Perfect simulation for image restoration

Mark Huber □ *Duke University, Durham, North Carolina, USA*

□ *The coupling method has been an enormously useful tool for studying the mixing time of Markov chains and as the basis of perfect sampling algorithms such as Coupling From the Past. Several methods such as Wilson's layered multishift coupling and Breyer and Roberts' catalytic coupling have been introduced to use the coupling approach on continuous state spaces. This work builds upon these approaches by using a simple coupling for small Metropolis moves together with catalytic coupling. As an application, the analysis of a Markov chain for the autonormal distribution in the Wasserstein metric of A. Gibbs is extended to an analysis in total variation distance. Moreover, a perfect sampling algorithm is constructed that has mean running time $O(N \ln N)$ time for fixed values of the parameters of the model.*

Keywords Coupling, perfect simulation, autonormal model, image analysis

AMS Subject Classification 68U20; 60J10.

This is an Author's Accepted Manuscript of an article published in Stochastic Models July 2007, available online: `http://www.tandfonline.com/10.1080/15326340701471117`.

1 Introduction

Monte Carlo algorithms generate random variates from a target distribution π. These variates can be used as the basis for approximation algorithms for $\#P$ complete problems [15], finding exact p values for statistical tests (see for example [7]), or sampling from the posterior distribution in a Bayesian analysis [8].

In these problems, the target distribution π is typically described through a density whose normalizing constant is unknown. The classical approach to this type of problem is to construct a Markov chain whose stationary distribution matches π, and then run the chain forward for a number of steps. The central question with this method is: how many steps must be taken in the Markov chain before the state is close to stationarity?

Let π and μ be probability distributions on a common measurable space $(\Omega, \mathcal{F})$. A common way of measuring how close π is to μ is the total variation distance:

$$d_{TV}(\pi, \mu) = \sup_{A \in \mathcal{F}} |\mu(A) - \pi(A)|. \tag{1}$$

Let $x_0 \in \Omega$, and $\{X_t\}_{t=0}^{\infty}$ be a Markov chain on $(\Omega, \mathcal{F})$. Then let $\mathcal{L}(X_t|X_0 = x_0)$ denote the distribution of X_t given that the chain started at x_0, and

$$\tau_{x_0}(\epsilon) = \inf\{t : d_{TV}(\mathcal{L}(X_t|X_0 = x_0), \pi) \leq \epsilon\}, \tag{2}$$

is the *mixing time* of the Markov chain. In order for the Markov chain to be used as a means of generating random variates, practitioners need some means for bounding this mixing time.

One commonly used method for upper bounding the mixing time is coupling. A bivariate process (X_t, Y_t) is a *coupling* of a Markov chain if the marginal processes $\{X_t\}_{t=0}^{\infty}$ and $\{Y_t\}_{t=0}^{\infty}$ are both Markov chains with the same transition probabilities as the original Markov chain. The coupling lemma [5, 1] states that if $Y_0 \sim \pi$, then

$$d_{TV}(\mathcal{L}(X_t|X_0 = x_0), \pi) \leq \mathbb{P}(X_t \neq Y_t). \tag{3}$$

Coupling has been used very successfully in discrete spaces (see for example [11, 12]), but A. Gibbs [9] pointed out that in practice it can be difficult to bound for continuous state spaces. Therefore she considered the Wasserstein metric:

$$d_W(\mu, \pi) = \inf \mathbb{E}[d(X, Y)], \tag{4}$$

where d is a metric on Ω, and the infimum is taken over all bivariate random variables (X, Y) where X has distribution μ and Y has distribution ν.

The advantage of $d_W(\mu, \pi)$ for continuous state spaces is that a coupling (X_t, Y_t) need not completely come together in order to get an upper bound. In fact, what Gibbs showed was that if for a coupling (X_t, Y_t) with $X_0 = x_0$ and $Y_0 \sim \pi$ and

$$\mathbb{E}[d(X_{t+1}, Y_{t+1})|X_t, Y_t] \leq c \cdot d(X_t, Y_t), \tag{5}$$

for all t and some $c \in (0, 1)$, then after $(\ln c^{-1})^{-1}(\ln \epsilon^{-1} + \ln \sup_{x,y \in \Omega} d(x, y))$ steps $d_W(\mathcal{L}(X_t|X_0 = x_0), \pi) < \epsilon$. In other words, the fact that the X_t and Y_t processes are coming together exponentially is enough to upper bound Wasserstein distance, having $X_t = Y_t$ is not needed.

Gibbs applied her result to the autonormal distribution that arises in statistical physics and in Bayesian image restoration. More details on this distribution are provided in Section 2. For the Gibbs sampler Markov chain for this distribution, Gibbs showed the Wasserstein distance between $X_t|X_0 = x_0$ and π is at most ϵ after

$$N(\Delta\gamma^2 + \sigma^{-2})(\Delta\gamma^2)^{-1}[\ln(\Delta N) + \ln \epsilon^{-1}] \tag{6}$$

time steps, where $N\Delta$ is a measure of the input size of the problem, and γ and σ are parameters of the model. In addition, Gibbs noted that because the Markov chain studied is monotonic, ideas such as multigamma coupling of Murdoch and Green [13] can be used to create a perfect simulation algorithm for this problem, although she did not complete an analysis of the running time.

In this paper the following results are shown.

1. When a result similar to (5) is possible for monotonic chain, it is possible to obtain a multishift coupler applicable in more situations than the one employed by Wilson [17] by combining the catalytic coupling idea of Breyer and Roberts [3] with a small nonmonotonic Metropolis chain move. This coupler can then be used to create a perfect sampling algorithm for the problem.

2. For a bounded continuous state space like those studied by Gibbs, a result similar to (5) can usually be used to obtain a bound on the total variation distance in addition to the Wasserstein metric bound.

3. For the particular autonormal distribution studied by Gibbs, it is possible to obtain perfect samples in $O(N \ln N)$ expected time for fixed values of σ and γ, and $O(N^2 \ln N)$ expected time for arbitrary σ and γ.

The essential idea of the new procedure is to first use a coupling for a Gibbs sampler Markov chain for a fixed number t of steps. The value of t should be large enough that starting from any possible state, after t steps the resulting Markov chain states are likely to be close to one another. Then one more step is taken using a Metropolis Markov chain that proposes a new state that is likely to be accepted from all of the closely bunched states.

The next section describes the autonormal distribution and image restoration application in detail. Section 3 illustrates the basics of perfect sampling with coupling from the past, and Section 4 shows how a coupling satisfying (5) can serve as the basis of a perfect sampling algorithm by modifying the Markov chain utilized. Section 5 shows how the mixing time for the original Markov chain can be analyzed in terms of total variation distance instead of Wasserstein.

2 Image restoration

In order to illustrate the methodology, consider the following approach to image restoration. Suppose that a greyscale image is represented by assigning each of N pixels a color from 0 to 1, where a white pixel is 0, a black pixel is 1, and in between values are grey. Therefore the state space is $\Omega = [0,1]^N$. The pixels are connected to their nearest neighbor by a set of edges E. Let Δ be the maximum degree of the graph. Typically (N, E) is a 2 dimensional square lattice in which case $\Delta = 4$.

A Bayesian statistical model first puts a prior distribution on the set of configurations. This represents the fact that even though the state of the configuration is unknown before the data/picture is taken, some configurations of pixels are relatively more likely to occur than others. Then data is collected, and the prior distribution conditioned on the data becomes the posterior distribution.

One such prior of Besag [2] puts greater probability on configurations where neighboring pixels are given similar greyscale values. This prior has density

$$\pi_\gamma(x) = Z_\gamma^{-1} \exp\left(- \sum_{\{i,j\} \in E} (1/2)\gamma^2 (x(i) - x(j))^2 \right) \mathbf{1}(x \in [0,1]^N). \tag{7}$$

The parameter γ measures how much influence the neighbors of a node have upon its value. When $\gamma = 0$ the nodes are independent, as γ goes to infinity they become more tightly clustered. A more general model assigns different values of γ to different edges. While the methods presented here can be easily generalized to varying γ, for purposes of presentation only the γ constant across edges case is dealt with here.

The second component of a Bayesian approach is a model of how the data is gathered conditioned on the value of the true picture. In the model used here, independent Gaussian errors are placed on each pixel, conditional on the value of the pixel remaining in $[0,1]$. Hence the density for the distribution of the data given the true picture X is:

$$\pi_\sigma(d|X = x) = Z_{x,\sigma}^{-1} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (d(i) - x(i))^2 \right) \mathbf{1}(y \in \Omega). \tag{8}$$

Combining the two using Bayes' rule gives the final posterior (target) density for the true picture given the data D:

$$\pi(x|D = d) = Z^{-1} \exp(-H(x,d)) \mathbf{1}(x \in \Omega), \text{ where} \tag{9}$$

$$H(x,d) = -\frac{1}{2\sigma^2} \sum_i (x(i) - d(i))^2 - \sum_{\{i,j\}} \frac{1}{2} \gamma^2 (x(i) - x(j))^2. \tag{10}$$

The dependence of the normalizing constant Z on σ, γ, and d is suppressed in the notation.

The random scan Gibbs sampler technique chooses a pixel uniformly at random at each step, and then updates the value for that pixel conditioned on its neighbors. Let $x^{(-i)}$ denote the vector $x(1), x(2), \ldots, x(i-1), x(i+1), \ldots, x(N)$. The form of π is such that the distribution of x given $x^{(-i)}$ will have a normal distribution. Let

$$b(i) = (\sigma^{-2} + n(i)\gamma^2)^{-1}, \quad a(i) = \left[\sigma^{-2} d(i) + \gamma^2 \sum_{j:\{i,j\} \in E} x(j) \right] b(i), \tag{11}$$

where $n(i)$ is the number of neighbors of i in the graph. Then $x(i)$ given $x^{(-i)}$ has the distribution of a normal random variable with mean $a(i)$ and variance $b(i)$ conditioned to lie in $[0,1]$.

3 Perfect simulation using CFTP

Perfect simulation algorithms are a means around the problem of finding the mixing time of a Markov chain. These algorithms draw samples exactly from π, the target distribution, with no measure of closeness needed. Unfortunately, they have the drawback that their running time is itself a random variable that could be arbitrarily large. Therefore these algorithms have expected running times rather than deterministic running times.

The perfect sampling protocol used here is the Coupling From the Past (CFTP) idea of Propp and Wilson [14]. There are many different forms of CFTP [10, 16, 6], here a framework is presented that allows the most flexibility in design. Call $\phi : \Omega \times [0,1] \to \Omega$ stationary with respect to a distribution π if for U uniform over $[0,1]$ and $X \sim \pi$, then $\phi(X,U)$ also has distribution π. For example, taking 17 steps in a Markov chain which has π as its stationary distribution yields a stationary function.

The powerful observation of Propp and Wilson that drives CFTP is that if for the random choice of U, $\phi(\cdot,U)$ maps the entire state space into a single state, then that state has the stationary distribution π. If $\phi(\cdot,U)$ does not map the entire state space into a single state, then recursively find a stationary state and map it forward using the same U in order to generate a stationary state.

CFTP *Output: Y*
1) **Generate** $U \leftarrow \text{Unif}([0,1])$
2) **If** $|\phi(\Omega,U)| = 1$
3) **Let** Y be the sole element of $\phi(\Omega,U)$
4) **Else**
5) **Let** $Z \leftarrow$ **CFTP**
6) **Let** $Y \leftarrow \phi(Z,U)$

Propp and Wilson showed that if $\phi(\cdot,U)$ has positive probability of collapsing to a single state, then the above **CFTP** procedure terminates with probability 1 and outputs a random variate with distribution π. Suppose that $\phi(\Omega,U)$ consists of running a Markov chain forward in time for a fixed number of steps t using the random bits encoded in the uniform random variable U. Let $F_a^b(\cdot,U)$ be a function such that for a Markov chain $\{X_t\}$, $F_a^b(x,U)$ has the same distribution as X_b conditioned on $X_a = x$. Then $\phi(\Omega,U)$ will be a singleton if and only if for the fixed number of steps t, $F_0^t(\Omega,U)$ is a singleton.

One situation where it is possible to quickly determine if $F_0^t(\Omega,U)$ is a singleton is when the function F is monotonic. Suppose that the state space has a partial order $\preceq$ on Ω. Say that F_a^b is *monotonic* if

$$(\forall u \in [0,1])(\forall x,y \in \Omega)(x \preceq y \Rightarrow F_a^b(x,u) \preceq F_a^b(y,u)). \tag{12}$$

That is, if x is smaller than y in a monotonic Markov chain, then after taking a fixed

number of steps in the Markov chain x is still smaller than y. For the autonormal distribution application, it was noted in [9] that the Gibbs sampler run using an inverse cdf method for choosing the new value of a node conditioned on its neighbors yields a monotonic Markov chain under the partial order where $x \preceq y$ if and only if $x(i) \leq y(i)$ for all $i \in \{1, \ldots, N\}$.

Suppose that Ω has a minimum element x_{min} and a maximum element x_{max} under the partial order. Then to test whether or not $|F_a^b(\Omega, U)| = 1$, simply evaluate $F_a^b(x_{min}, U)$ and $F_a^b(x_{max}, U)$. Note for any $x \in \Omega$, $x_{min} \preceq x \preceq x_{max}$. So monotonicity gives

$$F_a^b(x_{min}, U) \preceq F_a^b(x, U) \preceq F_a^b(x_{max}, U).$$

Therefore if $F_a^b(x_{min}, U) = F_a^b(x_{max}, U)$, then this common value is the singleton element of $F_a^b(\Omega, U)$.

For monotonic CFTP to work, there must be a positive chance that $F_a^b(x_{min}, U) = F_a^b(x_{max}, U)$. For the autonormal distribution, the Gibbs sampling chain is monotonic, and $x_{min} = \{0\}^N$ and $x_{max} = \{1\}^N$, so it is set up to utilize monotonicity. However, it is easy to see that running the Markov chain forward from x_{min} and x_{max} using the inverse cdf method, the two chains will never meet one another. They can get very close to one another, but will never match exactly.

Wilson [17] introduced the idea of layered multishift coupling to deal with this problem for certain distributions. Wilson's method applies when the marginal distribution at each node is a shifted (or scaled) distribution whose mean depends on the neighbors of the node. For instance, in the autonormal distribution where the value of each node is unbounded, the marginal distribution is a normal distribution with variance that depends only on the model, and mean that depends on the neighbors.

Unfortunately, in the bounded case, the marginal distributions are no longer simply shifted (or scaled) versions of a reference distribution. Instead, they are shifted normal distributions conditioned to lie between 0 and 1, which means that both the mean and variance of these marginal distributions are changing. Therefore layered multishift coupling cannot be applied here.

4 Perfect simulation using a small Metropolis move

To solve the problem of the chain never quite coalescing, an idea of Breyer and Roberts [3] called catalytic coupling can be used. They noted that instead of running the same Markov chain for a fixed number of steps and letting that be the stationary update ϕ, for continuous state spaces it is helpful to take a number of steps using one Markov chain, and then a final step taken using a different Markov chain which actually brings all the states together completely. They utilized an Independence Sampler chain for their purposes, here a small Metropolis move is introduced and used so that a trivial form of multishift coupling can be utilized.

An alternate approach to the Gibbs sample in designing Markov chains is the Metropolis method, where a proposal state is created from the current state. Let X_t be the current state and $b(y|X_t)$ the density of the proposed point Y. Then $X_{t+1} = Y$ with probability

$$\min\{1, b(X_t|Y)\pi(Y)/[b(Y|X_t)\pi(X_t)]\}, \tag{13}$$

otherwise $X_{t+1} = X_t$.

The new idea in this work is to couple the choice of proposal state together by making a very small uniform move. Here is how it operates in one dimension. Given $X_t = x$, the proposal state Y will have the uniform distribution over $[x - \epsilon, x + \epsilon]$.

In order to deal with the bounded state space, the uniform over $[x - \epsilon, x + \epsilon]$ will be generated in a two step process. First, flip a fair coin. If the coin comes up heads, generate the uniform over $[x, x + \epsilon]$, and on tails generate over $[x - \epsilon, x]$. The reason for this coin flip step will be explained below when multiple dimensions are considered.

Now suppose that X is bounded so that $a \le X \le b$. Suppose that $b - a < \epsilon$ and that the coin flip is heads so that Y is uniform over $[x, x+\epsilon]$. The choice of Y for different values of $x \in [a, b]$ can be coupled in the following fashion, called multishift coupling. Let $\mathcal{L}$ be the set of points $\{a, a + \epsilon, a + 2\epsilon, \ldots\}$, and U be uniform on $[0, \epsilon]$. Then for any x in $[a, b]$, let Y be the point of $\mathcal{L} + U$ that lies in $[x, x + \epsilon]$. This point Y has the uniform distribution over $[x, x + \epsilon]$, moreover, if $U > b - a$ the same choice of Y will be used for every $x \in A$. This happens with probability $1 - (b - a)/\epsilon$. A similar approach is used when the coin flip is tails, to obtain Y uniform over $[x-\epsilon, x]$.

That covers the choice of proposal state–the second piece of Metropolis then accepts or rejects the proposal state with probability (13). In this case $b(y, x) = b(x, y)$ so the probability of accepting the state is just $\pi(x)/\pi(y)$. In the case that $\pi(x) = Z^{-1}f(x)$ for known f and unknown Z^{-1}, this makes the probability that everything moves to the same state in one dimension is at least

$$\left(1 - \frac{b - a}{\epsilon}\right)\left(\min_{a \le x, y \le b+\epsilon} \frac{f(x)}{f(y)}\right). \tag{14}$$

Now consider the multidimensional case, and suppose that a and b are N dimensional vectors where $a(i) \le X(i) \le b(i)$ for all i. Then in some dimensions $a(i)$ is close to 0, in which case the good coin flip from a coupling point of view is to propose a point uniformly over $[X(i), X(i) + \epsilon]$. In other dimensions, $b(i)$ is close to 1, and a good coin flip is to propose a point uniformly over $[X(i) - \epsilon, X(i)]$.

Therefore the choice of coin flip is coupled across dimensions. When the coin flip is heads, then in every dimension, the proposal state is uniform over $[x(i), x(i)+\epsilon]$ if $a(i)$ is close to 0, and is uniform over $[x(i) - \epsilon, x(i)]$ when $b(i)$ is close to 1. This means that when the coin flip is heads, the proposal state tries to move the state away from the boundary in every dimension. Since the proposal state is uniform over

$[x(i), x(i)+\epsilon]$ or $[x(i)-\epsilon, x(i)]$ for each dimension, multishift coupling can be used, as is summarized in the following pseudocode (**SMM** stands for Small Metropolis Move).

SMM *Input:* a, b *Output:* $CoupleFlag$, Y
1) **Let** $H \leftarrow \text{Uniform}(\{-1, 1\})$, **let** $CoupleFlag \leftarrow TRUE$
2) **For** $i \in \{1, \ldots, N\}$ do
3) **Let** $U(i) \leftarrow \text{Uniform}([0, \epsilon])$
4) **Let** $B(i) \leftarrow H\mathbf{1}(b(i) < 1/2) - H\mathbf{1}(b(i) \geq 1/2)$
5) **Let** $Y(i) \leftarrow (a + U(i))\mathbf{1}(B(i) = 1) + (b - U(i))\mathbf{1}(B(i) = -1)$
6) **If** $U(i) < b - a$ then **let** $CoupleFlag \leftarrow FALSE$
7) **End** for
8) **Let** $U \leftarrow \text{Uniform}([0, 1])$
9) **If** $U > f(Y)/\max_{\{x: a \preceq x \preceq b\}} f(x)$ then **let** $CoupleFlag \leftarrow FALSE$

Lemma 4.1 *Suppose that the input vectors* a *and* b *to* **SMM** *satisfy* $b(i) - a(i) < \min\{N^{-2}[1.5\sigma^{-2} + 2.25\gamma^2\Delta]^{-1}, N^{-1}\}$ *for all* i*, and suppose* $\epsilon = N \max_i\{b(i) - a(i)\}$ *in line 3) of* **SMM**. *Then the probability that* **SMM** *coalesces the states (so* $CoupleFlag = TRUE$*) is at least* $(1/2)\exp(-1)$.

Proof. When $H = 1$, this is a "good coin flip" in the sense that in every dimension, the proposal state moves away from the boundary at 0 or 1. This event occurs with probability 1/2. Conditioned on $H = 1$, each node i has a $(1 - (b(i) - a(i))/\epsilon) \geq 1 - 1/N > \exp(-1/N)$ chance of coalescing at the proposal point. There are N nodes, and each choice is independent, so that leaves at least an $[\exp(-1/N)]^N = \exp(-1)$ chance of choosing the same proposal point for each node.

Finally there is the Metropolis acceptance step. Let x satisfy $a \preceq x \preceq b$, and y be any state with $|x(i) - y(i)| \leq \epsilon$. Note $|(y(i) - d(i))^2 - (x(i) - d(i))^2| < 2\epsilon d(i) + \epsilon^2 < 3\epsilon$. Also for any edge $\{i, j\}$, $|(y(i) - y(j))^2 - (x(i) - x(j))^2| < 4\epsilon(x(i) - x(j)) + (1/2)\epsilon$. Hence

$$\begin{aligned} |H(x, d) - H(y, d)| &\leq \frac{1}{2\sigma^2} \sum_i 3\epsilon + \sum_{\{i,j\}} \frac{\gamma^2}{2}(4.5\epsilon) \\ &\leq N1.5\sigma^{-2}\epsilon + N\Delta 2.25\gamma^2\epsilon. \end{aligned}$$

But this last expression is at most 1 from the bound on $b(i) - a(i)$ and choice of ϵ. Hence the chance of accepting y is least $\exp(-1)$, and the lemma is proved.

Now **SMM** is one step in the Markov chain within the larger **CFTP** algorithm from Section 3. When the output of **SMM** has CoupleFlag = TRUE, line 3) from **CFTP** is executed and the output Y of **SMM** is the output of **CFTP**. When CoupleFlag is FALSE, lines 5) and 6) of **CFTP** are executed: the output Y of **SMM**

is not used, but instead **CFTP** is called again recursively by line 5) to generate Z. Line 6) then uses the same updates so that Z is updated to state Y, which is the final output of **CFTP**.

The stationary update ϕ for **CFTP** proposed here has two parts: run the Gibbs sampler monotonic chain forward a fixed number t steps, and then take one step in **SMM**. The chance that this leads to coupling is bounded below by the chance that 1) the Gibbs sampler brings x_{min} and x_{max} close together and 2) **SMM** finishes the job.

Theorem 4.1 *Let*

$$t = N\sigma^2(\sigma^{-2} + \Delta\gamma^2)^{-1}) \ln[2N^2(1.5\sigma^{-2} + 2.25\gamma^2\Delta)] \tag{15}$$

After running the monotonic Gibbs sampler for t steps and then taking one step of **SMM***, there is at least a* $(1/4)\exp(-2)$ *chance that the state has coalesced.*

Proof. A. Gibbs showed in Theorem 3.1 of [9] that running the Gibbs sampler with $X_0 = x_{min}$ and $Y_0 = x_{max}$ gives

$$\mathbb{E}[d(Y_t, X_t)] \leq c^t d(x_{max}, x_{min}), \text{ for } c = 1 - N^{-1}\sigma^{-2}(\sigma^{-2} + \Delta\gamma^2)^{-1}, \tag{16}$$

where $d(x, y) = \sum_{i=1}^{N} \text{degree}(i)|x(i) - y(i)|$. Hence $d(x_{max}, x_{min}) \leq \Delta N$ and if $d(Y_t, X_t) < \delta$ then $|X_t(i) - Y_t(i)| < \delta$ for all i.

Let $\delta = (2N^2[1.5\sigma^{-2} + 2.25\gamma^2\Delta])^{-1}$, so after $\ln \delta^{-1} / \ln c^{-1}$ steps $\mathbb{E}[d(X_t, Y_t)]$ is at most δ. By Markov's inequality there is at least a $1/2$ chance that $d(X_t, Y_t) \leq 2\delta$, which is bound needed by Lemma 4.1 so that **SMM** coalesces with probability at least $(1/2)\exp(-1)$. Using $1/\ln c^{-1} \leq 1/(1-c)$ completes the proof.

Remark 4.1 *This gives an* $O(N \ln N)$ *expected time algorithm for fixed values of* σ*, and* γ*. When* σ *is allowed to vary and becomes large so that the dependence on the data is very weak, simply insert arbitrary data with* $\tilde{\sigma}^{-2} = 1/N$*. Once a sample using* $\tilde{\sigma}$ *is obtained, acceptance/rejection can be used to decide whether to use it as a sample from* σ*. Acceptance occurs with probability at least* $\exp(-1/2)$*, and so this gives an* $O(N^2 \ln N)$ *expected time algorithm for arbitrary* σ *and* γ*.*

5 Total variation bound

Section 4 shows that it is possible to generate perfect samples from the model distribution in (9) in bounded expected time by using a modified Markov chain, but what about the mixing time of the original Gibbs sampler Markov chain? Bounding this mixing time can be useful in obtaining knowledge about the spectral gap of the chain for comparison purposes [4].

In this section it is shown how the ideas of the previous section can be utilized to show a total variation bound in the original Markov chain. The idea is as follows. Again follow an upper (Y_t) and lower (X_t) process and when X_t is close to Y_t change the method by which the marginal distribution is generated at each step. First generate from the marginal distribution exactly, then take one step in the Metropolis Markov chain that has the marginal distribution as its stationary distribution. The Metropolis proposal moves will just be the same as before: flip a fair coin and then based on the coin flip either add or subtract a value uniformly drawn from $[0, \epsilon]$. Using these moves from the last section allows the reuse of the analysis there to prove the following.

Theorem 5.1 *Let $\epsilon_{TV} > 0$, $t_1 := 2N \ln N$, and*

$$t_2 := N\sigma^2(\sigma^{-2} + \Delta\gamma^2) \ln(\Delta N + t_1 + 160Nt_1^2(1.5\sigma^{-2} + 2.25\Delta\gamma^2)) \tag{17}$$

Let Z_t be the random scan Gibbs sampler Markov chain where at each step a node is chosen uniformly at random and updated by a draw from π conditioned on the values of its neighbors. For any $z_0 \in \Omega$, after $t = \lceil 1.6 \ln \epsilon_{TV}^{-1} \rceil (t_1 + t_2)$ steps, the total variation distance between $\mathcal{L}(Z_t | Z_0 = z_0)$ and π is at most ϵ_{TV}.

Proof. Let $z_0 \in \Omega$, and let (X_t, Y_t, Z_t, S_t) be four monotonically coupled Markov chains whose transition probabilities are given by the random scan Gibbs sampler for π, with initial states $X_0 = x_{min}$, $Y_0 = x_{max}$, $Z_0 = z_0$, and $S_0 \sim \pi$. Equation (3) says that the total variation distance between $\mathcal{L}(Z_t | Z_0 = z_0)$ and $\mathcal{L}(S_t) = \pi$ is bounded above by the probability that $Z_t \neq S_t$ which is in turn bounded above by the probability that $X_t \neq Y_t$. The goal of the proof is to show that after t steps this probability is at most ϵ_{TV}.

As mentioned in the paragraph immediately before the theorem statement, the coupling will be monotonic until the gap between the upper and lower processes is small. At this point the coupling switches over so that a single move in the Markov chain is accomplished by choosing a node i uniformly at random followed by a random draw from the marginal distribution, followed by the small Metropolis move that preserves that marginal distribution.

As with the **SMM** moves of the previous section, the Metropolis move here only effectively couples the state space when the moves are away from the boundary. This is where the coin flip comes into play: with probability 1/2 the proposal state moves away from the boundary and coalesces the states. Let τ be the number of steps needed to randomly select each node and have the coin flip go the right way so that the proposal state moves away from the boundary. Then finding $\mathbb{E}[\tau]$ is a variant of the famous Coupon Collector problem. After t_1 steps the chance that a particular node has not been selected with a good coin flip is at most $(1 - 1/[2N])^{t_1} \leq \exp(-t_1/[2N])$. So for $t_1 = 2N(\ln N)$, there is at most an $\exp(-1)$ chance that any of the N nodes have not been selected with the good coin flip.

Recall that $\mathbb{E}[d(X_t, Y_t)]$ is bounded above by $c^t \Delta N$, where $c = 1 - N^{-1}\sigma^{-2}(\sigma^{-2} + \Delta\gamma^2)^{-1}$. Let $\epsilon := [N(20t_1)(1.5\sigma^{-2} + 2.25\Delta\gamma^2)]^{-1}$ Therefore, after $t_2 = N\sigma^2(\sigma^{-2} + \Delta\gamma^2)\ln(\Delta N + t_1 + 20t_1/\epsilon)$ steps, $\mathbb{E}[d(X_t, Y_t)] \leq t_1^{-1}\epsilon(20t_1)^{-1}$ for all t from t_2 to $t_2 + t_1 - 1$. Hence the probability that $d(X_t, Y_t) > \epsilon t_1^{-1}$ for any of these time steps is (by Markov and Bonferroni inequalities) at most $1/20$.

Now consider the probability that each of the steps from time t_2 to $t_2 + t_1 - 1$ is a good event in the sense that every proposal state coalesces, and is accepted if they do coalesce. Let $b - a$ be the difference between the upper and lower process at the chosen node. Then the probability that this move is a good move is at least (see (14) and the proof of Theorem 4.1)

$$\left(1 - \frac{b-a}{\epsilon}\right) \exp(-\epsilon N(1.5\sigma^{-2} + \Delta 2.25\gamma^2)) > \exp(-2(20t_1)^{-1}). \tag{18}$$

Since there are t_1 such moves, the chance that all of them are good moves is at least $\exp(-.1)$, so the chance of failure is at most $1 - \exp(-.1)$.

Combining these three bounds, the total chance of not coupling after $t_1 + t_2$ moves is at most $\exp(-1) + 1/20 + 1 - \exp(-.1) < .52$. If the chain fails to coalesce, begin the process over again independently. Therefore after $t = k(t_1 + t_2)$ steps, the chance of failure to coalesce is at most $.52^{\lfloor k \rfloor}$ steps, and for $k = \lceil \ln \epsilon_{TV} / \ln .52 \rceil$ this is at most ϵ_{TV}. Note $-(\ln .52)^{-1} < 1.6$ to finish the proof.

Remark 5.1 *Whether t_1 or t_2 is larger depends on the size of σ and γ. For fixed σ and γ the mixing time bound is $\Theta(N \ln N)$.*

6 Conclusions

A. Gibbs [9] pointed out that in the continuous state space situation, it is often straightforward to bring a chain started at x_{max} and one started at x_{min} close together, yielding a bound on the mixing time using the Wasserstein metric. This work illustrates that the final piece of bringing them exactly together can be accomplished in an automatic fashion, by utilizing a small Metropolis move with multishift coupling on the proposal state. This not only yields a perfect sampling algorithm, but the idea can also be used to show bounds on the total variation distance (which is a stronger metric than Wasserstein.) Therefore this method is helpful from a practical simulation standpoint as well as in analysis of Markov chains.

Acknowledgments

Support for this work comes from NSF CAREER grant DMS-05-48153. The anonymous referee also provided many helpful comments.

References

1. D. Aldous. Some inequalities for reversible Markov chains. *J. London Math. Soc.*, 25(2):561–576, 1982.

2. J. Besag. On the statistical analysis of dirty pictures. *J. Roy. Stat. Soc. B*, 48:259–302, 1986.

3. L.A. Breyer and G. O. Roberts. Catalytic perfect simulation. *Methodology and Computing in Applied Probability*, 3(2):161–177, 2001.

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

5. W. Doeblin. Exposé de la théorie des chains simples constantes de Markov à un nombre fini d'états. *Rev. Math. de l'Union Interbalkanique*, 2:77–105, 1933.

6. J. A. Fill, M. Machida, D. J. Murdoch, and J. S. Rosenthal. Extension of Fill's perfect rejection sampling algorithm to general chains. *Random Structures Algorithms*, 17:290–316, 2000.

7. P. Galbusera, L. Lens, T Schenck, E. Waiyaki, and E. Matthysen. Genetic variability and gene flow in the globally, critically-endangered Taita thrush. *Conservation Genetics*, 1:45–55, 2000.

8. A.E. Gelfand and A.F.M. Smith. Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, 85:398–409, 1990.

9. A. Gibbs. Convergence in the wasserstein metric for markov chain monte carlo algorithms with applications to image restoration. *Stochastic model*, 20:473–492, 2004.

10. O. Häggström and J. E. Steif. Propp-Wilson algorithms and finitary codings for high noise Markov random fields. *Combin. Probab. Computing*, 9:425–439, 2000.

11. M. Jerrum. A very simple algorithm for estimating the number of k-colourings of a low-degree graph. *SIAM J. Comput.*, 22:1087–1116, 1995.

12. M. Luby and E. Vigoda. Fast convergence of the Gauber dynamics for sampling independent sets. *Random Structures Algorithms*, 15:229–241, 1999.

13. D. J. Murdoch and P. J. Green. Exact sampling from a continuous state space. *Scand. J. Statist.*, 25(3):483–502, 1998.

14. J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures Algorithms*, 9(1–2):223–252, 1996.

15. A. Sinclair. *Algorithms for random generation and counting: a Markov chain approach.* Birkhäuser, 1993.

16. D. B. Wilson. How to couple from the past using a read-once source of randomness. *Random Structures Algorithms*, 16(1):85–113, 2000.

17. D.B. Wilson. Layered multishift coupling for use in perfect sampling algorithms (with a primer on cftp). *Fields Institute Communications*, 26:141–176, 2000.