Exact Sampling with Coupled Markov Chains and Swendsen-Wang Cluster Sampling of the Ising Model

Eric M. Fischer
University of California Los Angeles
Los Angeles, CA 90095
emfischer712@ucla.edu
303 759 361

Abstract

We perform a comparative analysis of two sampling methods: 1) the exact sampling method using coupled Markov chains and the Gibbs distribution proposed by Propp and Wilson in 1996 and 2) a cluster sampling method using the Swendsen-Wang algorithm proposed in 1987. We sample the (2-D) Ising model of ferro-magnetism in statistical mechanics and thus briefly motivate and describe the model. In great detail, we expound the sampling methods as a foundation for understanding the results. Namely, after examining the coalesence times of coupled Markov chains for exact sampling and the convergence times of chains to their sufficient statistics for cluster sampling, it became clear that cluster sampling exhibits much faster convergence rates toward a desired equilibrium distribution π.

1. Introduction

Random sampling has found numerous applications in statistics, computer science, physics, and other fields. In this study, we sample the Ising model and compare exact sampling using a Gibbs sampler to cluster sampling using the Swendsen-Wang algorithm.

First, we define the model from which we sample in this study: the 2-D Ising model. We give an explanation for selecting it and discuss its architecture. We then proceed to outline exact sampling with a Gibbs sampler and cluster sampling with the Swendsen-Wang algorithm. Understanding the motivation for these sampling methods and the problems they overcome will be crucial to understanding the results of this study.

For exact sampling with a Gibbs sampler, we cover three core ideas underlying the exact sampling method: the coalesence of coupled Markov chains, coupling from the past, and monotonicity. We also introduce the Gibbs sampler and then elucidate the exact sampling method used on the Ising model in this study.

For cluster sampling with the Swendsen-Wang algorithm, we first explain the idea of cluster sampling and then discuss the motivation for the Swendsen-Wang algorithm. We review the structure of the algorithm and its clustering and flipping steps in detail, after which we specifically discuss the Swendsen-Wang sampling of the Ising model in this study.

With the key characteristics of our chosen sampling methods expounded, we move on to our specific problem formulation. Here we discuss the details of the experiments we perform for both exact and cluster sampling. We discuss the metrics used for measuring the coalesence of coupled Markov chains in exact sampling and the convergence of chains to their sufficient statistics in cluster sampling with the Swendsen-Wang algorithm.

Results and analysis follow, first presenting the experimental results for exact sampling and then for cluster sampling. Cluster sampling proves to have less demanding sampling requirements for all values of ferro-magnetic strength β used in the Ising model.

2. Ising Model

The Ising model was invented by physicist Wilhelm Lenz, who gave the problem to his student Ernst Ising who solved it in his 1924 thesis. The one-dimensional Ising model has no phase transition though, like the two-dimensional square lattice model used in this study. Considerably more difficult, the two-dimensional model was given an analytic description much later by Lars Onsager in 1944.

The Ising model is a mathematical model of ferromagnetism in statistical mechanics. The model has discrete variables representing magnetic dipole moments of atomic spins that can be in one of two states, 1 or -1. The spins are arranged in a graph, customarily a lattice, so that each spin interacts with its neighbors. Well-suited for experimen-

tal studies, the two-dimensional square lattice Ising model is one of the simplest statistical models to observe a phase transition [10].

To define the Ising model, let $\mathbf{G} = < V, E >$ be a lattice with 4 nearest neighbor connections. Each vertex $v_i \in V$ has a state variable x_i with the label 0 or 1, which is represented by the color black or white, respectively. Let $\mathbf{X} = (x_1, x_2, ..., x_{|V|})$ denote the labeling of the graph.

The Ising model assigns positive energy to spins in opposite directions, i.e. to edges between vertices that have dissimilar labels. Formally, the total energy of the system is given by

$$H(X) = -\sum_{<s,t>\in E} \beta_{st} x_s x_t$$

where E represents all of the 4 nearest neighbors of the lattice and β is the ferro-magnetic interaction strength.

β can be inhomogeneous but is not in this study – we use a consistent positive β value throughout each simulation. A value $\beta \geq 0$ represents a system that prefers similar labels for neighboring vertices.

Accordingly, the probability measure for each possible state of the lattice is

$$\pi(X) = \frac{1}{Z}\exp^{-H(X)}$$

and thus the full Ising model can be defined as follows:

$$\pi(X) = \frac{1}{Z}\exp\{-\sum_{<s,t>\in E} \beta_{st}\mathbf{1}(x_s \neq x_t)\}$$

Notably, the distribution π is an Ising model when the number of possible labels $L = 2$ and a Potts model when $L \geq 3$. We use an Ising model in this study as we have 2 labels: 0 (black) or 1 (white).

3. Exact Sampling

For high-dimensional problems, widely used random sampling methods include Markov Chain Monte Carlo (MCMC) methods like the Metropolis-Hastings method, Gibbs sampling, and slice sampling. One can run an ergodic, i.e. irreducible aperiodic, Markov chain whose stationary distribution is the desired distribution of the set. These methods are guaranteed to produce samples from a target density asymptotically, as long as the Markov chain has converged to the equilibrium, or stationary, distribution π.

Naturally, the principal concern with these methods is how many iterations M the Markov chain should be run to reach the stationary distribution. This can often be very difficult to determine [11]. Presenting a method that solves for this during runtime, James Propp and David Wilson introduced exact sampling with coupled Markov chains in 1996.

Propp and Wilson's *exact sampling method*, also known as perfect simulation or coupling from the past, depends on three key ideas: coalesence of coupled Markov chains, coupling from the past, and monotonicity.

3.1. Coalesence of Coupled Markov Chains

If several Markov chains start from different initial conditions and share a single random-number generator, then their trajectories in state space may *coalesce* and by definition not separate again. If all initializations create trajectories that coalesce into a single trajectory, then it is said the Markov chain "forgets" its initial condition. We refer to the Markov chains as coupled because they share the same random number each sweep of the sampling method and may coalesce [11].

3.2. Coupling from the Past

The coupling from the past procedure returns an exact sample of the equilibrium distribution π of a finite-state, ergodic Markov chain. In principle it gives a perfect, or exact, sample of the equilibrium distribution π [7]. The motivation for coupling from the past comes from the realization that sampling forward in time until coalesence occurs is deficient. The state of a system at the moment coalesence occurs is not guaranteed to be a valid sample of the equilibrium distribution.

Although couplings are key to other sampling methods, in the exact sampling method proposed by Propp and Wilson the coupled chains are uniquely run from a time T_0 in the past up to the present, rather than from the present to a time in the future. Notably, the time T_0 in the past is determined during the running of the algorithm [12].

The idea that we can obtain exact samples by sampling from the a time T_0 in the past up to the present is central to exact sampling. If coalesence occurs, the present coalescent state can be output as an unbiased sample of the equilibrium distribution. If not, one restarts the simulation at a time T_0 further into the past, reusing the same random numbers, and prepends new random moves to the old ones. The simulation is repeated at a sequence of ever more distant times T_0, with a doubling of T_0 from one run to the next commonly serving as a convenient increment. With enough moves prepended, coalesence will occur at a time before the present and we can output $x(0)$ as an exact sample of the equilibrium distribution of the Markov chain [11].

Once T_0 is found, i.e. coalesence for all the Markov chains under review is observed given any state initialization, we can theoretically pick any earlier time and the chains are still guaranteed to coalesce (given the same random numbers). This is because they will ultimately pass through the exact same state T_0, and that state has already

been shown to lead to coalesence.

Thus, we can start from any point further into the past than T_0 and, given any initialization, we will arrive at the same state. The next intuition, then, is wondering how useful this can really be in practice. If we have to simulate chains from *any* initial state, which would be infeasible for most realistic sampling tasks, then what benefit do we receive from this MCMC method? The principal motivation for MCMC methods is to avoid having to visit every state of a system [11].

3.3. Monotonicity

Hence, we have established that exact samples can be guaranteed by simulating forward from a time T_0 in the past, given that coalesence is observed for all possible state initializations from that time. The third key component of the exact sampling method which makes it practical, eliminating the need to test all state initializations, is *monotonicity*.

The key idea is that we can assume coalesence for all state trajectories, i.e. from any state initialization, without actually simulating all of the trajectories. We can do this by taking advantage of a property, often true, that there exists an implicit partial ordering of the state space. Imposing partial ordering in conjunction with coupling, one can determine whether coalesence has occurred by determining whether it has occurred for the two state histories whose initial states were the maximal and minimal elements of the state space. Often, there are indeed a unique maximal and minimal element, so only two state histories need to be simulated [8].

In our current problem of applying a Gibbs sampling method to a ferro-magnetic Ising model, the partial ordering of states can again be defined as

state **x** is "greater than or equal to" state **y** if $x_i \geq y_i, \forall i$

Thus the maximal and minimal states are the all-up (all spins 1) and all-down (all spins -1) states. For this reason, in our experiments we initialize an all-white chain (all labels 1) and an all-black chain (all labels 0). We only need to simulate these two state histories to derive an upper and lower bound for the number of sweeps necessary by the Gibbs sampler to ensure coalesence of the coupled Markov chains. Clearly, only having to run two Markov chains creates significant computational savings.

Summarizing, applying the principles of the coalesence of coupled Markov chains, coupling from the past, and monotonicity, we can ensure that after a finite number of rounds of simulation M of just two coupled Markov chains, our measure $\rho(i)$ of the resulting state i will be sufficiently close to the equilibrium distribution $\pi(i)$ of the chains [5]. That is,

$$||\rho(i) - \pi(i)|| \leq \epsilon$$

And we can hence claim to obtain *exact samples* from the distribution π.

3.4. Gibbs Sampler

The Gibbs sampler an MCMC algorithm for obtaining samples that approximate a given multivariate distribution in cases in which direct sampling is difficult. It was proposed by brothers Stuart and Donald Geman in 1984, roughly eight decades after the passing of Gibbs. Gibbs distributions commonly appear in hard or soft constraint satisfaction problems, e.g. in image denoising or in Bayesian inference. Broadly speaking, it is considered a randomized alternative to deterministic algorithms for statistical inference, such as the expectation-maximization (EM) algorithm.

Gibbs sampling is appropriate when the joint distribution is not known explicitly or is difficult to sample directly, but the conditional distribution of each variable is known and easier to sample. The Gibbs sampler generates an instance of each variable in turn, conditional on the current value of the other variables. The sequence of samples will constitute a Markov chain, and the stationary distribution of the Markov chain is the desired joint distribution [9].

Often the distributions are written in the Gibbs form:

$$\pi(\mathbf{x}) = \frac{1}{Z}\text{exp}^{-E(\mathbf{x})}$$

where $\mathbf{x} = (x_1, ..., x_d) \in \Omega$. The goal of the Gibbs sampler is to sample a joint probability,

$$X = (x_1, x_2, ..., x_d) \sim \pi(x_1, x_2, ..., x_d)$$

It samples in each dimension according to the conditional probability,

$$x_i \sim \pi(x_i|x_{-i}) = \frac{1}{Z}exp(-E[x_i|x_{-i}]), \forall i$$

where $\pi(x_i|x_{-i})$ is the conditional probability at a site i given the other sites.

We are now in position to formally define the Gibbs sampler. Suppose Ω is d-dimensional and each dimension is discretized into L finite states such that the total number of states is L^d. The Gibbs sampling algorithm is as follows:

Input: Probability function $\pi(\mathbf{x})$, current state $\mathbf{x}^{(t)} = (x_1, ..., x_d) \in \Omega$
Output: New state $\mathbf{x}^{t+1} \in \Omega$

1. Select a variable $i \in \{1, ..., d\}$ at random, taking L values $y_1, ..., y_L$
2. Compute the conditional probability vector $\mathbf{u} = (u_1, ...u_L)$ with $u_k = \pi(x_i = v_k|x_{-i})$

3. Sample $j \sim \mathbf{u}$ and set $\mathbf{x}_{-i}^{(t+1)} = \mathbf{x}_{-i}^{(t)}$,$x_i^{(t+1)} = y_j$

The order in which the variables are selected in Step 1 above can be random or follow a predefined schema [5].

A sweep of the Gibbs sampler entails a sequential visit to all of the sites once. Note that although the transition matrix K_i for one Gibbs step may not be ergodic, the total transition matrix $K = K_1 \cdot K_2 \cdots K_d$ is ergodic after one complete sweep.

A well-known problem of the Gibbs sampler is that it has difficulty sampling probability distributions with two tightly coupled variables, or in general for more dimensions, data that is concentrated on a lower-dimensional manifold in a d-dimensional space. This is because to make a sweep through all the sites, the Gibbs sampler updates different dimensions within the state space *independently*. In two dimensions, one can visualize how this would be inefficient if all our data, for example, were focused on a 1-D line. A jagged update pattern forms, when clearly it would be more efficient to make updates that move along the direction of the line [5].

Distributions that can be difficult to sample using the Gibbs sampler include Markov random fields and the Ising/Potts model in particular. In this paper, the key to obtaining an exact sample of the Ising model using a Gibbs sampler lies in exploiting coupling from the past, given an implicit partial ordering of chain states [13].

3.5. Exact Sampling of Ising Model

For our current problem of sampling the Ising model, we can take advantage of a partial ordering with unique maximal and minimal elements:

state **x** is "greater than or equal to" state **y** if $x_i \geq y_i, \forall i$

Thus the maximal and minimal states are the all-up (all spins 1) and all-down (all spins -1) states. For this reason, in our experiments we initialize an all-white chain (all labels 1) and an all-black chain (all labels 0). We only need to simulate these two state histories to derive an upper and lower bound for the number of sweeps necessary by the Gibbs sampler to ensure coalesence of the coupled Markov chains [15].

Due to its high dimensionality, sampling the Ising model is not trivial. The Gibbs sampler updates the chain based on the conditional distribution of each particular spin of the lattice, $P(s/\partial_s)$, where ∂_s represents the 4 nearest neighbors of s. It is very easy to sample this distribution $P(s/\partial_s)$, and it has been demonstrated that if a deterministic or semi-deterministic schema for updating lattice points is used, the induced Markov chain will converge to the joint distribution for the lattice, $P(I)$, i.e. the stationary distribution of the Markov chain $\pi(X)$ [5].

4. Cluster Sampling

In cluster sampling, a researcher divides a population into separate groups called clusters, after which a simple random sample of clusters can be drawn for analysis. Ideally, populations within clusters are as heterogeneous as possible, with homogeneity between clusters. Clusters should also be mutually exclusive and collectively exhaustive. Compared to stratified sampling, cluster sampling draws multiple clusters (but not all of them) for each sample, while stratified sampling draws random samples from *each* strata for each sample.

Cluster sampling has several advantages and disadvantages in comparison with simple random sampling and stratified sampling. A well-known advantage is that cluster sampling reduces computational costs by increasing sampling efficiency. For example, if transitions between clusters are computationally expensive, cluster sampling can be more cost-effective than other methods. This is at the expense of sampling precision, for which stratified sampling is advantageous. Given equal sample sizes, cluster sampling is known to provide less precision than either simple random sampling or stratified sampling [4].

Usually, cluster sampling is performed as either one-stage or two-stage cluster sampling. In one-stage cluster sampling, all of the elements from selected clusters are selected for the sample. In two-stage cluster sampling, just a subset of the elements from selected clusters is randomly selected for the sample.

The Swendsen-Wang algorithm was the first non-local or cluster algorithm for Monte Carlo simulation of large systems near criticality. By near criticality, we mean near a phase transition during which sampling requirements change, usually increasing dramatically. Non-locality refers to the property that in one sweep of the sampling method all spin variables of the system are collectively updated [6].

4.1. Introducing Swendsen-Wang Algorithm

Introduced by Robert Swendsen and Jian-Sheng Wang in 1987, the Swendsen-Wang (SW) algorithm was initially designed to address a well-known, critical slow-down in effective sampling that occurs for the Ising/Potts model. Specifically, near critical temperatures in which phase transitions occur, there is a dramatic increase in the number of samples required to obtain valid random samples from the model [5].

The key feature of the Swendsen-Wang algorithm was the random cluster model, introduced by Kees Fortuin and Piet Kasteleyn in 1969. The random cluster model is a representation of the Ising/Potts model through percolation models of connecting bonds. The SW algorithm has since been generalized by Adrian Barbu and Song-Chun Zhu in 2005 to arbitrary sampling probabilities. This requires interpreting the SW algorithm in a Metropolis-Hastings fashion

by computing acceptance probabilities of proposed Monte Carlo moves [14] [3].

The percolation model is defined as a set of nodes, which is commonly organized onto a lattice structure, in which each node has a label sampled independently from a Bernoulli distribution. In the physics sense, a label 1 represents a pore through which liquid can percolate. During sampling, any two adjacent nodes both assigned a label 1 are automatically connected by their edge. Hence, random clusters of nodes can be obtained by sampling the node labels and automatically connecting adjacent nodes that both have labels 1. In this model, a large pore probability indicates it is very likely a large cluster will form connecting the left and right edges of the lattice. This is what is referred to as percolation [5].

4.2. Swendsen-Wang Algorithm

The Swendsen-Wang algorithm introduces a set of auxiliary variable on the edges. Each edge $e = < s, t >$ is augmented with a binary variable $\mu_e \in \{0, 1\}$. The set of auxiliary variables can be denoted as

$$\mathbf{U} = \{\mu_e : \mu_e \in \{0, 1\}, \forall e \in E$$

An edge e is "turned off", i.e. disconnected, if and only if its auxiliary variable $\mu_e = 0$. μ_e follows a Bernoulli distribution conditional on the vertex labels edge e connects, x_s and x_t:

$$\mu_e|(x_s, x_t) \sim \text{Bernoulli}(q_e \mathbf{1}(x_s = x_t))$$

where $q_e = 1 - e^{-\beta_{st}}$, $\forall e \in E$. Recall that β is the ferromagnetic strength of the system.

From this expression, we can see that $\mu_e = 1$ with probability q_e if $x_s = x_t$, and $\mu_e = 0$ if $x_s \neq x_t$.

With this structure in mind, we can review the two steps the SW algorithm performs each iteration: a clustering step and a flipping step.

In the **clustering step**, given the current state X, the SW algorithm samples the auxiliary variables in $\mathbf{U}$ according to the expression given for $\mu_e|(x_s, x_t)$. This involves several substeps.

First, any edge $e = < s, t >$ with auxiliary variable $\mu_e = 0$ (because its associated vertices $x_s \neq x_t$) is turned off deterministically. After this, the full set of edges can be expressed as

$$E = E_{\text{on}}(\mathbf{X}) \cup E_{\text{off}}(\mathbf{X})$$

Second, the remaining "on" edges E_{on} are turned off with probability $1 - q_{st} = \exp(-\beta_{st})$, dividing them into another "on" and "off" set depending on their respective values for μ_e. As a result, the edge set $E_{on}(\mathbf{X})$ can be further expressed as

$$E_{\text{on}}(\mathbf{X}) = E_{\text{on}}(\mathbf{U}, \mathbf{X}) \cup E_{\text{off}}(\mathbf{U}, \mathbf{X})$$

The edges in $E_{\text{on}}(\mathbf{U}, \mathbf{X})$ will form a number of connected components in which vertices are guaranteed to have the same label, or color. We denote the set of connected components in $E_{\text{on}}(\mathbf{U}, \mathbf{X})$ by

$$\text{CP}(\mathbf{U}, \mathbf{X}) = \{\text{cp}_i : i = 1, 2, ..., K, \text{with} \cup_{i=1}^{K} \text{cp}_i = V\}$$

In the **flipping step**, the SW algorithm randomly selects one connected component $V_o \in$ CP and assigns a common color, or label l, to all vertices, or sites s, in V_0. The new label l follows a discrete uniform distribution

$$x_s = l, \forall s \in V_o$$

where $l \sim \text{uniform}\{1, 2, ..., L\}$. Note, however, that there are only two labels 0 and 1 in this study, so more accurately $l \sim \text{uniform}\{0, 1\}$.

As the set of connected components CP($\mathbf{U}, \mathbf{X}$) is decoupled, one may perform color assignments for some or all of the connected components independently. By making independent updates amongst all the connected components in one sweep, we can draw an analogy with Gibbs sampling. In Gibbs sampling, updates for all the sites, or dimensions, are also made independently in just one sweep of the algorithm [5] [1].

4.3. Swendsen-Wang Sampling of Ising Model

For various values of β, we can observe consecutive samples of the Ising model obtained by the SW algorithm for an example lattice 256x256:

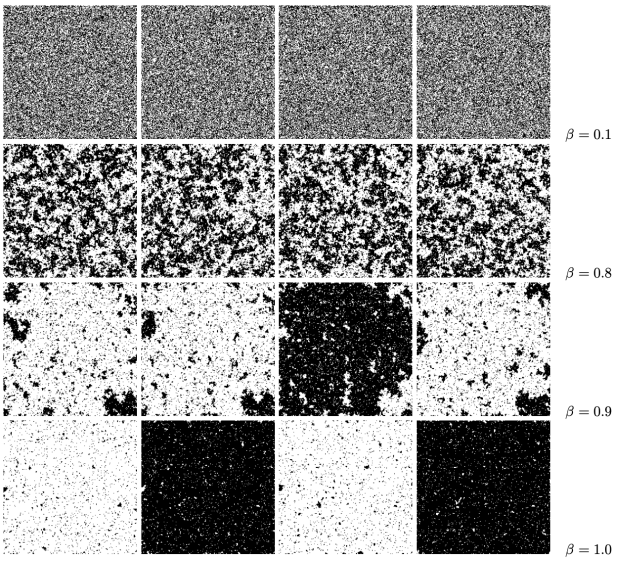

Figure 1. Consecutive samples of Ising model by SW algorithm for different β values. From top to bottom, β = 0.1, 0.8, 0.9, 1 [5].

For small values of β such as 0.1, the SW samples not only appear random, but relatedly, consecutive samples are almost indistinguishable from each other. It reminds us of a television signal after losing connection to a satellite.

For larger values of β such as 1, most vertices within the samples have the same label. And consecutive samples alternate in predominant label, white or black, reflective of the stronger ferro-magnetism strength β. (A larger β corresponds to a system that more strongly prefers similar labels for neighboring vertices.)

Very interestingly, we can also observe a phase transition in the above SW samples. In between 0.8 and 0.9, there is a value β, commonly denoted β_0, that corresponds to a phase transition. We can observe a relatively dramatic change, from a random phase to a solid unicolor phase, in the samples returned between these two values. In physics, the value $\frac{1}{\beta_0}$ is referred to as the critical temperature [5].

5. Problem Statement

In this study we sample a 2-D Ising model using two sampling methods: exact sampling with coupled Markov chains and the Gibbs distribution and cluster sampling with the Swendsen-Wang algorithm. We compare coalescence times of coupled Markov chains using exact sampling to convergence times to sufficient statistics using cluster sampling.

We use an Ising model in a **64**x**64** lattice in which every site has 4 nearest neighbors. We define a state X as the current binary image on the lattice and the variable X_s as a binary value **0** or **1** at a particular pixel, or site s, of the image. The Ising model is again defined as such:

$$\pi(X) = \frac{1}{Z}\exp\{-\sum_{<s,t>\in E} \beta_{st}\mathbf{1}(x_s \neq x_t)\}$$

For both exact and cluster sampling, multiple values of β are used to observe how it influences either coalesence times in exact sampling or convergence times in cluster sampling.

5.1. Exact Sampling

We simulate two coupled Markov chains with the Gibbs sampler. The first chain, whose state is denoted as X^1, is initialized with all sites equal to 1. The second chain, whose state is denoted as X^2, is initialized with all sites equal to 0.

Given this initialization schema, we call the first chain the white chain and the second the black chain. This parallels the convention for RGB pixel values, in which white is encoded as (255, 255, 255) and black as (0, 0, 0).

At each step the Gibbs sampler picks up a site s from both states X^1 and X^2 and calculates their respective conditional probabilities, which are only dependent on their respective 4 nearest neighbors. If we denote a set of 4 nearest neighbor sites as ∂s, then the 2 conditional probabilities calculated at each step can be denoted as:

$$\pi(X_s^1|X_{\partial s}^1) \text{ and } \pi(X_s^2|X_{\partial s}^2)$$

The Gibbs sampler at each step updates the states X_s^1 and X_s^2 according to these two conditional probabilities, using the same random number in [0,1] for both update operations. Under this process the two Markov chains are said to be coupled [5].

5.2. Exact Sampling Coalesence

The statistic of interest is the coalesence time τ: the number of sweeps necessary for the Markov chains X_1 and X_2 to converge to a similar value for the cumulative sum of their respective states. By this time τ, the states X_1 and X_2 are said to represent *exact samples* from the Ising model. The two chain states will simply remain constant, repeatedly determined by the same random number $r \in [0, 1]$ each step.

To evaluate coalesence times, we plot the cumulative sums denoted below of the respective Markov chain states over sweeps of the sampling method.

$$\Sigma_s X_s^1 \text{ and } \Sigma_s X_s^2$$

These sums are the total magnetization at any state of the Ising model for given values of β. We will additionally display an image of the state X when the two chains coalesce.

For varying values of β = [**0.5**, **0.6**, **0.7**, **0.8**, **0.83**, **0.84**, **0.85**, **0.9**] used in the Ising model, we observe plots of τ over β. We will note that once β is increased to a certain value **0.84**, there is a phase transition, causing a dramatic slow-down in effective sampling, requiring many more samples to get an accurate random sample of the distribution [5].

5.3. Cluster Sampling

For the two Markov chains simulated with cluster sampling, the first chain with state X^1 is initialized as a constant black or white image, i.e. with all sites equal to **0** or **1**, respectively. The second chain with state X^2 is initialized as a checkerboard image, forming an alternating pattern of black and white sites with respect to the rows and columns of the lattice.

Accordingly, the first chain has the property $h = 0$, as it is entirely homogeneous with respect to site values. As all site values are equal, we say there are no cracks. Conversely, the second chain initialized as a checkerboard has the property $h = 1$, indicating it possesses the maximal amount of cracks, i.e. it has maximum entropy.

Convergence with cluster sampling is determined by whether $H(X)$, the sufficient statistics of X, converges to a constant value h over time. The underlying idea from

physics is that as long as a lattice nxn is large enough, the probability mass of $\pi(x)$ concentrates around some set uniformly, having zero probability outside of the set. We can denote this set as $\Omega(h)$:

$$\Omega(h) = \{X : H(X) = h\}$$

The sufficient statistics $H(X)$ measures the length of the total boundaries, or cracks, in X and is normalized by the number of edges. $H(X)$ is formally defined as

$$H(X) = \frac{1}{2n^2} \sum_{<s,t>} 1(X_s \neq X_t)$$

We consider two images X_1 and X_2 to have the same probability distribution if their sufficient statistics are equal: $H(X_1) = H(X_2)$.

Theoretically, in the absence of a phase transition, there is a one-to-one correspondence between β (in the Ising model) and h, i.e. $h = h(\beta)$. Thus, again we can empirically diagnose convergence by monitoring whether $H(X)$ converges to a specified constant h over time [5].

5.4. Cluster Sampling Convergence

To evaluate convergence to a sufficient statistic using cluster sampling, we use three values of β in the Ising model: **0.6**, **0.8**, and **0.84**. This lends three images X_1, X_2, and X_3 for the lattice states at the respective convergence times t_1, t_2, and t_3, corresponding to the three different values of β. As we used all three β values for the exact sampling experiment as well, we will be able to make direct comparisons.

From these images X_1, X_2, and X_3 we compute their respective sufficient statistics h_1^*, h_2^*, and h_3^*. These sufficient statistics give the value at which the coupled Markov chains meet. When the two chains meet at h_i^*, one initialized as a constant black or white image and the other as a checkerboard image, we will believe they have converged to $\Omega(h_i^*)$.

Hence for cluster sampling, we plot the sufficient statistics $H(X)$ of the current state that we denote as $X(t)$ over the time, or sweeps, t. Convergence, and a stopping of the sampling method, is determined by h approximating within a certain distance ϵ the sufficient statistic h_i^*. We set ϵ to be **0.001**.

We compare the cluster sampling convergence times for $\beta = [0.6, 0.8, 0.84]$ to the exact sampling method using the Gibbs sampler. Note that this comparison may be slightly unfair to the Gibbs sampler, as it is possible it converges to some sufficient statistics $\Omega(h_i^*)$ before coalesence in some cases [5].

Lastly, we will plot the average sizes of the connected components (CP), or the number of pixels flipped together at each sweep, for each of the three values of β: β_1, β_2, and β_3.

6. Results and Analysis

We first observe results for exact sampling with coupled Markov chains and the Gibbs sampler, and then observe results for cluster sampling using the Swendsen-Wang algorithm.

6.1. Exact Sampling

For each of the values of ferro-magnetic strength β = [**0.5**, **0.6**, **0.7**, **0.8**, **0.83**, **0.84**, **0.85**, **0.9**] used in the Ising model, we display two figures: a plot of the coalesence of the coupled Markov chains and a sample of the Ising model at coalesence. At coalesence or after, the image samples are said to be *exact samples* from the Ising model.

The plots of coalesence have on the y-axis the total magnetization $\sum_s X_s$ of the Ising model and on the x-axis the number of sweeps τ of the sampling method. Recall that the total magnetization $\sum_s X_s$ is the cumulative sum of the states X_s of a Markov chain over sweeps. We expect the respective total magnetizations for coupled Markov chains to converge in value, causing the chains to ultimately move in a nearly identical fashion, signifying coalesence.

To be concrete, when two chains meet each other such that $X_s^1 = X_s^2, \forall s$ after many sweeps, they have coalesced. They then remain in the same state permanently, as due to the nature of the Gibbs sampler, the chains are determined by the same random number $r \in [0, 1]$ at each step.

As a last note for the plots, as the white chain X^1 was initialized with all sites equal to 1, and the black chain X^2 was initialized with all sites equal to 0, it can be shown by induction that $X_s^1 \geq X_s^2, \forall s$ in any step of the sampling method. Accordingly, we label the white chain as the upper bound and black chain as the lower bound for the sum of an image $\sum_s X_s$ for a given sweep, or iteration, of the sampling method [5].

6.2. Coalescence and Samples

For β = **0.5**, we display below the coalesence of the white and black coupled Markov chains. They coalesce at τ = **25** sweeps of the exact Gibbs sampling method.

This smallest value of $\beta = 0.5$ lends the smallest coalesence time $\tau = 25$ observed. This is intuitive, as smaller values of beta promote clustering less, which in turn causes the image sums of the coupled chains to converge fastest to a similar value.

To expound, recall larger values of $\beta \in [0, 1]$ configure Ising models to more strongly promote clustering, i.e. prefer similar labels for neighboring vertices. And recall the coupled Markov chains are initialized as white and black images, with image sums of 4096 and 0 respectively (as we use 64x64 lattices labels 1 and 0). The smallest value of β hence lends a model that, due to promoting clustering the least, allows the sites of the two images, at first entirely 0

or 1, to more frequently make color updates. This in turn causes the image sums to coalesce more quickly and reach a similar value in between 0 and 496. Simply put, the chains can more quickly meet each other in the middle due to less of an influence of the ferro-magnetic strength β.

The weaker preference for clustering is reflected in the Ising model sample drawn at coalesence, which displays less clustering of the white and black sites than the samples for other values of β.

Hence, the coupled chains for $\beta = 0.5$ most quickly reach a similar value for their respective total magnetizations, i.e. the sum of their respective images, due to a weaker preference for clustering which in turn causes a quicker convergence of the image sums.

Below we display a sample of the Ising model at coalesence for $\beta = 0.5$ and $\tau = 25$. Note this sample displays the least clustering of the white and black sites of any in this study, reflective of the smallest β value used for this study.

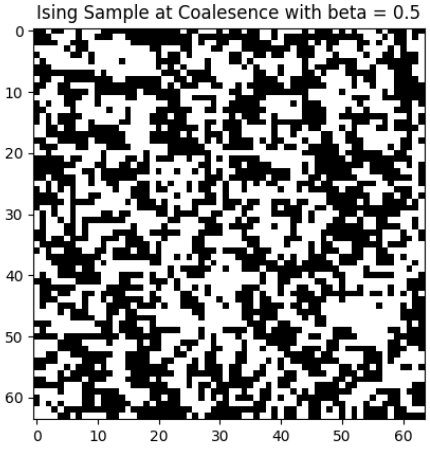

Figure 2. Ising model sample for $\beta = 0.5$, $\tau = 25$

For $\beta = \mathbf{0.6}$, we display the coalesence of the white and black coupled Markov chains. They coalesce at $\tau = \mathbf{53}$ sweeps of the exact Gibbs sampling method.

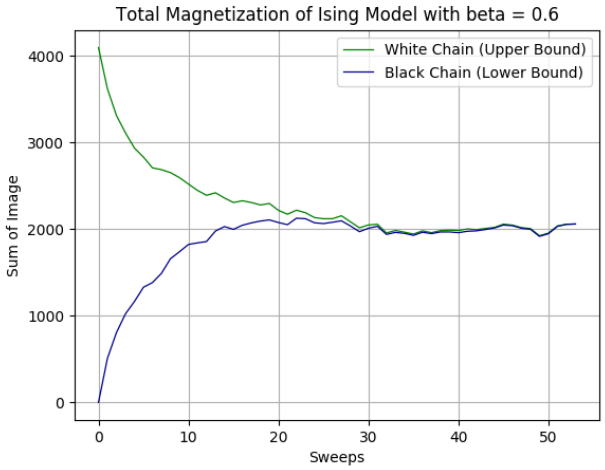

Figure 3. Coalesence at $\tau = 53$ sweeps for $\beta = 0.6$

Below we display a sample of the Ising model at coalesence for $\beta = 0.6$ and $\tau = 53$.

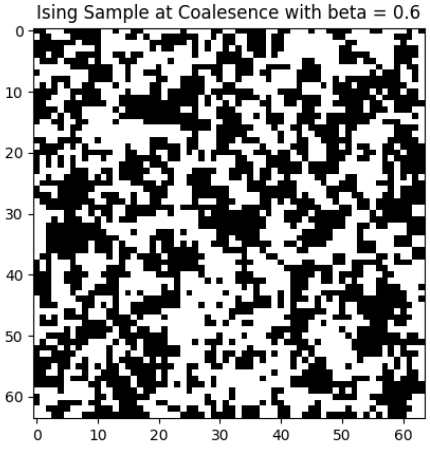

Figure 4. Ising model sample for $\beta = 0.6$, $\tau = 53$

For $\beta = \mathbf{0.7}$, we display the coalesence of the white and black coupled Markov chains. They coalesce at $\tau = \mathbf{69}$ sweeps of the exact Gibbs sampling method.

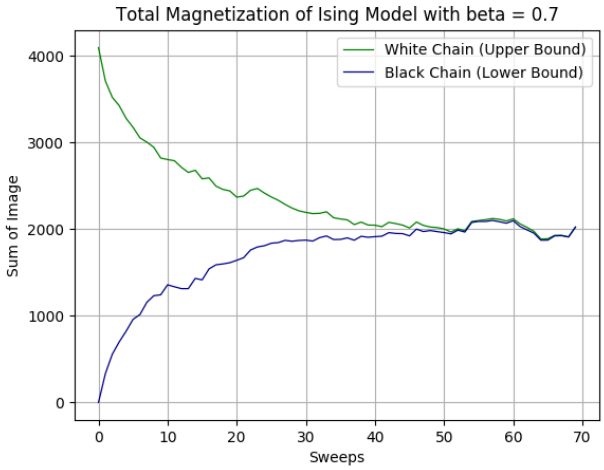

Figure 5. Coalesence at $\tau = 69$ sweeps for $\beta = 0.7$

Increasing β from 0.5 to 0.7 has so far not caused any dramatic increases in the number of sweeps necessary for the coupled chains to coalesce.

Below we display a sample of the Ising model at coalesence for $\beta = 0.7$ and $\tau = 69$. Now how the samples, as values of β increase, become increasingly clustered.

The larger ferro-magnetism strength β at this point is still just gradually encouraging more neighboring vertices to have similar labels.

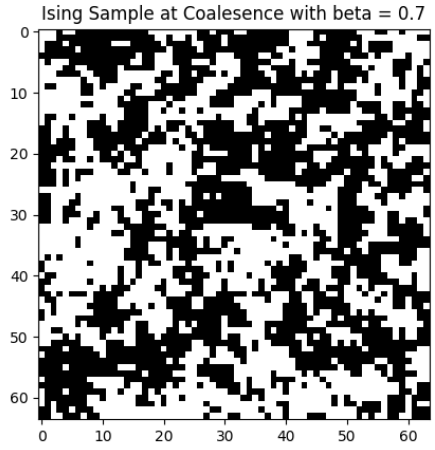

Figure 6. Ising model sample for $\beta = 0.7$, $\tau = 69$

For $\beta = \mathbf{0.8}$, we display the coalesence of the white and black coupled Markov chains. They coalesce at $\tau = \mathbf{458}$ sweeps of the exact Gibbs sampling method.

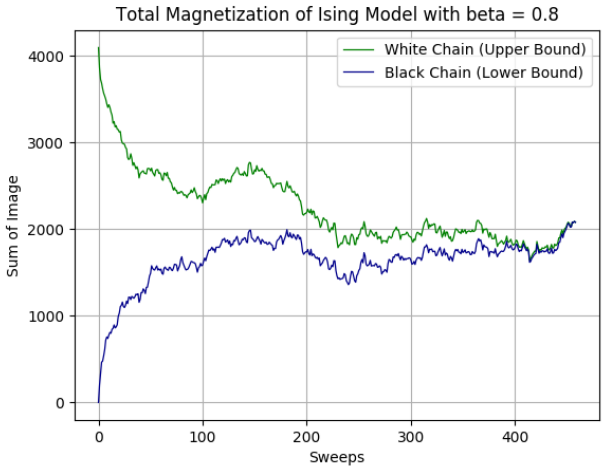

Figure 7. Coalesence at $\tau = 458$ sweeps for $\beta = 0.8$

Increasing β from 0.7 to 0.8 caused a larger relative increase in the number of sweeps necessary for coalesence, as compared to the transition from 0.6 to 0.7.

Below we display a sample of the Ising model at coalesence for $\beta = 0.8$ and $\tau = 458$.

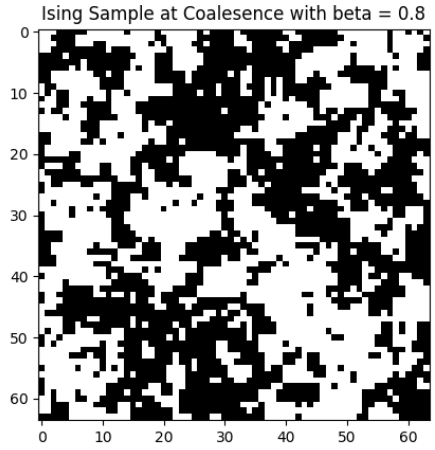

Figure 8. Ising model sample for $\beta = 0.8$, $\tau = 458$

For $\beta = \mathbf{0.83}$, we display the coalesence of the white and black coupled Markov chains. They coalesce at $\tau = \mathbf{372}$ sweeps of the exact Gibbs sampling method.

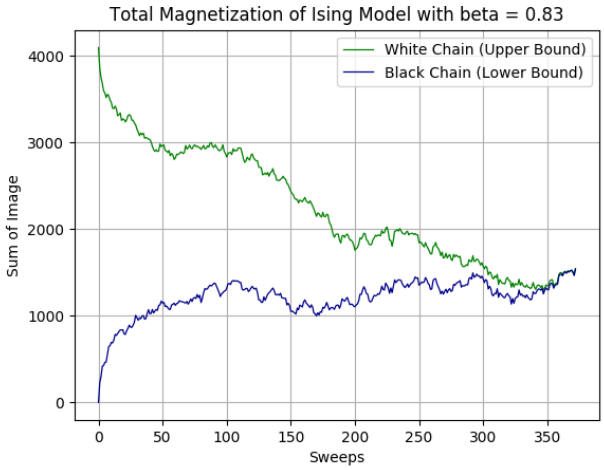

Figure 9. Coalesence at $\tau = 372$ sweeps for $\beta = 0.83$

Below we display a sample of the Ising model at coalesence for $\beta = 0.83$ and $\tau = 372$.

Note the number of sweeps necessary for coalesence is slightly lower here for $\beta = 0.83$ as opposed to $\beta = 0.8$, but this is merely due to random chance. The Gibbs sampler uses a (shared) random number in [0,1] for each sweep in which it updates the states X^1 and X^2. On most runs of the exact Gibbs sampler through the increasing β values, the number of sweeps necessary for coalesence increased monotonically.

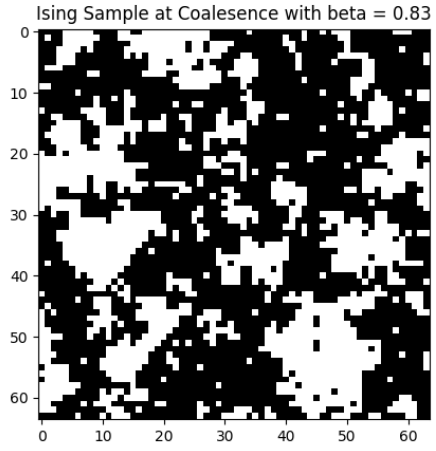

Figure 10. Ising model sample for $\beta = 0.83$, $\tau = 372$

For $\beta = \mathbf{0.84}$, we display the coalesence of the white and black coupled Markov chains. They coalesce at $\tau = \mathbf{887}$ sweeps of the exact Gibbs sampling method.

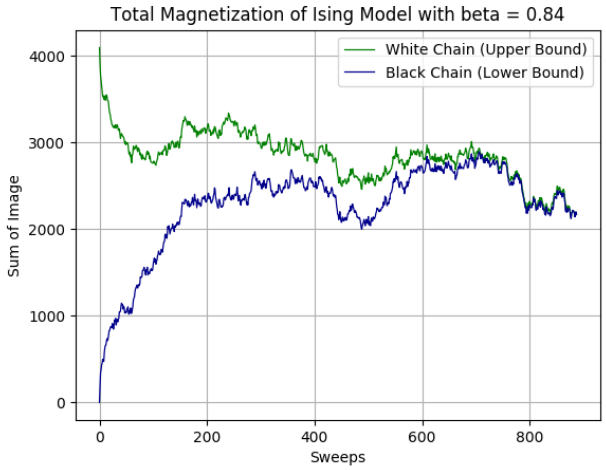

Figure 11. Coalesence at $\tau = 887$ sweeps for $\beta = 0.84$

Increasing β from 0.83 to 0.84, a seemingly small transition, caused a roughly 2.5x increase in the number of sweeps necessary for coalesence. Roughly at this value of $\beta = 0.84$, the Ising model undergoes a phase transition. The sampling demands have increased greatly for just a small increase in the ferro-magnetism strength β.

Below we display a sample of the Ising model at coalesence for $\beta = 0.84$ and $\tau = 887$.

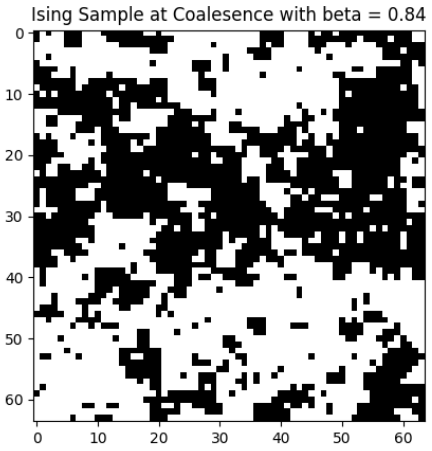

Figure 12. Ising model sample for $\beta = 0.84$, $\tau = 887$

For $\beta = \mathbf{0.85}$, we display the coalesence of the white and black coupled Markov chains. They coalesce at $\tau = \mathbf{883}$ sweeps of the exact Gibbs sampling method.

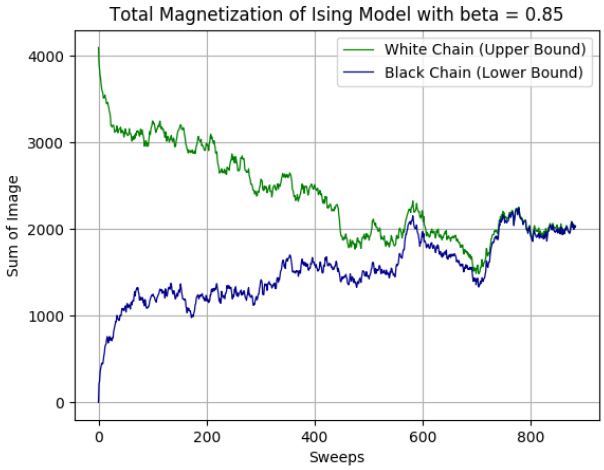

Figure 13. Coalesence at $\tau = 883$ sweeps for $\beta = 0.85$

By chance, this value $\tau = 883$ sweeps is close to the value of $\tau = 887$ sweeps for the previous value of $\beta = 0.84$. We can still be confident the system undergoes a clear phase transition by observing the number of sweeps τ necessary for $\beta = 0.9$, next.

Below we display a sample of the Ising model at coalesence for $\beta = 0.85$ and $\tau = 883$.

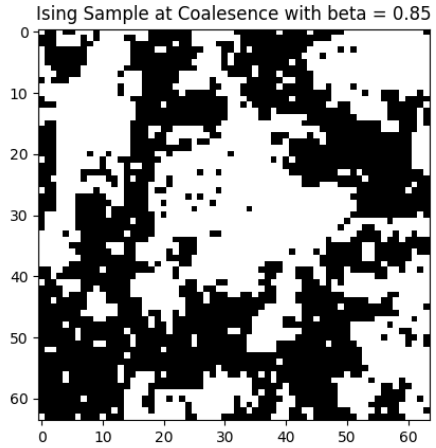

Figure 14. Ising model sample for $\beta = 0.85$, $\tau = 883$

For $\beta =$ **0.9**, we display the coalesence of the white and black coupled Markov chains. They coalesce at $\tau =$ **15330** sweeps of the exact Gibbs sampling method.

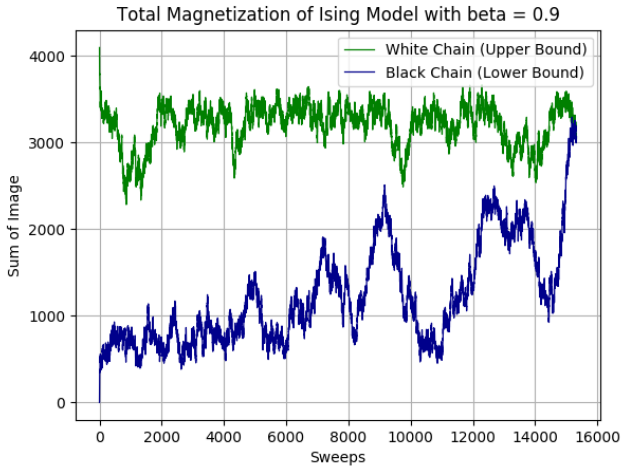

Figure 15. Coalesence at $\tau = 15330$ sweeps for $\beta = 0.9$

This is a very large number of sweeps required for coalesence compared to any previous β value and confirms the Ising model undergoes a phase transition before this $\beta = 0.9$ value.

Below we display a sample of the Ising model at coalesence for $\beta = 0.9$ and $\tau = 15330$.

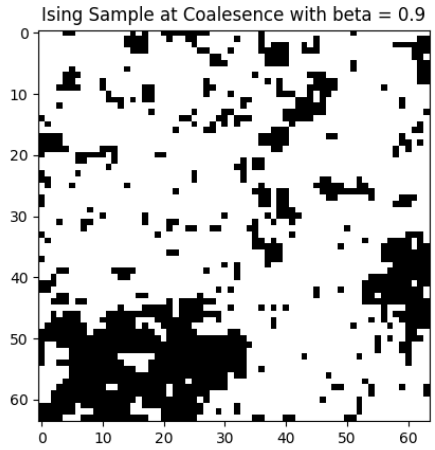

Figure 16. Ising model sample for $\beta = 0.9$, $\tau = 15330$

Lastly, we display a plot of the coalesence times τ for each of the values of $\beta =$ [**0.5**, **0.6**, **0.7**, **0.8**, **0.83**, **0.84**, **0.85**, **0.9**] used in the Ising model.

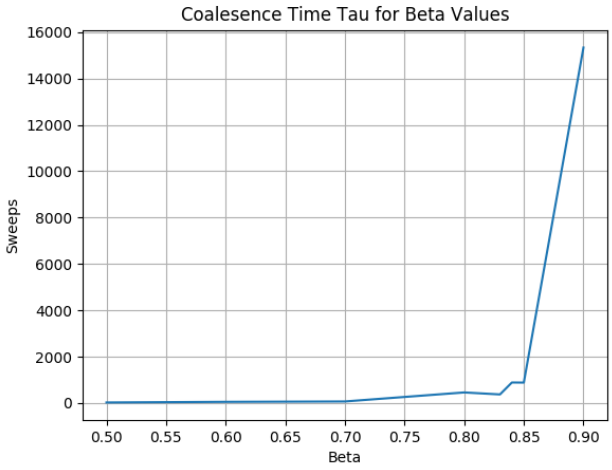

Figure 17. Coalesence times tau for beta values

This final plot of τ over β shows that at about $\beta = 0.85$, a phase transition occurs which causes a critical slow-down in the sweeps required for coalesence. In effect, many more sweeps of the Gibbs sampler will be necessary for β values larger than 0.85 to ensure we have accurate random samples of the distribution.

6.3. Cluster Sampling

Recall that for cluster sampling we use a different initialization. The first Markov chain with state X^1 is initialized as a constant black or white image, and the second chain with state X^2 is initialized as a checkerboard image.

Convergence with cluster sampling is determined by whether $H(X)$, the sufficient statistics of X that measures the length of total boundaries (or cracks), converges to a constant value h over time.

We use the following values of β in the Ising model for cluster sampling: **0.6**, **0.8**, and **0.84**. These lend three images X_1, X_2, and X_3 for the lattice states at the respective coalesence times t_1, t_2, and t_3. As we used these same values of β for exact sampling, we will be able make a direct comparison.

From these images X_1, X_2, and X_3 corresponding to the different β values, we compute their respective sufficient statistics h_1^*, h_2^*, and h_3^*. These are equal to **0.3194**, **0.2231**, and **0.1966**, respectively.

These sufficient statistics give the value at which the Markov chains meet. Again, when the two chains meet at h_i^*, one initialized as a constant black or white image and the other as a checkerboard image, we believe they have converged to $\Omega(h_i^*)$.

6.4. Convergence using Sufficient Statistics

We first plot the sufficient statistics $H(X)$ of the current state $X(t)$ over time, or sweeps, t. Recall that convergence, and a stopping of the sampling method, is determined by h approximating within a certain distance ϵ the sufficient statistic h_i^*. We again set $\epsilon = \mathbf{0.001}$.

Below we plot the sufficient statistics $H(X)$ over sweeps for $\beta = \mathbf{0.6}$. We do this for both initializations, the constant image and checkerboard image. We placed red markers for the times at which the Markov chains converge to the sufficient statistic $h_1^* = \mathbf{0.3194}$ given this beta value.

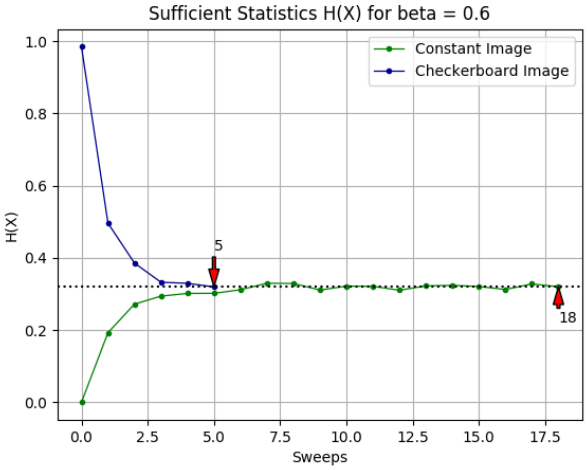

Figure 18. Constant and checkerboard images converge in 18 and 5 sweeps respectively to sufficient statistic $h_1^* = 0.3194$ for $\beta = 0.6$

Note the much faster convergence compared to the exact sampling method, which required $\tau = 53$ sweeps for the coupled Markov chains to coalesce given $\beta = 0.6$.

Also note how the constant (white or black) image requires more sweeps to converge, as confirmed by all the following plots as well. This is intuitive, as the constant image should exhibit a greater mixing time to reach the equilibrium distribution than a checkerboard image. As the checkerboard image is initialized in a maximally mixed fashion (equal white and black sites), it should display faster convergence for most equilibrium distributions.

Below we plot the sufficient statistics $H(X)$ over sweeps for $\beta = \mathbf{0.8}$. We do this for both initializations, the constant image and checkerboard image. We placed red markers for the times at which the Markov chains converge to the sufficient statistic $h_2^* = \mathbf{0.2231}$ given this beta value.

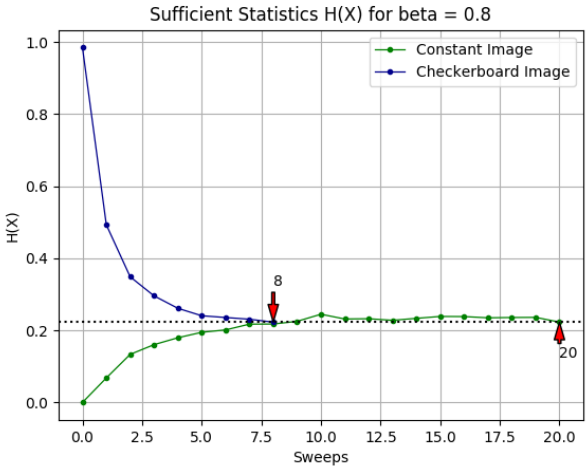

Figure 19. Constant and checkerboard images converge in 20 and 8 sweeps respectively to sufficient statistic $h_2^* = 0.2231$ for $\beta = 0.8$

Similar to the results seen with exact sampling, notice how a larger value of β increases the sweeps τ required for convergence. This is to be expected, as an increase in the ferro-magnetism strength β more strongly configures neighboring vertices to have similar labels, encouraging clustering and slowing the mixing time of the chains.

Lastly, we plot the sufficient statistics $H(X)$ over sweeps for $\beta = \mathbf{0.84}$. We do this for both initializations, the constant image and checkerboard image. We placed red markers for the times at which the Markov chains converged to the sufficient statistic $h_3^* = \mathbf{0.1966}$ given this beta value.

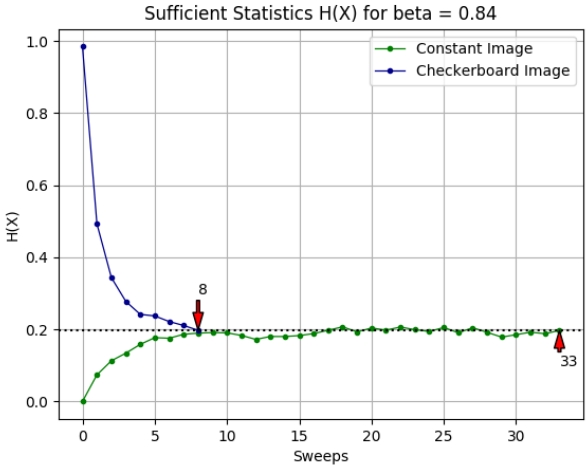

Figure 20. Constant and checkerboard images converge in 33 and 8 sweeps respectively to sufficient statistic $h_3^* = 0.1966$ for $\beta = 0.84$

This is the value of $\beta = 0.84$ for which we previously observed a phase transition using the Gibbs sampler to perform exact sampling. Note that in cluster sampling, the number of sweeps required for convergence does not dramatically increase around this value of $\beta = 0.84$.

6.5. Comparison with Exact Sampling

Now we can compare the exact sampling coalesence times for the three values of $\beta = [0.6, 0.8, 0.84]$ to the convergence times found for cluster sampling with SW.

Using the Gibbs sampler for exact sampling, the Markov chains coalesced in 53, 458, and 887 sweeps for the three values of β, respectively. For cluster sampling, the slowest convergence times for each β value were for the constant image: 18, 20, and 33 sweeps. Overall, we can see that cluster sampling gives much faster convergence rates than does the exact sampling method with coupled Markov chains.

Note, however, that this comparison is slightly unfair to the Gibbs sampler, as it is possible that either or both of the coupled Markov chains in the exact sampling experiments converge to the respective sufficient statistic $\Omega(h_i^*)$ *before* coalesence [5]. Coalesence displayed by coupled Markov chains and convergence to a sufficient statistic are two different ways to confirm one is sampling from an equilibrium distribution π. In any case, because the disparity is so large, it is safe to assume that cluster sampling does indeed provide faster convergence rates than exact sampling with coupled Markov chains.

6.6. Average Connected Component

Lastly, we also plot the average sizes of the connected components (CP), or the number of pixels flipped together at each sweep, for each of the three values $\beta = 0.6$, 0.8, and 0.84.

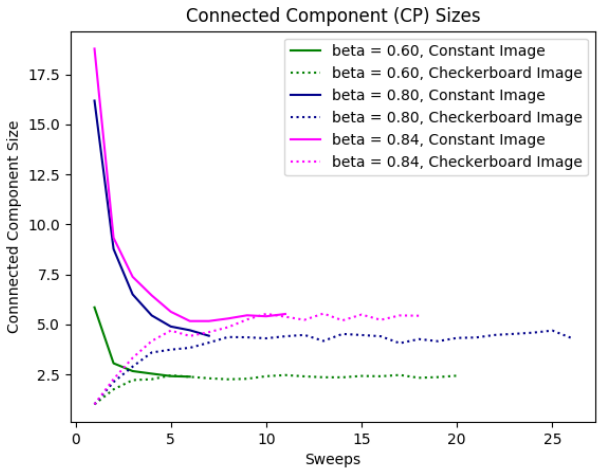

Figure 21. Connected component (CP) sizes for β = 0.6, 0.8, and 0.84

The outstanding observation is that as β increases from 0.6 to 0.84, the average CP size also increases. This is intuitive, as we know that a stronger ferro-magnetic strength β

more strongly promotes neighboring vertices to have similar labels and hence it promotes clustering. With larger clusters on average, the average CP size, or the average number of pixels flipped together at each sweep, will certainly be larger.

It also seems that larger values of β correspond to greater variation in average CP sizes. The value $\beta = 0.84$ caused the most variation and the largest number of sweeps necessary for the constant and checkerboard images to converge to roughly the same CP size per sweep. This is also reasonable, as stronger β values promote stronger clustering, which promotes the formation of more and larger clusters, which in turn produces greater variation in cluster sizes.

7. Future Work

The Swendsen-Wang clustering method is limited in two ways we have not mentioned. First, it is only valid for the Ising and Potts models. Second, it requires that the number of labels, or colors, L be known. In applications such as image analysis, L may represent the number of objects (or image regions) that have to be inferred from input data.

Hence for future work, we would like to explore Data Driven Markov Chain Monte Carlo (DDMCMC) methods, which do not require that the number of labels L be known. Utilizing a Bayesian statistical framework, DDMCMC methods perform image segmentation and hence labeling in a purely data-driven manner [2].

8. Conclusion

We have carried out a full implementation and comparative analysis of exact sampling with coupled Markov chains proposed by Propp and Wilson and cluster sampling with the Swendsen-Wang algorithm. Understanding the motivation behind the methods, their use-cases, their weaknesses, and other key qualities helped us to differentiate and further characterize these methods. Cluster sampling with the Swendsen-Wang algorithm proved certainly that it is a superior method for sampling the 2-D Ising model with regard to computational expense.

References

[1] Adrian Barbu. Cluster sampling and its applications to segmentation, stereo and motion. https://www.researchgate.net/publication/239155422_Cluster_Sampling_and_its_Applications_to_Segmentation_Stereo_and_Motion.

[2] Adrian Barbu and Song-Chun Zhu. Generalizing swendsen-wang for image analysis. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 27, August 2005. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.162.2784&rep=rep1&type=pdf.

[3] Adrian Barbu and Song-Chun Zhu. Generalizing swendsen-wang to sampling arbitrary posterior probabilities. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 27, August 2005. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.454.965&rep=rep1&type=pdf.

[4] Adrian Barbu and Song-Chun Zhu. Interpretations of cluster sampling by swendsen-wang. 2013. https://pdfs.semanticscholar.org/4256/35f354bab06f3811eb0093d67ba37445fe12.pdf.

[5] Adrian Barbu and Song Chun Zhu. *Monte Carlo Methods*. Springer, March 2019. http://www.stat.ucla.edu/~sczhu/courses/ucla/Stat_202C/MCMC_book.pdf.

[6] Cluster Sampling. Cluster sampling — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Cluster_sampling.

[7] Coupling from the Past. Coupling from the past — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Coupling_from_the_past.

[8] F. Friedrich, G. Winkler, O. Wittich, and V. Liebscher. An elementary rigorous introduction to exact sampling. *Institute of Biomathematics and Biometry*, 2003. https://core.ac.uk/download/pdf/12162754.pdf.

[9] Gibbs Sampling. Gibbs sampling — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Gibbs_sampling.

[10] Ising Model. Ising model — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Ising_model.

[11] David J.C. MacKay. Exact monte carlo sampling. *Cambridge University Press*, 2003. http://www.inference.org.uk/mackay/itprnn/ps/413.435.pdf.

[12] James Gary Propp and David Bruce Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Massachusetts Institute of Technology*, July 1996. https://www.stat.berkeley.edu/~aldous/206-RWG/RWGpapers/propp_wilson.pdf.

[13] Julien Stoehr1. A review on statistical inference methods for discrete markov random fields. https://arxiv.org/pdf/1704.03331.pdf.

[14] Swendsen-Wang Algorithm. Swendsen-wang algorithm — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Swendsen-Wang_algorithm.

[15] David Bruce Wilson. Exact sampling with markov chains. *Massachusetts Institute of Technology*, June 1996. https://dspace.mit.edu/bitstream/handle/1721.1/38402/36023178-MIT.pdf;sequence=2.