CONVERGENCE IN THE WASSERSTEIN METRIC FOR MARKOV CHAIN MONTE CARLO ALGORITHMS WITH APPLICATIONS TO IMAGE RESTORATION

ALISON L. GIBBS

Abstract. In this paper, we show how the time for convergence to stationarity of a Markov chain can be assessed using the Wasserstein metric, rather than the usual choice of total variation distance. The Wasserstein metric may be more easily applied in some applications, particularly those on continuous state spaces. Bounds on convergence time are established by considering the number of iterations required to approximately couple two realizations of the Markov chain to within ϵ tolerance. The particular application considered is the use of the Gibbs sampler in the Bayesian restoration of a degraded image, with pixels that are a continuous grey-scale and with pixels that can only take two colours. On finite state spaces, a bound in the Wasserstein metric can be used to find a bound in total variation distance. We use this relationship to get an $O(N \log N)$ bound on the convergence time of the stochastic Ising model that holds for appropriate values of its parameter as well as other binary image models. Our method employing convergence in the Wasserstein metric can also be applied to perfect sampling algorithms involving coupling from the past to obtain estimates of their running times.

1. Introduction

A critical issue for users of Markov chain Monte Carlo (MCMC) algorithms is the number of iterations that should be run until the result can be considered a sample from the stationary distribution. Convergence diagnostics do not guarantee convergence, and are known to introduce bias into the results (Cowles, Roberts and Rosenthal 1997). The theoretical results that exist are difficult to apply in practice. Perfect simulation algorithms (Propp and Wilson 1996, Fill 1998) have made a important contribution to meeting this need, although difficult

1991 *Mathematics Subject Classification.* 65C05, 60J10, 68U10, 62-04.

Key words and phrases. Bayesian image restoration; Coupling; Gibbs sampler; Ising model; Perfect simulation; Probability metrics; Total variation distance.

rigorous analysis is still necessary if one wishes to determine their expected running times. See Häggström and Nelander (1999) and Huber (1998) for some work in this direction.

In this paper 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 a Markov chain's distance from its stationary distribution, the Wasserstein metric has a coupling characterization. However, the Wasserstein metric may be more useful on continuous state spaces than the total variation distance. The coupling characterization for total variation distance is the probability that two random variables with the relevant distributions become equal, while the Wasserstein coupling characterization 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, such as the monotone chain on a continuous state space examined in Section 3.

Our results hold for bounded state spaces. The number of iterations for approximate convergence of the Markov chain to its stationary distribution is bounded by the time necessary for the expected distance between coupled chains to be ϵ, where ϵ is user-specified. Convergence to within ϵ precision has been considered by Møller (1999) in the context of applying perfect simulation algorithms which require coalescence of coupled Markov chains to continuous state spaces. A discussion of the application of the results of this paper to perfect simulation is given in Section 5.

We examine in detail a particular application: the Bayesian restoration of a noisy image. See Green (1996) for a discussion of the Markov chain Monte Carlo approach to this problem. We consider an image composed of pixels taking on values either in a $[0, 1]$ grey-scale or a binary image. 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 neighbors.

The theory that leads to the convergence bound and more discussion of the choice of probability metric are provided in Section 2. In Section 3, 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. The method presented in Section 2 also allows an improvement to the bound on the convergence time found in Gibbs (2000) for a similar image restoration problem where the pixels are binary. This improvement is given in Section 4.

The results of Section 4 include a precise $O(N \log N)$ bound on the convergence time of the stochastic Ising model, where N is the number of pixels. Häggström and Nelander (1999) also give an $O(N \log N)$ bound on the expected running time of a corresponding perfect sampling algorithm. Their results apply to the particular case of Gibbs samplers for Markov random fields. Our method of Section 2 provides an alternative approach which can be more broadly applied to any Markov chain Monte Carlo algorithm on a bounded state space.

2. Convergence in the Wasserstein Metric

There exist dozens of distance measures to quantify closeness between two probability measures (see, for example, Rachev (1991)). 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, 1996)). The total variation distance between two probability measures μ and ν on a space $\mathcal{X}$ is

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

Some authors (for example, Tierney (1996)) define total variation distance as twice our definition.

Much of the success in bounding convergence in total variation distance arises from its coupling characterization

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

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 (1995), Luby, Randall and Sinclair (1995), and Gibbs (2000).

In this paper, we introduce the use of the Wasserstein metric to assessing convergence of Markov chain Monte Carlo algorithms. If μ, ν are two probability measures on the same space $\mathcal{X}$, the Wasserstein metric is

$$W(\mu, \nu) = \inf \mathbf{E}[d(X, Y)] \tag{2.1}$$

where the infimum is taken over all random variables X, Y with $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$ and d is any given metric on $\mathcal{X}$. The Wasserstein metric was chosen because of its coupling characterization. On continuous state spaces, the Wasserstein metric may be more practical than total variation distance because it does not require exact coalescence and because the Wasserstein metric metrizes convergence in distribution on bounded state spaces (see, for example, Dudley (1989)) while convergence in total variation distance is stronger than convergence in distribution.

When estimating one measure by another, the error in expectations with respect to these measures of a bounded function is small when the total variation between the two measures is small. This is because there exists the following equivalent formulation of total variation distance

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

where the supremum is taken over functions $h : \mathcal{X} \to \mathbf{R}$ satisfying $|h(x)| \le 1$. A similar result exists for the Wasserstein metric. By the Kantorovich-Rubinstein Theorem (see, for example, Dudley (1989) Theorem 11.8.2), for $\mathcal{X}$ a separable metric space,

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

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

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

The following relationship exists between total variation distance and the Wasserstein metric:

$$W \le \operatorname{diam}(\mathcal{X})\, TV$$

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

$$d_{\min}\, TV \;\le\; W. \tag{2.2}$$

On an infinite set no such general relation can occur as it is possible for W to go to 0 while TV remains fixed at 1. The relationship (2.2) will be proven in Section 4 and applied to give improved precise bounds for results in Gibbs (2000).

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.

Consider a Markov chain on a bounded state space $\mathcal{X}$ with $\operatorname{diam}(\mathcal{X}) = \sup_{x,y\in\mathcal{X}} d(x,y)$ where d 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. We consider two coupled realizations of the chain; each of these chains can be started in any initial state and their updating may or may not be independent of each other, but their marginal distributions must each follow the transition probabilities of the original Markov chain. We describe in detail the coupling used in our image restoration problem in Section 3. If, on average, these two coupled realizations of the Markov chain are getting closer together at each iteration, we can bound the time until the Wasserstein metric, $W(P^T(x^0,\cdot),\pi(\cdot))$, is small as follows.

Theorem 2.1. *Consider two coupled realizations 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*

$$\mathbf{E}[d(X_{t+1},Y_{t+1})|X_t,Y_t] \le c\, d(X_t,Y_t) \tag{2.3}$$

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

$$T > \frac{\ln(\epsilon/\mathrm{diam}(\mathcal{X}))}{\ln c}$$

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

The proof uses the following lemma.

Lemma 2.1. *Suppose $\{X_t\}$, $\{Y_t\}$ are two coupled Markov chains for which there exists a positive constant c such that*

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

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

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

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

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

Then

$$\begin{aligned}\mathbf{E}[d(X_{k+1},Y_{k+1})|X_0,Y_0] &= \mathbf{E}\left[\,\mathbf{E}[d(X_{k+1},Y_{k+1})|X_k,Y_k]\,|X_0,Y_0\right]\\ &\le \mathbf{E}[c\, d(X_k,Y_k)|X_0,Y_0]\\ &\le c^{k+1}\, d(X_0,Y_0). \qquad (2.4)\end{aligned}$$

□

Proof of theorem: We wish to bound the number of iterations, T, which guarantee $W(P^T(x^0,\cdot),\pi(\cdot)) \le \epsilon$ where x^0 is any initial state. Consider a coupled realization of the Markov chain started in its stationary distribution π. Let x denote the initial state af this second chain. Applying the lemma,

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

Thus $W(P^T(x^0,\pi(\cdot))$ is less than or equal to ϵ for

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

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

Thus if we can find a value of c satisfying (2.3), we can find a bound on the convergence time guaranteeing the Wasserstein metric is less than a specified tolerance ϵ.

3. Restoring a Grey-Scale Image

3.1. **The Model and Algorithm.** We now apply the convergence bound in the Wasserstein metric of Theorem 2.1 to an algorithm used to restore a distorted continuous grey-scale 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, and values in between represent the various shades of grey. Our belief about the image represented in this manner is that pixels tend to be like their nearest neighbors. This belief is modeled by the pairwise-difference prior distribution on the value of the image $x=\{x_i\}_{i=1}^N$ which has density

$$\pi_\gamma(x) \propto \exp\left\{-\sum_{<i,j>}\frac{1}{2}\left[\gamma(x_i-x_j)\right]^2\right\} \tag{3.1}$$

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

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

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

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

$$\pi_{posterior}(x|y) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-y_i)^2 - \sum_{<i,j>}\frac{1}{2}\left[\gamma(x_i-x_j)\right]^2\right\} \tag{3.2}$$

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 neighbors that are influential for pixel i, and $n_{\max} = \max_i n_i$. 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 (3.2) the full conditional densities for x_i, the value of pixel i, given the values of all other pixels, x_{-i}, and the observed image y are

$$\pi_{FC}(x_i|x_{-i},y) \propto \exp\left\{-\frac{(\sigma^{-2}+n_i\gamma^2)}{2}\left[x_i-\left(\sigma^{-2}+n_i\gamma^2\right)^{-1}\left(\sigma^{-2}y_i+\gamma^2\sum_{j\sim i}x_j\right)\right]^2\right\} \tag{3.3}$$

on $[0,1]$ and 0 elsewhere. The pixels $\{x_j,\ j\sim i\}$ are those which are neighbors of the ith 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 (3.3). To sample from the normal distribution restricted to $[0,1]$ we use an inverse normal cumulative distribution function approximation, with the adjustment for restricted sampling described in Fishman (1996, p. 152). 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 (3.2). 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 3.2, we require that our Markov chain is monotone, i.e. a partial order exists on our state space, and the Markov chain transitions preserve this order. 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 used in the generation of the corresponding samples from the distribution of interest.

3.2. **The Convergence Result.** The following theorem gives a precise $O(N \log N)$ result for the number of iterations required for convergence to ϵ precision in the Wasserstein metric the algorithm described Section 3.1.

Theorem 3.1. *Consider sampling using the random scan Gibbs sampler from (3.2), the posterior distribution for an image of N pixels, each taking a value in $[0,1]$, where the observed data are the random distortion of the image modelled by the addition of $N(0,\sigma^2)$ noise independently at each pixel. The distance in the Wasserstein metric between the distribution of the Markov chain and its stationary distribution will be less than ϵ at iteration T for*

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

where $n_{\max}$ is the maximum over all pixels of the number of influential neighboring pixels, and γ is the value of the smoothing parameter from the prior distribution (3.1).

Before we prove this theorem, we will show that coupled realizations of the Markov chain are monotone 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 order exists on the state space where one configuration is greater than or equal to another when each corresponding pixel has this ordering, i.e.

$$x \geq y \Longleftrightarrow x_i \geq y_i, i = 1, \dots, N.$$

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, with $d(x^{\min}, x^{\max}) = \mathrm{diam}(\mathcal{X})$. The following lemma indicates that the update function we use for our Markov chain is monotone.

Lemma 3.2. *Consider two coupled realizations of the Markov chain described in Section 3.1. 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 Lemma 3.2 we require the following lemma which we quote in the form given in Roberts and Rosenthal (1999, Lemma 5).

Lemma 3.3. *Suppose that* μ_1 *and* μ_2 *are two probability measures on* $\mathbf{R}$, *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* $\mathbf{R}$ *into* $\mathbf{R}^+$. *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 3.2: Consider two coupled realizations of the Markov chain started in initial states x^{high} and x^{low}. At each iteration one pixel, say the ith, is randomly chosen for updating as a sample from the full conditionals (3.3). For each coupled realization 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 coupled chains 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 $\alpha_i^{t,\,low}$. 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 monotone, i.e. we will show

$$\Pr(x_i^{t,\,high} > a \,|\, x^{t-1,\,high}) \geq \Pr(x_i^{t,\,low} > a \,|\, x^{t-1,\,low}) \tag{3.4}$$

for any $a \in \mathbf{R}$ whenever $x_i^{t-1,\,high} \geq x_i^{t-1,\,low}$, 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 can show that

$$\mathbf{E}[\mathbf{1}_{\{z_i^{t,\,high}a\}}|z_i^{t,\,high} \in [0,1]] \geq \mathbf{E}[\mathbf{1}_{\{z_i^{t,\,low}>a\}}|z_i^{t,\,low} \in [0,1]] \tag{3.5}$$

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 $x_i^{t-1,\,high} \geq x_i^{t-1,\,low}$, 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 3.3 gives (3.5) giving (3.4).

Using the inverse normal distribution function for updating as described in Section 3.1, 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 ith pixels are $x^{*,\,high}$, $x^{*,\,low}$, chosen to satisfy

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

and

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

Then by (3.4) $x^{t-1,\,high} \geq x^{t-1,\,low}$ results in $x^{t,\,high} \geq x^{t,\,low}$. □

The following lemma shows that the difference in the means of normal distributions with the same variance is at least as great as the difference in the means of the corresponding normal distributions restricted to $[0,1]$. It will be used in the proof of Theorem 3.1.

Lemma 3.4. *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 3.4: We will first define the following quantities:

$$\begin{aligned} f_\beta(\alpha) &= e_\beta(\alpha) - \alpha \\ p_\beta(s,t) &= \Pr(Z \in [s,t]) \text{ where } Z \sim N(0,\beta^2) \end{aligned}$$

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

Now

$$\begin{aligned} f_\beta(\alpha) &= \int_0^1 x\, e^{-(x-\alpha)^2/(2\beta^2)} dx \Big/ \int_0^1 e^{-(x-\alpha)^2/(2\beta^2)} dx \ - \ \alpha \\ &= \int_{-\alpha}^{1-\alpha} u\, e^{-u^2/(2\beta^2)} du \Big/ \int_{-\alpha}^{1-\alpha} e^{-u^2/(2\beta^2)} du \\ &= m_\beta(-\alpha, 1-\alpha). \end{aligned} \tag{3.6}$$

And if $s < t < u$,

$$
\begin{aligned}
& m_\beta(s,u) \\
&= \int_s^u x\, e^{-x^2/(2\beta^2)} dx \Big/ \int_s^u e^{-x^2/(2\beta^2)} dx \\
&= \left\{ \int_s^t x\, e^{-x^2/(2\beta^2)} dx + \int_t^u x\, e^{-x^2/(2\beta^2)} dx \right\} \\
&\quad \times \left\{ \int_s^u e^{-x^2/(2\beta^2)} dx \right\}^{-1} \\
&= \left\{ \frac{\int_s^t \frac{1}{\sqrt{2\pi}\beta} e^{-x^2/(2\beta^2)} dx \int_s^t x e^{-x^2/(2\beta^2)} dx}{\int_s^t e^{-x^2/(2\beta^2)} dx} \right. \\
&\qquad \left. + \frac{\int_t^u \frac{1}{\sqrt{2\pi}\beta} e^{-x^2/(2\beta^2)} dx \int_t^u x e^{-x^2/(2\beta^2)} dx}{\int_t^u e^{-x^2/(2\beta^2)} dx} \right\} \\
&\qquad \times \left\{ \int_s^u \frac{1}{\sqrt{2\pi}\beta} e^{-x^2/(2\beta^2)} dx \right\}^{-1} \\
&= \frac{p_\beta(s,t) m_\beta(s,t) + p_\beta(t,u) m_\beta(t,u)}{p_\beta(s,t) + p_\beta(t,u)}. \qquad (3.7)
\end{aligned}
$$

Now for $s < t < u$,

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

and from (3.7)

$$m_\beta(s,t) \le m_\beta(s,u) \le m_\beta(t,u). \qquad (3.8)$$

Now let $s < t < s+1$. Then $-t < -s < -t+1 < -s+1$ and using (3.6) and (3.8) 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. $\square$

Proof of Theorem 3.1: 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

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

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 (2.1).

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

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

where the pixelwise distances are weighted by the number of neighbours of the pixel. By Lemma 3.2

$$d(x^{t,\max},\, x^{t,\min})=\sum_{i=1}^{N} n_i(x_i^{t,\max}-x_i^{t,\min}).$$

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

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

Now

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

where the first inequality uses Lemma 3.4. The final sum over $i \sim j$ is less than n_j since the summand is less than 1 giving

$$\begin{aligned}
&\mathbf{E}\left[d(x^{t+1,\max}, x^{t+1,\min})|x^{t,\max}, x^{t,\min}\right] \\
&\leq \sum_{j=1}^{N}\left(\frac{N-1}{N}+\frac{1}{N}\frac{n_{\max}\gamma^2}{\sigma^{-2}+n_{\max}\gamma^2}\right)n_j(x_j^{t,\max}-x_j^{t,\min}) \\
&= \frac{N-1+n_{\max}\gamma^2/(\sigma^{-2}+n_{\max}\gamma^2)}{N}d(x^{t,\max}, x^{t,\min}).
\end{aligned}$$

Noting that $\operatorname{diam}(\mathcal{X}) \leq n_{\max}N$ gives our result. □

The fourth column of Table 3.1 shows the theoretical results for various values of N, ϵ, $n_{\max}$, γ and σ. Values are compared to a 32×32 grid with $\epsilon = 0.1$, $n_{\max} = 4$, $\gamma = 1$, and $\sigma = 0.2$. The required number of iterations until convergence to within ϵ tolerance, 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.

3.3. **Results from Simulations.** The final column of Table 3.1 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 restricted normal distribution were obtained using the inverse normal cumulative distribution function 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 similar to that illustrated in Figure 3.1. 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 is reassuring. Note that one simulation 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 realization 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 realizations of our Markov chain, started in the maximal and minimal states. We set the prior smoothing parameter γ at 1.0 and the neighborhood structure such that interior pixels were influenced by their four nearest neighbors. These simulations were each run for 12610

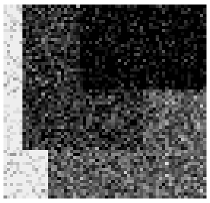
N	ϵ	$n_{\max}$	γ	σ	Theoretical value of T	Number of iterations required in simulation
32×32	0.1	4	1	0.2	12610	11293
10×10					958	572
256×128					535417	445301
32×32	0.1	4	1	0.2	12610	9932
	0.01				15344	11442
	1				9877	8558
32×32	0.1	4	1	0.2	12610	10618
		2			10974	9709
		8			15287	12603
32×32	0.1	4	1	0.2	12610	12760
			0.1		10888	6593
			3		26531	22186
			10		184874	163364
32×32	0.1	4	1	0.2	12610	10534
				0.1	11305	9053
				0.3	14786	11714

TABLE 3.1. Convergence times for the restoration of a grey-scale image.

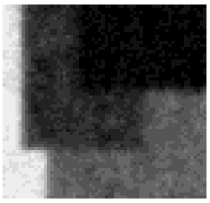
iterations, the number our theory requires for convergence to within $\epsilon = 0.1$ precision in the Wasserstein metric. For each simulation, the actual distance between the states of two Markov chains after iteration 12610 was recorded. The distribution of these distances was right-skewed, with a mean of 0.049540, 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 some high values, including 5 values greater than 3.0. Approximately one-tenth of the simulations (93 of 1000) had not coupled to within 0.1 precision.

Figure 3.1 gives an example of the image restoration process. Figure 3.1(a) shows the true image, drawn on a 64×64 grid. It was randomly degraded with $N(0, 0.15^2)$ noise, added independently to each pixel. The degraded image is shown in Figure 3.1(b). In the prior distribution, γ was set at 4 and neighbors were considered to be the pixels above, below, and beside ($n_{\max} = 4$). The specified accuracy was $\epsilon = 0.1$. Note that this error is distributed across 4096 pixels. The algorithm was run for 491788 iterations, the number that our theory requires for convergence to within ϵ accuracy, and our initial state was every pixel black. Figure 3.1(c) is an estimate of the posterior mean; it

(a)



(b)



(c)

FIGURE 3.1. A simulated restoration of a 64×64 image. (a) true image; (b) observed image; (c) the mean of 10 independent samples from the posterior distribution.

is the pixelwise average of 10 independent samples from the MCMC approximation to the posterior distribution.

4. Results for the Restoration of a Binary Image

The case where the image is a grid of binary pixels was considered in Gibbs (2000). Consider a configuration $\{x_i\}_{i=1}^N$ of N pixels which take on values $+1$ or -1. The following prior distribution, the Ising model, assigns greater probability to configurations where neighboring pixels are alike

$$\pi_\beta(x) = \frac{1}{Z} \exp\left\{ -\beta \sum_{<i,j>} x_i x_j \right\} \tag{4.1}$$

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

In Gibbs (2000), precise $O(N^2)$ bounds were found on the convergence time of the random scan 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 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 Gibbs (2000) 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 Gibbs (2000) to get a result for convergence in the Wasserstein metric. The theorem that follows gives these results for the three cases considered in Gibbs (2000).

Theorem 4.1. *Consider the use of the random scan Gibbs sampler for sampling from the posterior distribution for an image of N pixels in arbitrary dimension where each pixel can take on the values $\{-1, +1\}$ and each interior pixel is influenced by its n nearest neighbors. The algorithm will have converged in the sense that the Wasserstein metric will be less than ϵ at time T for*

$$T > \frac{\ln(\epsilon/N)}{\ln(1 - f(\beta)/N)} \tag{4.2}$$

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 (4.1). Then*

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

for

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

2. *The distribution of interest is the posterior distribution with Ising model prior and the observed image is modeled 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

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

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 modeled 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

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

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, 4, and 5 of Gibbs (2000), it was shown how the expected change in the distance function in one iteration can be bounded by the product of $-\frac{f(\beta)}{N}$ and the current distance. Thus,

$$\begin{aligned}
&\mathbf{E}[d(X_{t+1}^{\max}, X_{t+1}^{\min}) | X_t^{\max}, X_t^{\min}] \\
&\quad = d(X_t^{\max}, X_t^{\min}) \\
&\qquad + \mathbf{E}[d(X_{t+1}^{\max}, X_{t+1}^{\min}) - d(X_t^{\max}, X_t^{\min}) | X_t^{\max}, X_t^{\min}] \\
&\quad \le d(X_t^{\max}, X_t^{\min}) - \frac{f(\beta)}{N} \cdot d(X_t^{\max}, X_t^{\min}).
\end{aligned} \tag{4.3}$$

This expression can then be applied with Theorem 2.1 to give (4.2). 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, 4, and 5 of Gibbs (2000) respectively. □

By Taylor series expansion of the above result, it is easily seen that the bound 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 4.1. *On a finite set $\mathcal{X}$, the following relationship exists between total variation distance and the Wasserstein metric:*

$$d_{\min} TV \ \leq \ W$$

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 $\mathcal{L}(Y) = \nu$,

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

and

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

But

$$d(X, Y) \geq d_{\min} \mathbf{1}_{\{X \neq Y\}}. \tag{4.4}$$

□

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 4.1 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 (4.1). Then*

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

   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 modeled as the true image with each pixel incorrectly observed with probability α. Then*

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

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 modeled as the true image with* $N(0, \sigma^2)$ *noise added independently to each pixel. Then*

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

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 4.1 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 neighbors, 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 36281. In Gibbs (2000) the upper bound on the convergence time was 72246394.

5. Application to Perfect Simulation

Perfect simulation algorithms (Propp and Wilson 1996, Fill 1998) have recently generated a great deal of interest in the Markov chain Monte Carlo literature. See Casella, Lavine and Robert (2000) for a descriptive review. 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 realizations 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 update function of the Markov chain is monotone 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 Gibbs (2000) and Section 3, the image restoration examples are monotone Markov chains.

Our theory 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 perfect simulation 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 a perfect 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. If these chains can be shown to be closer together on average at each iteration, the convergence theory in Section 2 can be used to give an indication of the expected running time.

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. The layered multishift coupler of Wilson (2000) maintains monotonicity. Another approach to achieving exact coalescence on continuous state spaces has been suggested by Neal (1999).

The binary image restoration algorithm of Section 4 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 4.1 give an indication of the required starting time in the past necessary to achieve coalescence at time 0. These are examples of Markov random fields. For the random scan Gibbs sampler used for Markov random fields on discrete state spaces, including the stochastic Ising model of Section 4, Häggström and Nelander (1999) give an $O(N \log N)$ bound on the expected coupling time. For the Ising model, the restrictions on their result correspond to the values of the parameter β for which our result holds.

Acknowledgements. This work is part of my doctoral research under the direction of Jeffrey S. Rosenthal at the University of Toronto. I wish to thank him, Radford Neal, Neal Madras and Francis Su for helpful discussions.

References

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

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

Casella, G., Lavine, M. and Robert, C. (2000). Explaining the perfect sampler. Preprint.

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

Dudley, R. M. (1989). *Real Analysis and Probability*, Chapman and Hall, New York.

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

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

Gibbs, A. L. (2000). Bounding the convergence time of the Gibbs sampler in Bayesian image restoration, *Biometrika*. To appear.

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.

Häggström, O. and Nelander, K. (1999). On exact simulation of Markov random fields using coupling from the past, *Scandinavian Journal of Statistics* **26**: 395–411.

Huber, M. L. (1998). Exact sampling and approximate counting techniques, *Proceedings of the 30th ACM Symposium on the Theory of Computing*. To appear.

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

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.

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

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

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

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.

Rachev, S. T. (1991). *Probability metrics and the stability of stochastic models*, John Wiley and Sons, New York.

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

Thisted, R. (1988). *Elements of Statistical Computing*, Chapman-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.

Wilson, D. B. (2000). Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP), *in* N. Madras (ed.), *Monte Carlo Methods*, Vol. 26 of *Fields Institute Communications*, American Mathematical Society, pp. 141–76.

Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada M3J 1P3

E-mail address: agibbs@mathstat.yorku.ca