ideas for future work and extensions of the ideas and results in this thesis.

- Rather than just approaching the convergence issue as finding the number of iterations to ensure that the total variation distance or Wasserstein metric is below a specified tolerance, exploring the complete distribution of the coupling time can provide guidance in the design of an optimal strategy for the coupling-from-the-past algorithm. For example, if there is only a small chance of coupling quickly, it would be worthwhile to start the algorithm at a time far in the past. Understanding the coupling distribution may also generate ideas for how the joint updating can be modified to encourage fast coalescence.

- For chains for which there exists no maximal state, the idea of a dominating chain of Møller (1999) may be useful. In our context, this may be very worthwhile for single photon emission computed tomography (SPECT) in which the pixel values are counts of gamma-rays, which are modelled with a Poisson distribution.

- In Section 6.3 we referred to an example of a random walk which converges to its stationary distribution in both total variation distance and discrepancy, but is known to converge at different rates. Other examples of this type, particularly examples that converge at different rates in total variation distance and the Wasserstein metric, will help clarify the choice of metric in assessing convergence.

- Diaconis and Saloff-Coste (1993) develop inequalities that give bounds on the eigenvalues of a reversible Markov chain in terms of the eigenvalues of a second chain. These results can be applied to the comparison of the convergence times of two Markov chains. Future work could be carried out to determine if these or similar ideas can be applied to the results of this thesis, in order to achieve bounds on the convergence time of similar algorithms.

- Generalising the binary image model of Chapter 3 to a finite number of ordered colours with a Potts model prior (as used in Besag (1986)) is straightforward. Extension to models that do not maintain the partial order in the state space for coupled Markov chains is not readily available. This excludes us from considering, for example, models for multiple unordered colours.

Bibliography

Aldous, D. (1983). Random walks on finite groups and rapidly mixing Markov chains, *in* J. Azema and M. Yor (eds), *Séminaire de Probabilités XVII 1981/82*, Vol. 986 of *Lecture notes in mathematics*, Springer-Verlag, Berlin; New York, pp. 243–297.

Aldous, D. and Diaconis, P. (1987). Strong uniform times and finite random walks, *Advances in Applied Mathematics* **8**: 69–97.

Besag, J. (1986). On the statistical analysis of dirty pictures, with discussion, *Journal of the Royal Statistical Society* B **48**: 259–302.

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation, *Journal of the Royal Statistical Society* B **55**: 25–37.

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems, *Statistical Science* **10**: 3–66.

Billingsley, P. (1986). *Probability and Measure*, second edn, John Wiley and Sons.

Brooks, S. P. and Roberts, G. O. (1997). Assessing convergence of Markov chain Monte Carlo algorithms, *Statistics and Computing* **8**: 319–335.

Chung, F., Diaconis, P. and Graham, R. L. (1987). Random walks arising in random number generation, *The Annals of Probability* **15**: 1148–1165.

Cipra, B. A. (1987). An introduction to the Ising model, *American Mathematical Monthly* **94**: 937–959.

Corcoran, J. N. and Tweedie, R. L. (1998). Perfect sampling of Harris recurrent Markov chains, Preprint.

Cowles, M. K. (1999). MCMC sampler convergence rates for hierarchical normal linear models: A simulation approach. Preprint.

Cowles, M. K. and Carlin, B. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association* **91**: 883–904.

Cowles, M. K. and Rosenthal, J. S. (1998). A simulation approach to convergence rates of Markov chain Monte Carlo algorithms, *Statistics and Computing* **8**: 115–124.

Cowles, M. K., Roberts, G. O. and Rosenthal, J. S. (1997). Possible biases induced by MCMC convergence diagnostics, *Journal of Statistical Computing and Simulation*. To appear.

Diaconis, P. (1988). *Group Representations in Probability and Statistics*, Vol. 11 of *Lecture Notes – Monograph Series*, Institute of Mathematical Statistics.

Diaconis, P. and Saloff-Coste, L. (1993). Comparison theorems for reversible Markov chains, *The Annals of Applied Probability* **3**: 696–730.

Diaconis, P. and Stroock, D. (1991). Geometric bounds for eigenvalues of Markov chains, *The Annals of Applied Probability* **1**: 36–61.

Diaconis, P. and Zabell, S. L. (1982). Updating subjective probability, *Journal of the American Statistical Association* **77**: 822–830.

Dudley, R. M. (1989). *Real Analysis and Probability*, Wadsworth & Brooks/Cole, Belmont, CA.

Durrett, R. (1996). *Probability: Theory and Examples*, second edn, Duxbury Press, Belmont, California.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I, third edn, John Wiley and Sons.

Fill, J. A. (1998). An interruptible algorithm for perfect sampling via Markov chains, *The Annals of Applied Probability* **8**: 131–162.

Fishman, G. S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*, Springer-Verlag, New York.

Foss, S. and Tweedie, R. (1998). Perfect simulation and backward coupling, *Stochastic Models* **14**: 187–203.

Frieze, A., Kannan, R. and Polson, N. (1994). Sampling from log-concave distributions, *The Annals of Applied Probability* **4**: 812–834.

Frigessi, A., di Stefano, P., Hwang, C.-R. and Sheu, S.-J. (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics, *Journal of the Royal Statistical Society* B **55**: 205–219.

Frigessi, A., Martinelli, F. and Stander, J. (1997). Computational complexity of Markov chain Monte Carlo methods for finite Markov random fields, *Biometrika* **84**: 1–18.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association* **85**: 398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721–741.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Green, P. J. (1996). MCMC in image analysis, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, pp. 381–400.

Green, P. J. and Han, X.-L. (1992). Metropolis methods, Gaussian proposals, and antithetic variables, *in* P. Barone, A. Frigessi and M. Piccioni (eds), *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, Springer-Verlag, Berlin Heidelberg.

Green, P. J. and Murdoch, D. J. (1998). Exact sampling for Bayesian inference: towards general purpose algorithms, *in* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, Oxford University Press.

Guglielmi, A., Holmes, C. C. and Walker, S. G. (1999). Perfect simulation involving a continuous and unbounded state space, Preprint.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**: 97–109.

Huber, P. J. (1981). *Robust Statistics*, John Wiley & Sons, New York.

Ingrassia, S. (1994). On the rate of convergence of the Metropolis algorithm and Gibbs sampler by geometric bounds, *The Annals of Applied Probability* **4**: 347–389.

Jerrum, M. and Sinclair, A. (1993). Polynomial-time approximation algorithms for the Ising model, *SIAM Journal on Computing* **22**: 1087–1116.

Kullback, S. (1967). A lower bound for discrimination in terms of variation, *IEEE Transactions on Information Theory* **4**: 126–127.

LeCam, L. M. (1969). *Théorie Asymptotique de la Décision Statistique*, Les Presses de l'Université de Montréal, Montréal.

LeCam, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Liggett, T. M. (1985). *Interacting Particle Systems*, Springer-Verlag, New York.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods, *The Annals of Statistics* **22**: 1081–1114.

Lindvall, T. (1992). *Lectures on the Coupling Method*, John Wiley & Sons, New York.

Luby, M., Randall, D. and Sinclair, A. (1995). Markov chain algorithms for planar lattice structures (extended abstract), 36^{th} *Annual Symposium on Foundations of Computer Science*, pp. 150–159.

Madras, N. and Piccioni, M. (1999). Importance sampling for families of distributions, *The Annals of Applied Probability* **9**: 1202–1225.

Martinelli, F. (1997). Lectures on Glauber dynamics for discrete spin models, Lecture Notes, School in Probability Theory, Saint Flour.

Mengerson, K. and Tweedie, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms, *The Annals of Statistics* **24**: 101–121.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines, *The Journal of Chemical Physics* **21**: 1087–1092.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*, Springer-Verlag, London.

Møller, J. (1999). Perfect simulation of conditionally specified models, *Journal of the Royal Statistical Society* B **61**: 251–64.

Møller, J. and Nicholls, G. K. (1999). Perfect simulation for sample-based inference, Preprint.

Murdoch, D. J. (1999). Exact sampling for Bayesian inference: Unbounded state spaces, To appear in Proceeding of the Workshop on Monte Carlo Methods at the Fields Institute, October, 1998.

Murdoch, D. J. and Green, P. J. (1998). Exact sampling from a continuous state space, *Scandinavian Journal of Statistics* **25**: 483–502.

Neal, R. M. (1999). Circularly-coupled Markov chain sampling, *Technical Report 9910*, Department of Statistics, University of Toronto.

Polson, N. G. (1996). Convergence of Markov chain Conte Carlo algorithms, *in* J. M. Bernardo, A. P. Dawid and A. F. M. Smith (eds), *Bayesian Statistics 5*, Oxford University Press.

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics, *Random Structures and Algorithms* **9**: 223–252.

Propp, J. G. and Wilson, D. B. (1998). How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph, *Journal of Algorithms* **27**: 170–217.

Rachev, S. T. (1984). The Monge-Kantorovich mass transference problem and its stochastic applications, *Theory of Probability and its Applications* **29**: 647–676.

Rachev, S. T. (1991). *Probability Metrics and the Stability of Stochastic Models*, John Wiley & Sons, Chichester, New York.

Reiss, R.-D. (1989). *Approximate Distributions of Order Statistics*, Springer-Verlag, New York.

Roberts, G. O. and Rosenthal, J. S. (1998). Markov chain Monte Carlo: Some practical implications of theoretical results, with discussion, *Canadian Journal of Statistics* **26**: 5–31.

Roberts, G. O. and Rosenthal, J. S. (1999). Convergence of slice sampler Markov chains, *Journal of the Royal Statistical Society* B **61**: 643–660.

Rosenthal, J. S. (1995a). Convergence rates of Markov chains, *SIAM Review* **37**: 387–405.

Rosenthal, J. S. (1995b). Minorization condition and convergence rates for Markov chain Monte Carlo, *Journal of the American Statistical Association* **90**: 558–566.

Rosenthal, J. S. (1999). A review of asymptotic convergence for general state space Markov chains, Preprint.

Rüschendorf, L., Schweizer, B. and Taylor, M. D. (eds) (1996). *Distributions with Fixed Marginals and Related Topics*, Vol. 28 of *Lecture Notes – Monograph Series*, Institute of Mathematical Statistics, Hayward, California.

Sinclair, A. (1992). Improved bounds for mixing rates of Markov chains and multicommodity flow, *Combinatorics, Probability and Computing* **1**: 351–370.

Sinclair, A. and Jerrum, M. (1989). Approximate counting, uniform generation and rapidly mixing Markov chains, *Information and Computation* **1**: 93–133.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society* B **55**: 3–23.

Su, F. E. (1995). *Methods for Quantifying Rates of Convergence for Random Walks on Groups*, PhD thesis, Harvard University.

Su, F. E. (1998). Convergence of random walks on the circle generated by an irrational rotation, *Transactions of the American Mathematical Society* **350**: 3717–3741.

Szulga, A. (1982). On minimal metrics in the space of random variables, *Theory of Probability and its Applications* **27**: 424–430.

Thisted, R. A. (1988). *Elements of Statistical Computing*, Chapman and Hall, New York.

Tierney, L. (1994). Markov chains for exploring posterior distributions, with discussion, *The Annals of Statistics* **22**: 1701–1762.

Tierney, L. (1996). Introduction to general state-space Markov chain theory, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, pp. 59–74.

Zolotarev, V. M. (1983). Probability metrics, *Theory of Probability and its Applications* **28**: 278–302.