2024

Probing the Ising Model’s Thermodynamics through Restricted Boltzmann Machines

Xiaobei (Emma) Zhang

Probing the Ising Model's Thermodynamics through Restricted Boltzmann Machines

Xiaobei (Emma) Zhang

Weiqing Gu, Advisor

Daniel Tamayo, Reader

HARVEY MUDD COLLEGE

Department of Mathematics

May, 2024

Abstract

This thesis explores the connection between physics and machine learning by using Restricted Boltzmann Machines (RBMs) to study the thermodynamic properties of the Ising model. The Ising model is a simple but realistic model that captures the magnetic behavior of a system, where spins occupy a lattice of sites and different spin configurations correspond to different energies. The model exhibits phase transitions between ferromagnetic and paramagnetic phases as a function of temperature. RBMs are two-layered neural networks that can learn probability distributions over binary spins.

The study generates 2D Ising model data at different temperatures using Monte Carlo simulations, including the Metropolis algorithm and the Wolff algorithm. RBMs are trained on this data and validated by studying the learned weights and filters. We then use the trained RBMs to generate new Ising configurations. The quality of the RBM-generated configurations is assessed by comparing their probability distributions to those of the original configurations using the Wasserstein distance, a measure from optimal transport theory.

Interestingly, the Wasserstein distance between the generated and original configurations shows an unexpected trend, with lower values around the critical temperature and a sharp dip at $T = 2.0$. This suggests that the RBM is able to capture important features of the Ising model's thermodynamics, particularly near the phase transition. The next steps are to further investigate this finding, such as exploring the learned features in the RBM's hidden layer and generating configurations with more hidden units. Overall, this work demonstrates a promising approach for connecting physics and machine learning to gain insights into complex systems.

Contents

Abstract **iii**

Acknowledgments **ix**

1 Introduction **1**
1.1 The Ising Model 2
1.2 Probabilities in the Ising Model 4
1.3 Neural Networks 5
1.4 Restricted Boltzmann Machines 6
1.5 Overview 8

2 Background **9**
2.1 Statistical Thermodynamics of the Ising Model 9
2.2 Training RBMs 12
2.3 Optimal Transport in Machine Learning 14
2.4 Related Literature 17

3 Methods **19**
3.1 Generating 2D Ising Datasets 19
3.2 Training Specifications of RBM 24

4 Results **27**
4.1 Ising Model Simulations 27
4.2 RBM Training Results 30
4.3 RBM-Generated Ising Configurations 33

5 Conclusion **39**
5.1 Discussion of Key Results 39
5.2 Limitations 40

5.3 Future Work 41

A Thermodynamics of 2D Ising Model 43

B Energy and Probability of RBMs 45

Bibliography 47

List of Figures

1.1 A binary image created on a 10×10 lattice. 2

1.2 An example of an Ising model configuration. The spin variable assigned to a lattice site is either upward (indicated by green plus signs), $s_i = +1$, or downward (indicated by red minus signs), $s_i = -1$. 3

1.3 Diagram of a neural network. The layer with neurons colored red is the input layer that receives data, while the layer colored blue is the hidden layer. The layer colored green is the output layer, which delivers the final output. 6

1.4 Restricted Boltzmann Machines (RBMs) are a class of neural networks with two layers - the visible layer and the hidden layer. The units (nodes) across layers are connected, but no two units in the same layer are connected, which puts the restriction in an RBM. 7

2.1 Illustration of Optimal Transport. The goal of Optimal Transport is to find a transportation plan γ that minimizes the cost of transporting mass from μ to ν. 15

3.1 An example of a single-flip in the Metropolis algorithm. Plus signs indicate spins of +1, and negative signs indicate spins of -1. 21

3.2 Mean absolute magnetization per spin $\langle |m| \rangle$ as a function of temperature for 2D Ising models of sizes 8×8, 16×16, and 32×32 using the Metropolis algorithm. 22

3.3 An example of a cluster-flip in the Wolff algorithm. Plus signs indicate spins of +1, and negative signs indicate spins of -1. 23

4.1 Spin configurations of the 64×64 Ising model at different temperatures generated by Monte Carlo simulations. Black sites correspond to spin-up ($s_i = +1$) states, and white sites correspond to spin-down ($s_i = -1$) states. 28
4.2 Specific heat c_V as a function of temperature. 29
4.3 Training loss curves for RBMs with 900 hidden units at different temperatures. The losses are plotted on a logarithmic scale against the number of epochs. 31
4.4 Probability density function (PDF) of the distribution of w_{ij} of RBMs with 900 hidden units and trained at temperatures below, near, and above T_c. 32
4.5 Five sample filters $\mathbf{w}_i^T$ at temperatures $T = 1.0, 2.25, 3.5$. The color bar range is set to be two standard deviations of the distribution. 33
4.6 Distribution of the normalized filter sums for RBMs with 900 hidden units, trained at $T = 1.0, 2.25, 3.5$. 34
4.7 Samples of original Ising configurations (top row) and RBM-generated configurations (bottom row) at $T = 2.25$. 34
4.8 Wasserstein distance as a function of temperature for RBMs with 900 hidden units. 36
4.9 KL divergence and mean energy distance as functions of temperature for RBMs with 900 hidden units. 37

Acknowledgments

First and foremost, I'd like to express my gratitude to my advisor, Prof. Weiqing Gu, for her invaluable guidance, support, and encouragement throughout my research journey.

I'm also grateful to my second reader, Prof. Daniel Tamayo, for his thoughtful feedback and suggestions, which have greatly improved the quality of this thesis.

I'd also like to extend my appreciation to Prof. Jon Jacobsen for his support and organization throughout the year.

I'd also like to acknowledge Melissa Hernandez-Alvarez, DruAnn Thomas, and Jocelyn Olds-McSpadden for their work in supporting the thesis program.

I also appreciate all the faculty members of the math department for their support and for fostering a stimulating academic environment.

A special thanks goes to my fellow Class of 2024 thesis students for their encouragement and the many intellectual discussions we have had.

Finally, I am deeply grateful to my family and friends for their unwavering love, patience, and motivation throughout my time in college.

Chapter 1

Introduction

Imagine a tiny, mythical creature sitting in front of a simple machine with a button and a clock. Every minute as the clock ticks, the creature faces a simple yet important choice, as we will see later: to press the button or not. If it presses the button, the machine outputs 1. If it doesn't press the button, the machine outputs 0. Now, if we give it a few hours, the outputs of the machine will form a sequence of 0's and 1's. In a particular case, the sequence of outputs is

$$\{0, 1, 1, 0, 0, \dots\}$$

After a few days of meticulous button-pressing, the creature arranges the outputs into a 10×10 matrix with 1's colored black and 0's left blank. It reveals this extraordinary image, which looks like a cat!

This may seem like a mundane task, but we will think a bit deeper to see that it is not. Imagine there are N pixels in the image. Before the creature starts pressing or not pressing the button, each can be either 0 or 1. This means that there are 2^N possible images. Thus, much like the legendary Maxwell's demon, this creature needs to extract meaningful information from a sea of randomness, a concept at the heart of both statistical physics and machine learning.

Statistical physics focuses on understanding how simple principles give rise to complex behaviors. Similarly, machine learning involves creating algorithms that recognize patterns from data. Advances in one of these areas often enrich the other, as they are bound together through common ideas. In this thesis, we explore the dynamics of two frameworks that lie at the intersection of statistical physics and machine learning: the Ising model and a type of neural network called the Restricted Boltzmann Machine (RBM).

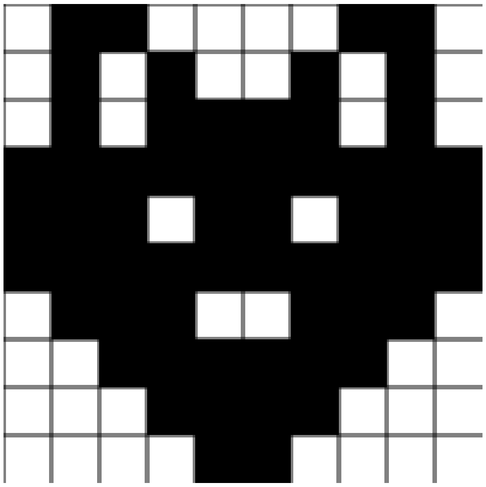

Figure 1.1 A binary image created on a 10×10 lattice.

1.1 The Ising Model

In his 1924 doctoral thesis, German physicist Ernst Ising proposed the Ising model — one of the first models in statistical physics that aimed to understand complex physical phenomena by simplifying the system as much as possible. The original purpose of the Ising model was to study phase transitions in magnetic materials. For this thesis, we will focus on the mathematical construction of the model.

The Ising model consists of "spin" variables, s, that occupy a lattice of N sites. For any given site i on the lattice where $i \in \{1, 2, \ldots, N\}$, the spin variable at that site is denoted by s_i. The Ising model allows the spins to orient in only two directions: up or down. We assign the upward spin with a value of $+1$ and the downward spin with a value of -1. Thus, the Ising spin variables are:

$$s_i = \pm 1.$$

A spin configuration $\{s_i\}$ of an Ising model is a specific assignment of $+1$ and -1 to each of the N lattice sites. For a lattice of N sites, the total number of possible spin configurations is 2^N. An example of a spin configuration is shown in Figure 1.2.

To see how the Ising model captures the magnetic behaviors of a system, we need to introduce interactions between the spin variables. These interactions will result in different spin configurations, producing different

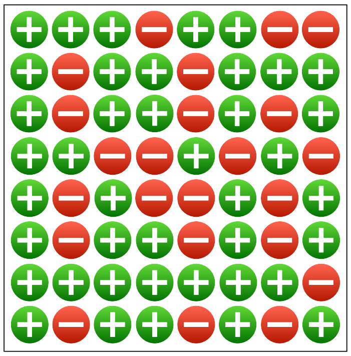

Figure 1.2 An example of an Ising model configuration. The spin variable assigned to a lattice site is either upward (indicated by green plus signs), $s_i = +1$, or downward (indicated by red minus signs), $s_i = -1$.

energies. In statistical physics, the energy of a system is quantified by the Hamiltonian. For the Ising model, $H(\{s_i\})$ is the Hamiltonian of a spin configuration $\{s_i\}$. We are mainly concerned with two kinds of interactions between the spin variables, and they each contribute to a term in $H(\{s_i\})$.

First, neighboring spins affect each other. This phenomenon is analogous to the attracting and repelling of two magnets put together. Neighboring spins will either tend to align or anti-align. For each pair of neighboring spins, s_i and s_j, we introduce an interaction strength J_{ij} that characterizes their interaction. Thus, interactions between neighboring spins introduce a term in the Hamiltonian:

$$H(\{s_i\}) = -\sum_{ij} J_{ij} s_i s_j.$$

Here, the minus sign ensures that the spin configuration that minimizes the energy is preferred.

The second type of interaction is between a spin and an external magnetic field B. The external magnetic field affects each spin independently, and each spin will try to align with the field. Summing over the spins, we obtain the Ising Hamiltonian with a second term:

$$H(\{s_i\}) = -\sum_{ij} J_{ij} s_i s_j - B \sum_i s_i.$$

1.2 Probabilities in the Ising Model

Now that we have the Hamiltonian that quantifies the energy of every assignment of +1 and −1 to the spin variables, how do we use it to describe the dynamics of a physical system? In statistical physics, the Hamiltonian is closely related to the probabilities of different states of the system.

1.2.1 The Boltzmann Factor

In the Ising model, the probability that a given spin configuration $\{s_i\}$ occurs is proportional to the exponential of $-H(\{s_i\})$. Mathematically, at temperature T

$$P(\{s_i\}) \propto e^{\frac{-H(\{s_i\})}{k_B T}}.$$

Here, $P(\{s_i\})$ is the probability of the spin configuration $\{s_i\}$. The exponential $e^{\frac{-H(\{s_i\})}{k_B T}}$ is called the Boltzmann factor, where $k_B \approx 1.38 \times 10^{-23} J \cdot K^{-1}$ is the Boltzmann constant.

To understand the physical intuition behind this relationship, we can think of the Boltzmann factor as a measurement of "likelihood." Fundamentally, nature prefers states with lower energy. For instance, we are more likely to find a ball sitting at the bottom of a slope than lying halfway on the slope. In the Ising model, a configuration with lower energy (where spins are more aligned) is more stable and thus more likely to occur than one with higher energy. Even small changes in energy can result in significant differences in likelihood. However, the temperature is a randomizing factor that affects the distribution of states. At higher temperatures, the value of $k_B T$ is larger, which means that the exponential term becomes less sensitive to the differences in energy between states. In other words, higher temperatures introduce more randomness into the system, allowing it to explore a wider range of states. Therefore, the exponential nature of the Boltzmann factor captures this relationship well.

1.2.2 The Partition Function

To convert the expression above to equality, we introduce another essential property in statistical mechanics: the partition function, Z. The partition function is the sum of the Boltzmann factors for all states of the system. For the Ising model, the partition function is

$$Z = \sum_{\{s_i\}} e^{\frac{-H(\{s_i\})}{k_B T}}.$$

The probability of the spin configuration $\{s_i\}$ is

$$P(\{s_i\}) = \frac{e^{\frac{-H(\{s_i\})}{k_B T}}}{Z}.$$

We can think of the partition function as a normalization factor, which gives us normalized probabilities.

1.3 Neural Networks

Now that we have introduced the Ising model, we will transition to introduce its counterpart in deep learning, which is a subfield of machine learning based on neural networks. We will start by getting a sense of what neural networks are.

Modeled loosely after the human brain, neural networks are algorithms designed to recognize patterns in data. The basic building blocks of neural networks are neurons. These are individual nodes in the networks that receive an input and produce an output. A neural network is simply a collection of neurons arranged into layers. Each neuron in a layer is connected to neurons in the layers before and after. The first layer that receives the data is called the input layer, while the last layer that delivers the final output is called the output layer. The layers between the input layer and the output layer are the hidden layers. Figure 1.3 is an example of a neural network with three layers.

How does a neural network learn to recognize patterns in data, such as images of dogs? It is trained on a dataset of many pictures of cats and non-cats. When these pictures are input into the neural network, they are passed from one layer to the next. In this process, each neuron assigns a weight to its input. At first, the guesses are random. However, the weights are adjusted later based on the error of previous guesses, producing more accurate predictions.

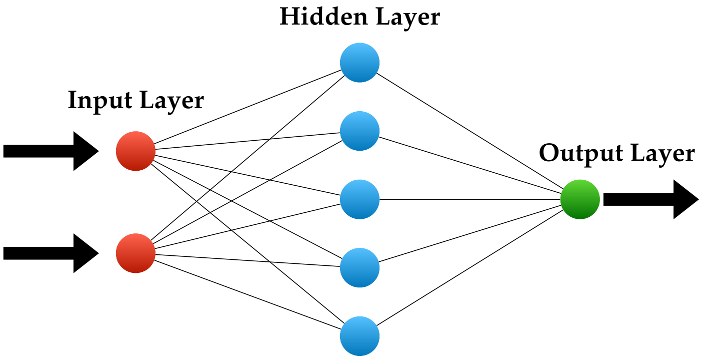

Figure 1.3 Diagram of a neural network. The layer with neurons colored red is the input layer that receives data, while the layer colored blue is the hidden layer. The layer colored green is the output layer, which delivers the final output.

1.4 Restricted Boltzmann Machines

A type of neural network that has been used extensively to study the Ising model is the Restricted Boltzmann Machine (RBM). The RBM itself is not a deep neural network but can be stacked to build deep models. As shown in Figure 1.4, an RBM consists of two layers of nodes — a visible layer and a hidden layer.

RBMs are energy-based models. This means that the state of an RBM is characterized by an energy function akin to the Hamiltonian of physical systems. We use $\{h_i\}$ to denote the hidden units and $\{v_j\}$ to denote the binary data in the visible layer. Therefore, the energy function that models the interactions between the visible and hidden layers is

$$E(\{v_j\},\{h_i\}) = -\sum_j b_j v_j - \sum_i c_i h_i - \sum_{ij} h_i W_{ij} v_j,$$

where b_j, c_i, and W_{ij} are real-valued, learnable parameters of the model.

Similar to the definition in Section 1.2.2, the probability distribution of a system with visible units $\{v_j\}$ and hidden units $\{h_i\}$ is given by

$$P(\{v_j\},\{h_i\}) = \frac{e^{-E(\{v_j\},\{h_i\})}}{Z},$$

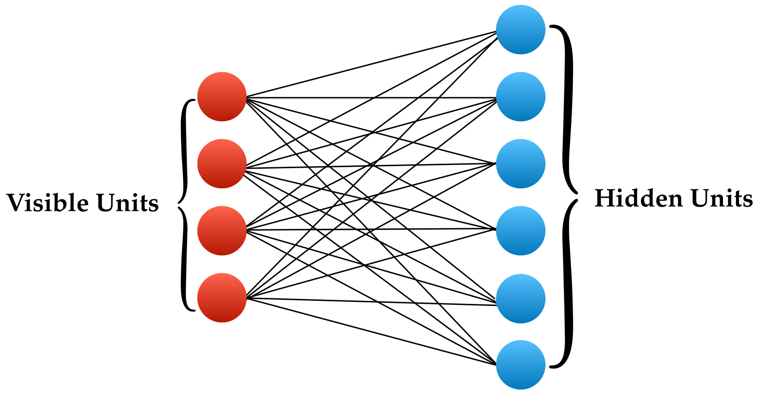

Figure 1.4 Restricted Boltzmann Machines (RBMs) are a class of neural networks with two layers - the visible layer and the hidden layer. The units (nodes) across layers are connected, but no two units in the same layer are connected, which puts the restriction in an RBM.

where the partition function of the RBM is

$$Z = \sum_{\{v_j\},\{h_i\}} e^{-E(\{v_j\},\{h_i\})},$$

analogous to the partition function.

Note that the energy-based formulation of RBMs has a striking resemblance to the Hamiltonian of the Ising model. In this analogy, the visible units of the RBM can be interpreted as the spins in the Ising model, taking on binary values that represent the state of each site in the lattice. The hidden units, on the other hand, capture the complex interactions and correlations between the visible units, akin to the coupling term in the Ising Hamiltonian. This correspondence suggests that RBMs have the potential to learn and represent the intricate statistical dependencies present in the Ising model configurations. By training RBMs on Ising model data, we aim to leverage the representational power of these energy-based models to gain insights into the collective behavior of the Ising model, opening up new avenues for the study of phase transitions and critical phenomena in statistical physics.

1.5 Overview

As summarized by Gu and Zhang (2022), it has been shown that with sufficient hidden units, RBMs can encode the Boltzmann distribution, construct the thermodynamic behaviors, and generate new configurations of an Ising model of small systems. However, the mechanism of RBMs' learning process has yet to be fully explored.

This thesis aims to numerically study the RBM learning of the Ising model of a bigger two-dimensional system. Specifically, we demonstrate the RBM's effectiveness in learning and generating configurations of the 2D Ising model at different temperatures in Chapter . In Chapter, we investigate how RBM-generated configurations capture the critical behavior and phase transitions of the Ising model. Finally, in Chapter , we use a mathematical framework called Optimal Transport as a metric to assess the quality of the RBM-generated Ising configurations.

Chapter 2

Background

As a fundamental model in statistical physics, the Ising model has been extensively studied for its ability to capture the essential features of phase transitions and critical phenomena. In this chapter, we dive into the statistical mechanics of the Ising model and the training and sampling procedures of RBMs.

2.1 Statistical Thermodynamics of the Ising Model

The Ising model captures the magnetic behaviors of systems, which is highly dependent on temperatures. In a nutshell, in such a magnetic system with up-spins and down-spins, up-spins want to be near up-spins, and down-spins want to be near down-spins. At high temperatures, the spins in the system are equally likely to be up and down. At low temperatures, the spins are either mostly up or mostly down. In this section, we will formalize the statistical mechanics of the Ising model.

2.1.1 Thermodynamic Properties of the Ising Model

As introduced above, temperature plays a crucial role in determining the system's behavior and the probability of different spin configurations of the Ising model. Recall that the energy of an Ising configuration $\{s_i\}$ is given by the Hamiltonian

$$H(\{s_i\}) = -\sum_{ij} J_{ij} s_i s_j - B \sum_{i} s_i.$$

For simplicity, we only consider the neighboring spin interactions for our 2D Ising model. Then our Ising Hamiltonian is

$$H(\{s_i\}) = -\sum_{ij} J_{ij} s_i s_j \tag{2.1}$$

where J is the interaction strength between neighboring spins. In 1.2.2, we also mentioned the probability of the spin configuration $\{s_i\}$ is

$$P(\{s_i\}) = \frac{e^{\frac{-H(\{s_i\})}{k_B T}}}{Z}$$

where Z is the partition function given by

$$Z = \sum_{\{s_i\}} e^{\frac{-H(\{s_i\})}{k_B T}}.$$

Thus, given temperature T, the mean energy of the system is

$$\langle E \rangle = \sum_{\{s_i\}} P(\{s_i\}) H(\{s_i\}) = \frac{\sum_{\{s_i\}} H(\{s_i\}) e^{\frac{-H(\{s_i\})}{k_B T}}}{Z_T}. \tag{2.2}$$

Similarly, we have

$$\langle E^2 \rangle = \sum_{\{s_i\}} P(\{s_i\}) H^2(\{s_i\}) = \frac{\sum_{\{s_i\}} H^2(\{s_i\}) e^{\frac{-H(\{s_i\})}{k_B T}}}{Z_T}. \tag{2.3}$$

For an infinitesimal increase in temperature, the increase in the mean energy is given by the specific heat c_V, which is

$$c_V = \frac{\beta^2}{N} \left(\langle E^2 \rangle - \langle E \rangle^2 \right) \tag{2.4}$$

where we defined $\beta = \frac{1}{k_B T}$. The entropy is defined by the Gibbs entropy equation:

$$S = -k_B \langle \ln(P(\{s_i\})) \rangle = -k_B \sum_{\{s_i\}} P(\{s_i\}) \ln P(\{s_i\}). \tag{2.5}$$

At low temperatures, configurations with neighboring spins aligned in the same direction are favored to minimize the system's energy. This leads to an ordered ferromagnetic phase, characterized by a non-zero magnetization, as shown by Kramers and Wannier (1941). The magnetization per spin is

$$\langle m \rangle = \frac{1}{N} \left\langle \sum_{i=1}^{N} s_i \right\rangle \tag{2.6}$$

where N is the total number of spins in the system.

On the other hand, at high temperatures, thermal fluctuations become more significant. The system is in a disordered paramagnetic phase, where spins are randomly oriented, and the magnetization vanishes. Next, we will explore this fascinating transition.

2.1.2 Phase Transitions and Critical Phenomena

In statistical physics, phase transitions occur when the system undergoes a sudden change in its macroscopic properties as a function of temperature. For the Ising model, the phase transition occurs between a ferromagnetic phase, characterized by a non-zero magnetization, and a paramagnetic phase, where the magnetization vanishes.

We see that the magnetization

$$m = \frac{1}{N} \sum_{i=1}^{N} s_i \tag{2.7}$$

is the order parameter that distinguishes the two phases. In the ferromagnetic phase, below a critical temperature T_c, the spins tend to align in the same direction, resulting in a non-zero magnetization. The system exhibits long-range order, where the spins are correlated over large distances.

As the temperature increases and approaches the critical temperature, the magnetization decreases. Finally, at the critical temperature, the system undergoes a continuous phase transition, and the magnetization vanishes. Fisher (1967) has shown that near the critical point, the magnetization follows a power-law scaling:

$$m \sim (T_c - T)^{\beta} \tag{2.8}$$

for $T < T_c$, where β is the critical exponent associated with the magnetization. This emergence of power-law behaviors of certain physical quantities is called critical phenomena.

One such quantity is the correlation length ξ, which measures the distance over which the spins are correlated. Near the critical point, the correlation length diverges as

$$\xi \sim |T - T_c|^{-\nu} \tag{2.9}$$

where ν is the critical exponent associated with the correlation length, as shown in Cardy (1996).

2.2 Training RBMs

In 1.4, we introduced the basic structure and formalism of the Restricted Boltzmann Machine (RBM). In the following section, we will delve into the architecture of an RBM in more detail, with a focus on the training process.

2.2.1 Sampling Procedures

For an RBM with n_h hidden units and n_v visible units, we can encode the two layers with state vectors. The hidden layer corresponds to the vector $\mathbf{h} = [h_1, h_2, \ldots, h_{n_h}]^T$, while the visible layer corresponds to $\mathbf{v} = [v_1, v_2, \ldots, v_{n_v}]^T$. Thus, the total energy of the RBM is given by

$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{b}^T\mathbf{v} - \mathbf{c}^T\mathbf{h} - \mathbf{h}^T\mathbf{W}\mathbf{v}. \tag{2.10}$$

Note that this is the vectorized version of the energy equation we saw in 1.4. Here, $\mathbf{b} = [b_1, b_2, \ldots, b_{n_v}]^T$ is the visible bias, and $\mathbf{c} = [c_1, c_2, \ldots, c_{n_h}]^T$ is the hidden bias. The weight matrix $\mathbf{W}$ encodes the connection between the visible and hidden units. The energy function measures the compatibility between the visible and hidden units, with lower energy configurations being more probable.

One of the key properties of RBMs is the conditional independence of the visible and hidden units given the state of the other layer, which allows for efficient inference and sampling. Since there are no direct connections between visible units in an RBM, we can easily get an unbiased sample of the state of a visible unit from a given $\mathbf{h}$:

$$P(v_i = 1|\mathbf{h}) = \sigma(b_j + \sum_{j=1}^{n_h} w_{ij}h_j) \tag{2.11}$$

where the sigmoid function is $\sigma(z) = \frac{1}{1+e^{-z}}$. Similarly, the conditional probability to generate **h** from **v** is

$$P(h_j = 1|\mathbf{v}) = \sigma(c_i + \sum_{i=1}^{n_v} w_{ij} v_i). \tag{2.12}$$

See Appendix B for detailed derivation.

2.2.2 Loss Function

The training objective of an RBM is to maximize the likelihood of the observed data under the model's probability distribution. Given a dataset $\mathcal{D} = [\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(M)}]$ consisting of M independent samples, the *loss function* is the log-likelihood of the data, given by

$$\mathcal{L}(\theta) = \frac{1}{M} \sum_{n=1}^{M} \ln P(\mathbf{v}^{(n)}; \theta) \tag{2.13}$$

where $\theta = \{W, \mathbf{a}, \mathbf{b}\}$ represents the model parameters (weights and biases) and $P(\mathbf{v}; \theta)$ is the marginal probability of the visible units, obtained by summing over all possible configurations of the hidden units:

$$P(\mathbf{v}; \theta) = \frac{1}{Z(\theta)} \sum_{\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h}; \theta)} \tag{2.14}$$

where the partition function is defined as

$$Z(\theta) = \sum_{\mathbf{v}} \sum_{\mathbf{h}} e^{-E(\mathbf{v}, \mathbf{h}; \theta)}. \tag{2.15}$$

The gradient of $\mathcal{L}(\theta)$ is given by

$$\nabla_\theta \mathcal{L}(\theta) = \langle \nabla_\theta \mathcal{E}(\mathbf{v}) \rangle_{data} - \langle \nabla_\theta \mathcal{E}(\mathbf{v}) \rangle_{model} \tag{2.16}$$

where $\mathcal{E}(\mathbf{v})$ is the effective energy for visible state $\mathbf{v}$. See Appendix B for detailed derivation of $\mathcal{E}(\mathbf{v})$.

Note computing the exact gradient is intractable due to the exponential number of terms in the partition function $Z(\theta)$, except for very small systems (see Oh et al. (2020)). Therefore, approximate training methods, such as contrastive divergence (CD), are often used to approximate the gradient.

2.2.3 Contrastive Divergence (CD)

The learning process of RBMs involves adjusting the weights and biases to minimize the difference between the true data distribution and the model's learned distribution. The most common learning algorithm for RBMs is the contrastive divergence (CD) algorithm, which approximates the gradient by replacing the expectation over the model distribution with samples obtained after a limited number of Gibbs sampling steps, starting from the data samples.

As demonstrated by Hinton (2012), the CD algorithm uses Gibbs sampling to start from the visible units, update the hidden units based on the conditional probabilities, and then reconstruct the visible units. Gibbs sampling is a Markov chain Monte Carlo (MCMC) method that samples from the joint distribution of the visible and hidden units by iteratively sampling from the conditional distributions. In the context of RBMs, Gibbs sampling alternates between sampling the hidden units given the visible units and sampling the visible units given the hidden units. This process allows the RBM to generate samples that approximate the learned distribution.

The CD algorithm starts by setting the visible units to a training example **v**. Then, it performs one or a few steps of Gibbs sampling to obtain samples from the model distribution. The sampling steps are based on the conditional probability distributions specified by Eq. 2.11 and Eq. 2.12.

After obtaining the samples from the model distribution, the weights and biases are updated based on the difference between the expectations under the data and model distributions:

$$\Delta w_{ij} = \epsilon(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}) \tag{2.17}$$

where ϵ is the learning rate, $\langle v_i h_j \rangle_{data}$ represents the expectation under the data distribution, and $\langle v_i h_j \rangle_{model}$ represents the expectation under the model distribution obtained after a few steps of Gibbs sampling.

2.3 Optimal Transport in Machine Learning

Optimal Transport (OT) is a mathematical framework for comparing and manipulating probability distributions. As summarized by Peyré and Cuturi (2020), OT has been applied in various fields, including machine learning, computer vision, and signal processing. In this section, we will introduce the concept of Optimal Transport and its applications in machine learning, focusing on the Wasserstein distance and its properties.

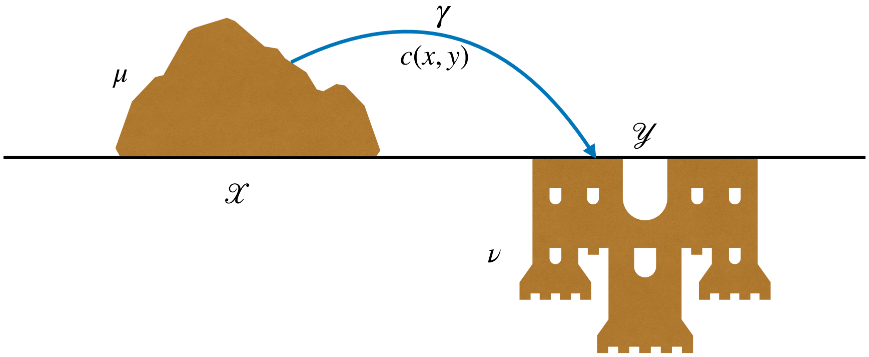

Figure 2.1 Illustration of Optimal Transport. The goal of Optimal Transport is to find a transportation plan γ that minimizes the cost of transporting mass from μ to ν.

2.3.1 Optimal Transport (OT)

Imagine we have a pile of sand. Our task is to transform the pile of sand into a different shape, such as a castle, by moving the sand in the most efficient way possible, as illustrated by Figure 2.1. The efficiency is measured by the total amount of work required to move the sand, which depends on the amount of sand being moved and the distance it needs to travel.

Now, let's think of the piles of sand as probability distributions. The goal of Optimal Transport is to find the most efficient way to redistribute the probability from one distribution to another, minimizing the total cost of transportation. More formally, given two probability measures μ and ν defined on measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, the goal of Optimal Transport is to find a transportation plan γ that minimizes the cost of transporting mass from μ to ν. The cost of transportation is defined by a cost function $c : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}_+$, where $c(x, y)$ represents the cost of transporting a unit of mass from x to y.

The set of all transportation plans between μ and ν is denoted as $\Pi(\mu, \nu)$, which consists of all joint probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν. Formally, $\Pi(\mu, \nu)$ is defined as:

$$\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) | \pi_{1\#}\gamma = \mu, \pi_{2\#}\gamma = \nu\} \tag{2.18}$$

where $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denotes the set of all probability measures on $\mathcal{X} \times \mathcal{Y}$, and $\pi_{1\#}\gamma$ and $\pi_{2\#}\gamma$ are the marginal measures of γ obtained by the projection maps $\pi_1 : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ and $\pi_2 : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$, respectively.

The Optimal Transport problem can be formulated as:

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y), d\gamma(x, y) \tag{2.19}$$

where the infimum is taken over all transportation plans γ in $\Pi(\mu, \nu)$, and the objective is to find the plan that minimizes the total transportation cost. See Santambrogio (2015) for detailed derivations.

2.3.2 Wasserstein Distance

When the cost function c is a metric on the space $\mathcal{X} = \mathcal{Y}$, the Optimal Transport problem gives rise to the Wasserstein distance (also known as the Earth Mover's distance by Rubner et al. (2000)). For two probability measures μ and ν on a metric space $(\mathcal{X}, d)$, the Wasserstein distance between μ and ν is defined as:

$$W(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y), d\gamma(x, y). \tag{2.20}$$

The Wasserstein distance quantifies the minimum cost of transforming one probability distribution into another, taking into account the underlying metric structure of the space. Ambrosio et al. (2005) have shown that the Wasserstein distance is sensitive to the geometry of the underlying space, capturing both the similarity in mass and the distance between the supports of the distributions.

2.3.3 Applications in Machine Learning

Optimal Transport (OT) and the Wasserstein distance have numerous applications in machine learning, offering a powerful framework for comparing and manipulating probability distributions. These tools have been particularly useful in unsupervised learning tasks, where the goal is to learn meaningful representations and capture the underlying structure of the data without explicit labels.

In the specific context of studying the Ising model with machine learning techniques, OT and the Wasserstein distance can be particularly useful. Specifically, OT and the Wasserstein distance can serve as valuable tools for assessing the quality of the learned representations. By comparing the Wasserstein distance between the generated samples from the trained model and the true Ising model configurations, we can evaluate how well the RBM captures the statistical properties and critical behavior of the Ising model.

2.4 Related Literature

Analytical results of the mean energy, specific heat capacity, and magnetization for the 2D Ising model have been derived by Kramers and Wannier (1941) and Yang (1952), as summarized in Appendix A. The study of phase transitions and critical phenomena in the Ising model has been a major focus of statistical physics research. Exact solutions for the one-dimensional and two-dimensional Ising models, as in Onsager (1944), have provided valuable insights into the nature of phase transitions.

The energy-based formulation of RBMs shares similarities with the Hamiltonian of the Ising model. Mehta and Schwab (2014) have found that visible units can be seen as analogous to the spins in the Ising model, while the hidden units capture the higher-order interactions and correlations between the visible units. This connection has motivated the use of RBMs to study the Ising model and its phase transitions, such as the work of Morningstar and Melko (2017), as the RBM can learn to represent the statistical dependencies present in the Ising model configurations.

The applications of OT and the Wasserstein distance extend beyond the study of the Ising model. As shown by Kolouri et al. (2017), these tools have been successfully applied to a wide range of machine learning problems, including domain adaptation, transfer learning, and distributionally robust optimization.

Chapter 3

Methods

This chapter presents the methodology used by Restricted Boltzmann Machines (RBMs) to study the Ising model and its critical behavior. We begin by describing the process of generating 2D Ising model datasets using Monte Carlo simulations, specifying the dataset parameters and specifications. Next, we discuss the training of the RBM, including its architecture, hyperparameters, and the optimization procedure used. We also outline the process of generating new configurations from the trained RBM, which will be crucial for evaluating the model's performance.

To assess the quality of the generated configurations and the RBM's ability to capture the Ising model's critical behavior, we use the Wasserstein distance as a metric based on Optimal Transport (OT). This metric provides insights into the RBM's performance across different temperature ranges. We also discuss the comparison of the Wasserstein distance with other evaluation metrics to validate our findings. By following this methodology, we aim to systematically investigate the effectiveness of RBMs in modeling the Ising model and gain new insights into its critical behavior.

3.1 Generating 2D Ising Datasets

To train the RBM and evaluate its performance in modeling the Ising model, we first need to generate a dataset of 2D Ising model configurations at various temperatures. In this section, we discuss the process of generating these datasets using Monte Carlo simulations, namely the Metropolis algorithm and the Wolff algorithm, which is known for its efficiency in sampling near the critical temperature. By carefully designing the dataset generation process, we ensure that the RBM is trained on a representative set of Ising model configurations, allowing it to learn the underlying statistical properties and critical behavior of the model.

3.1.1 Metropolis Algorithm

To generate 2D Ising model configurations, we employ Monte Carlo simulations, which are widely used in statistical physics to sample from complex probability distributions (see Newman and Barkema (1999) for details). For temperatures far from the critical point, we use the local Metropolis algorithm. The Metropolis algorithm is a Monte Carlo method for sampling from a probability distribution, and it is particularly useful when the distribution is known up to a normalization constant, as is the case with the Boltzmann distribution in the Ising model.

The Metropolis algorithm works as follows:

1. Start with an initial Ising configuration $\{s_i\}$
2. Randomly pick a site i and attempt to flip the spin at that site by $s_i \longrightarrow -s_i$
3. Calculate the change in energy ΔE caused by the proposed flip: $\Delta E = 2Js_i \sum_{j\in \mathrm{nn}(i)} s_j$ where $\mathrm{nn}(i)$ denotes the nearest neighbors of site i
4. Accept the flip with probability: $P_{\text{accept}} = \min\{1, e^{-\beta\Delta E}\}$ and update the configuration if accepted
5. Repeat steps 2-4 to reach equilibrium and obtain independent samples

See Figure 3.1 for a schematic of a local single flip in the Ising model.

The Metropolis algorithm successfully captures the phase transition between a paramagnetic phase and a ferromagnetic phase. For a 2D Ising model, Onsager (1944) first solved that this transition occurs at

$$T_c = \frac{2J}{k_B \ln(1+\sqrt{2})} \approx 2.269 J/k_B$$

where J is the coupling constant between neighboring spins. We plot the mean absolute magnetization as a function of temperature for 2D Ising models of sizes 8 × 8, 16 × 16, and 32 × 32 in Figure 3.2. We observe that the magnetization of the system drops near the critical temperature and approaches 0 as temperature increases.

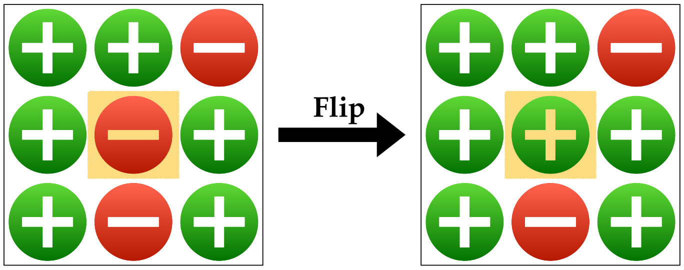

Figure 3.1 An example of a single-flip in the Metropolis algorithm. Plus signs indicate spins of $+1$, and negative signs indicate spins of -1.

Near the critical temperature, the Metropolis algorithm becomes inefficient. This is because of the significant changes in the total magnetization as the system transitions from high to low temperatures. At high temperatures, the distribution is sharply concentrated around zero magnetization. At low temperatures, the distribution exhibits a distinct double-peaked structure. However, during the transition between these two regimes, the system passes through an intermediate state where the magnetization probability distribution is nearly uniform across almost all possible magnetization values. This flat distribution poses a significant challenge for single spin-flip algorithms such as the Metropolis algorithm, making it extremely difficult to sample the configuration space effectively. For a more quantitative analysis of this point, see Krauth (2006).

3.1.2 The Wolff Algorithm

Near the critical temperature, we have noticed that the Metropolis algorithm becomes inefficient due to the presence of long-range correlations and the phenomenon of critical slowing down. To overcome this issue, we use the Wolff algorithm proposed by Wolff (1989), a cluster-flipping method that updates large clusters of correlated spins simultaneously, leading to faster convergence.

The Wolff algorithm introduces bond variables $b_{ij} \in \{0, 1\}$ between

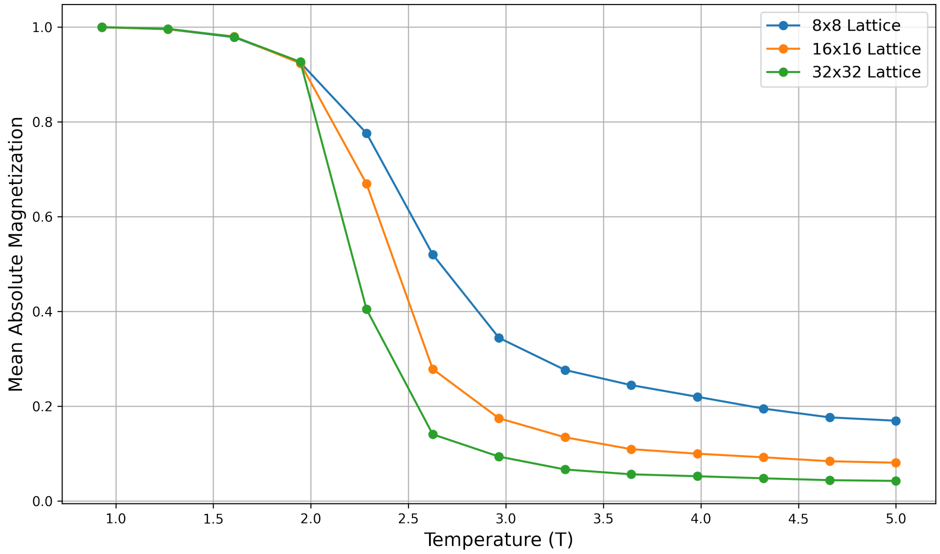

Figure 3.2 Mean absolute magnetization per spin $\langle|m|\rangle$ as a function of temperature for 2D Ising models of sizes 8×8, 16×16, and 32×32 using the Metropolis algorithm.

neighboring spins. The joint probability distribution of spins and bonds is given by:

$$P(\{s_i\}, \{b\}) = \frac{1}{Z} \prod_{\langle i,j \rangle} \left[p \delta_{b_{ij},1} \delta_{s_i,s_j} + (1-p) \delta_{b_{ij},0} \right] \tag{3.1}$$

where $p = 1 - e^{-2\beta J}$ is the probability of a bond being present and δ is the Kronecker delta function.

The Wolff algorithm grows clusters of spins connected by bonds and flips them simultaneously. It works as follows:

1. Choose a random initial spin as the starting point for the cluster
2. Create two lists: one for the cluster sites and another for the pocket sites. Add the initial spin to both lists
3. Define a probability p, which determines the likelihood of adding neighboring spins with the same orientation to the cluster
4. Begin the cluster construction process by selecting a pocket site from the pocket list and removing it from the pocket

5. For each neighboring spin with the same orientation as the pocket site, add it to the cluster and the pocket with probability p

6. If a neighboring spin is added to the cluster, also add it to the pocket list

7. Repeat steps 4-6 until the pocket list is empty

8. Once the cluster construction is complete, flip all the spins within the cluster simultaneously

9. Repeat steps 1-7 to generate a new configuration

See Figure 3.3 for a schematic of a cluster flip in the Ising model.

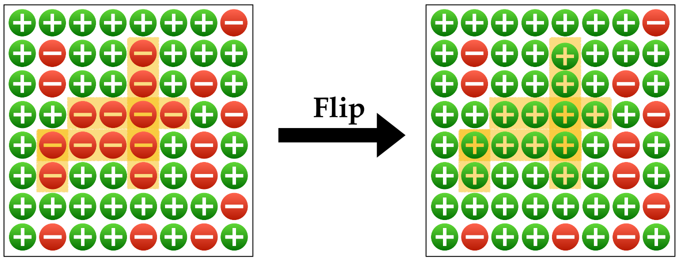

Figure 3.3 An example of a cluster-flip in the Wolff algorithm. Plus signs indicate spins of $+1$, and negative signs indicate spins of -1.

By following these steps, the Wolff algorithm efficiently generates new configurations of the system by flipping clusters of spins instead of individual spins. The approach is particularly efficient near the critical point because the average cluster size becomes comparable to the system size, allowing for large-scale updates that effectively decorrelate the configurations.

3.1.3 Dataset Specifications and Parameters

We simulate the 2D Ising model on a square lattice with periodic boundary conditions. The lattice size is chosen to be 64×64. In units of J/k_B, we choose a temperature range from $T = 0.25J/k_B$ to $T = 4.0J/k_B$, with a step size of $\Delta T = 0.25J/k_B$. The temperature range for the simulations is selected to cover both the high-temperature paramagnetic phase and the low-temperature ferromagnetic phase, as well as the critical region around the phase transition.

For each temperature, we generate an ensemble of $M = 10000$ independent configurations, using the Wolff algorithm for near-critical temperatures $T = 2.0, 2.25, 2.5$ and the Metropolis algorithm for other temperatures. We use 5000 Monte Carlo steps for equilibration and 10000 steps for data collection.

The generated dataset consists of binary spin configurations, where each spin is represented by a value of +1 or −1. The configurations are stored in a tensor of shape $(10000, 64, 64)$ for each temperature.

3.2 Training Specifications of RBM

In this section, we focus on training the Restricted Boltzmann Machine (RBM) using the dataset of 2D Ising model configurations. We begin by discussing the architecture of the RBM, including the number of visible and hidden units and the choice of activation functions. Next, we specify the hyperparameters used for training, such as the learning rate, batch size, and the number of training epochs. Finally, we outline the process of generating new configurations from the trained RBM, which will be essential for evaluating the model's performance in capturing the critical behavior of the Ising model.

3.2.1 RBM Architecture and Hyperparameters

In the context of modeling the 2D Ising model, the visible layer represents the spin configurations, while the hidden layer captures the underlying features and correlations present in the data.

For our Ising model on a 64 × 64 square lattice, the visible layer of the RBM consists of $N_v = 64^2$ units, each corresponding to a single spin. The visible units are binary, taking values of +1 or -1, consistent with the Ising spin variables. The number of hidden units, N_h, is a hyperparameter that can be adjusted to control the RBM's representational capacity. In this work, we choose $N_h = 900$. The RBM code is adapted from Gu and Zhang (2022).

The choice of hyperparameters can significantly impact the RBM's training process and its ability to learn the underlying distribution. In this work, we use the following hyperparameters:

- Learning rate: The learning rate determines the step size at each iteration. We employ the adaptive learning rate scheme Adam, developed by Kingma and Ba (2017), with an initial learning rate of 0.0001. This allows for faster convergence and automatic adjustment of the learning rate based on the gradient statistics.

- Batch size: The batch size determines the number of training examples used in each iteration of the optimization process. We use a batch size of 128.

- Number of training epochs: An epoch is a complete pass through the entire training dataset. We train the RBM for 200 epochs, which is sufficient for the model to learn the essential features of the Ising model configurations, according to Gu and Zhang (2022).

- Weight initialization: The initial values of the weights w_{ij} are drawn from a Gaussian distribution with zero mean and a standard deviation of 0.1.

3.2.2 Generating New Configurations

Once the RBM is trained on the Ising model dataset, it can be used as a generative model to produce new configurations that follow the learned probability distribution. This is achieved by exploiting the RBM's ability to capture the joint distribution of visible and hidden units and by performing Gibbs sampling to generate new configurations.

To generate new configurations, we start by randomly initializing the visible layer with binary values of -1 and 1. Next, we perform Gibbs sampling by iteratively updating the visible and hidden layers based on the learned weights and biases of the RBM. We first update the hidden layer given the current state of the visible layer and then update the visible layer given the new state of the hidden layer. The updated visible layer obtained after the Gibbs sampling steps represents a new configuration generated by the RBM.

The generated configurations will be used to assess the RBM's ability to capture the essential features and statistical properties of the 2D Ising model in the next chapter.

Chapter 4

Results

In this chapter, we present and analyze the results of training RBMs on 2D Ising model data generated by Monte Carlo algorithms. We start by examining our Monte Carlo simulation results. We do so by visualizing the Ising configurations at different temperatures and studying the presence of critical behavior. Next, we present the RBM training results, focusing on the model's learned filters. Finally, we assess the RBM's ability to generate new Ising configurations using the Wasserstein distance as a metric for similarity. Through these analyses, we can gain insights into the effectiveness of RBMs in capturing the critical behavior of the Ising model.

4.1 Ising Model Simulations

In this section, we present and analyze the results obtained from the Monte Carlo simulations of the 2D Ising model using the methods discussed in Section 3.1. These simulations serve as the ground truth for evaluating the performance of the RBM in capturing the critical behavior of the system. We begin by visualizing the spin configurations generated at different temperatures, focusing on the emergence of long-range correlations and the transition from the paramagnetic to the ferromagnetic phase. We then investigate the critical temperature and the phase transition behavior by examining various thermodynamic quantities as a function of temperature. The simulation results are compared with the exact solution and finite-size scaling theory to validate the accuracy of our approach and establish a reliable reference for assessing the RBM's performance.

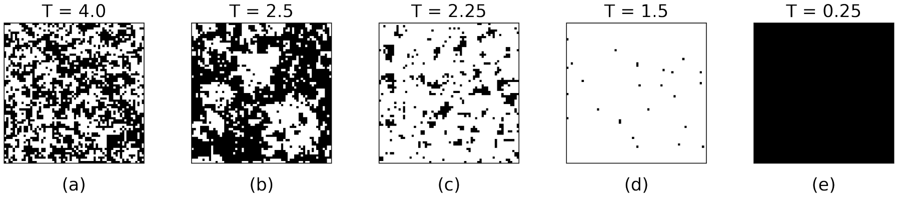

Figure 4.1 Spin configurations of the 64 × 64 Ising model at different temperatures generated by Monte Carlo simulations. Black sites correspond to spin-up ($s_i = +1$) states, and white sites correspond to spin-down ($s_i = -1$) states.

4.1.1 Data Visualization

As discussed in Section 2.1, the thermodynamic behavior of the Ising model heavily depends on the temperature. To gain insights into the system's behavior, we start by visualizing the spin configurations generated by the Monte Carlo simulations at various temperatures.

Figure 4.1 shows a series of representative Ising spin configurations for our 64 × 64 lattice at different temperatures. At high temperatures ($T \gg T_c$), the system is in the paramagnetic phase, characterized by disordered spin orientations (Figure 4.1(a)). The spins fluctuate randomly, and there is no discernible pattern or structure in the configuration. As the temperature decreases towards the critical temperature, the system undergoes a phase transition, and long-range correlations begin to appear (Figure 4.1(b)). As discussed in Cardy (1996), we can see that clusters of aligned spins start to form, and the configuration displays a mixture of ordered and disordered regions.

At the critical temperature ($T \approx T_c$), the system is scale-invariant, with spin clusters of various sizes coexisting (Figure 4.1(c)). Below the critical temperature ($T < T_c$), the system enters the ferromagnetic phase, where long-range order dominates (Figure 4.1(d)). The spins align in one of two possible orientations, forming large domains of either positive or negative magnetization. The size of these domains increases as the temperature decreases further, eventually leading to a fully ordered state at very low temperatures (Figure 4.1(e)).

These visualizations of the spin configurations confirm the effectiveness of our Monte Carlo simulations and thus can serve as a qualitative reference for assessing the RBM's ability to generate Ising configurations that resemble the true Ising model at different temperatures.

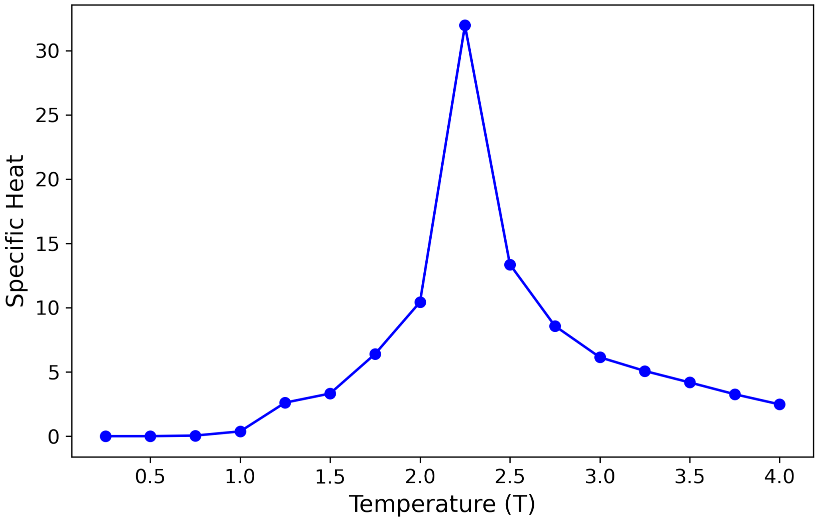

Figure 4.2 Specific heat c_V as a function of temperature.

4.1.2 Critical Behavior

A key signature of the Ising model is the phase transition at the critical temperature T_c, where the system switches from a low-temperature ferromagnetic phase to a high-temperature paramagnetic phase. We will investigate the critical temperature and the phase transition behavior by comparing the Monte Carlo simulation results with the exact solution of the 2D Ising model.

In Section 3.1.1, we have calculated the mean magnetization of the generated configurations as a function of temperature, as shown in Figure 3.2. The magnetization exhibits a sharp transition from non-zero values (ferromagnetic phase) to zero (paramagnetic phase) as the temperature increases. The transition becomes increasingly sharp as the lattice size increases, signifying the presence of a phase transition in the thermodynamic limit, as discussed by Newman and Barkema (1999).

Another way to witness the phase transition is by analyzing the temperature dependence of the specific heat (c_V), as we defined in Eq. 2.4, which measures the system's response to temperature changes. In the vicinity of the critical point ($T_c \approx 2.269J/k_B$), the specific heat exhibits a peak, as shown in Figure 4.2.

The analysis of the critical temperature and the phase transition behavior in the Monte Carlo simulations provides a solid foundation for evaluating the RBM's performance. We are now ready to compare the RBM-generated configurations and their statistical properties with the Monte Carlo results.

4.2 RBM Training Results

In this section, we analyze the results of training the Restricted Boltzmann Machine (RBM) on the 2D Ising model dataset. We focus on two key aspects of the training process: the convergence of the training loss and the learned representations of the RBMs. First, we investigate the training loss and discuss the convergence behavior. Second, we explore the learned weights and biases of the RBMs and interpret their physical significance in relation to the Ising model. The training results lay the foundation for evaluating the RBM's performance in generating new configurations in the next chapter.

4.2.1 Training Loss Convergence

During the training process, the RBM adjusts its weights and biases to minimize the difference between the true data distribution (the Ising configurations) and the distribution it learns. This difference is quantified by the training loss, which we have defined by Eq.2.13. Monitoring the training loss as a function of the number of epochs, we can assess the convergence behavior of the RBM and its ability to learn the underlying structure of the Ising model.

Figure 4.3 shows the training loss curves for RBMs with 900 hidden units. The losses are plotted on a logarithmic scale with respect to the first 20 epochs. As the training progresses, the loss decreases, indicating that the RBM is learning to better represent the data distribution.

For temperatures far from the critical point, the training loss converges quickly. This rapid convergence suggests that the RBM can easily learn the patterns in the Ising configurations at these temperatures [2]. However, near the critical temperature (e.g., $T = 2.0, 2.25$), the training loss converges more slowly, indicating that the RBM faces challenges in capturing the complex patterns near the critical point, as shown by Mehta and Schwab (2014). This is due to the complexity of the data distribution near the critical temperature, which requires the RBM to learn more intricate representations and demands more iterations.

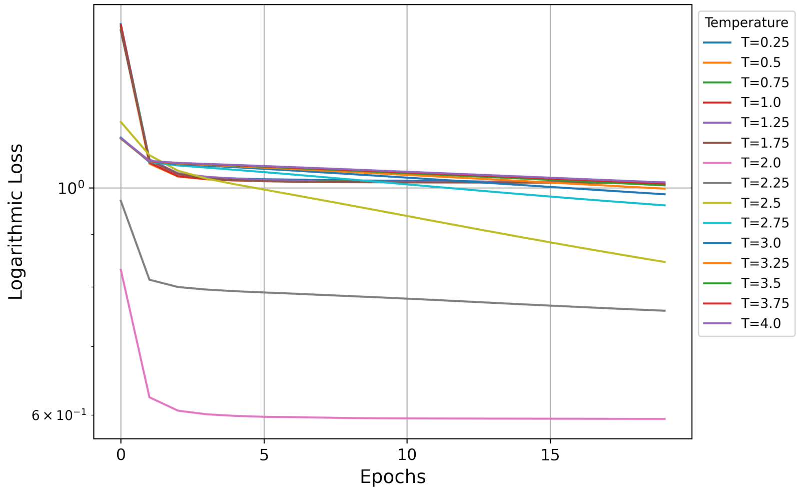

Figure 4.3 Training loss curves for RBMs with 900 hidden units at different temperatures. The losses are plotted on a logarithmic scale against the number of epochs.

4.2.2 Weights and Filters

Figure 4.4 visualizes the trained weight matrix elements w_{ij} of the RBMs with 900 hidden units trained below, near, and above the critical temperature T_c. We can observe that the PDFs follow the Gaussian distribution of zero mean, with the distribution for $T = 2.25$ having the largest variance. The distributions here are consistent with the results of Gu and Zhang (2022), but we do not observe the uniform distribution at low temperatures as shown in Torlai and Melko (2016).

For our training dataset of Ising configurations, the visible state vectors $\mathbf{v}$ are the Ising spin vectors $\mathbf{s}$. Therefore, the weights $\mathbf{w}_i^T$ connecting the visible units to the hidden units can be interpreted as *filters* that capture specific patterns or features in the Ising configurations. Each hidden unit learns to respond to a particular pattern in the visible units, and the weights determine the strength and nature of these connections. If we reshape $\mathbf{w}_i^T$ to match the spatial structure of the Ising lattice, we should expect patterns similar to the representative spin configurations at different temperatures.

In Figure 4.5, we show five sample filters for temperatures below, around, and above the critical temperatures. Each row of the matrix corresponds

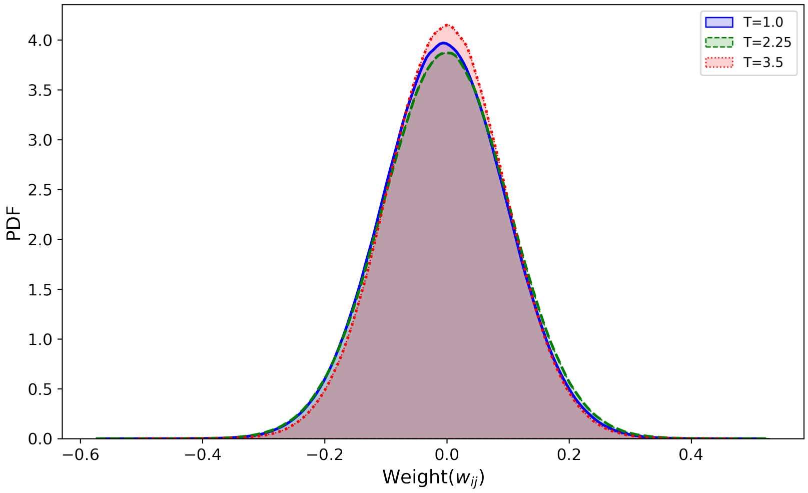

Figure 4.4 Probability density function (PDF) of the distribution of w_{ij} of RBMs with 900 hidden units and trained at temperatures below, near, and above T_c.

to a hidden unit, and each column corresponds to a visible unit. The color scale represents the strength and sign of the weights, with red indicating positive values and blue indicating negative values.

Below the critical temperature ($T = 1.0$), the Ising model is in the ferromagnetic phase, characterized by large domains of aligned spins. The components of $\mathbf{w}_i^T$ tend to be mostly positive or negative. Close to the critical temperature ($T = 2.25$), the Ising model exhibits critical behavior, with the emergence of scale-invariant correlations. Here, the patterns of $\mathbf{w}_i^T$ fluctuate, corresponding to the critical fluctuations of the system. Above the critical temperature ($T = 3.5$), the Ising model is in the paramagnetic phase, characterized by disordered spin configurations. The filters at this temperature have roughly equal numbers of well-mixed positive and negative values, agreeing with Gu and Zhang (2022).

The *filter sum* of the RBM, $\sum_{j=1}^{n_v} w_{ij}$, serves a similar role to the magnetization m in the Ising model. In Figure 4.6, we plot the PDFs of the normalized filter sums, $\sum_{j=1}^{n_v} w_{ij}/n_v$, of the RBMs trained at $T = 1.0, 2.25, 3.5$. As the temperature increases, the distribution changes from bimodal to unimodal. The bimodal distribution at low temperatures ($T < T_c$) corresponds to the two possible magnetization states in the ferromagnetic phase. Near the

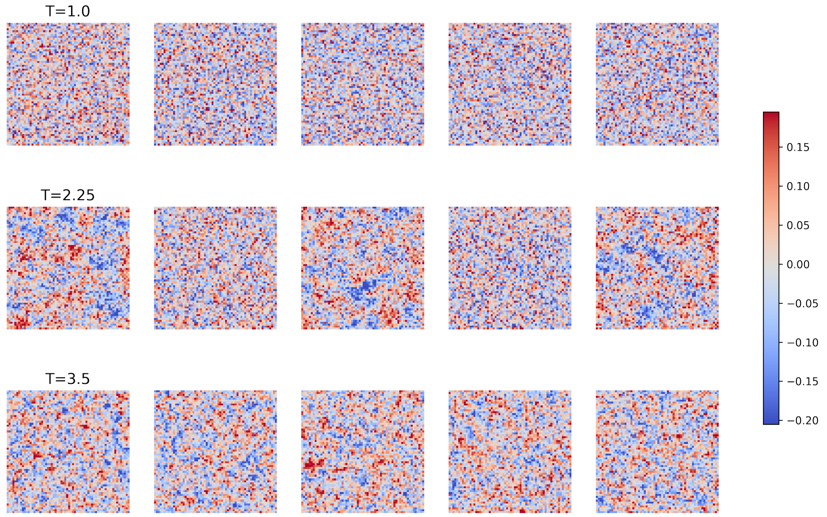

Figure 4.5 Five sample filters $\mathbf{w}_i^T$ at temperatures $T = 1.0, 2.25, 3.5$. The color bar range is set to be two standard deviations of the distribution.

critical temperature ($T \approx T_c$), the bimodal distribution of the filter sums becomes less pronounced, and the peaks start to merge, reflecting the increased fluctuations and the diminishing long-range order in the phase transition. At high temperatures ($T > T_c$), the unimodal distribution reflects the random spin orientations in the model with no preferred magnetization direction. We expect the distribution to be centered at zero. However, the shifted center of the PDF at $T = 1.0$ may be caused by uneven numbers of $m > 0$ and $m < 0$ configurations in our dataset.

4.3 RBM-Generated Ising Configurations

New data generation is an essential feature of neural networks. Using our trained RBMs, we generate new Ising configurations that mimic the statistical properties of the original Ising dataset that we generated using Monte Carlo simulations. To generate new configurations, we first initialize the RBM with random initial spin states. Then, we perform five Gibbs

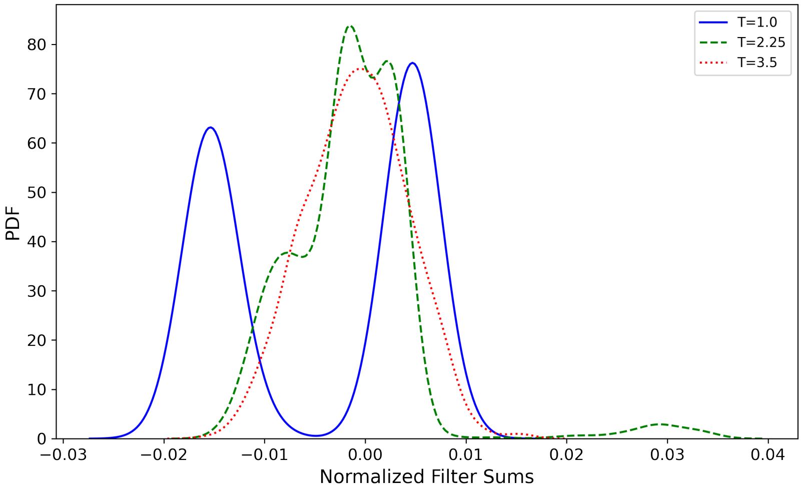

Figure 4.6 Distribution of the normalized filter sums for RBMs with 900 hidden units, trained at $T = 1.0, 2.25, 3.5$.

sampling steps to update the visible units based on the probabilistic influence of hidden units and vice versa. The Gibbs sampling procedure allows the RBM to walk through its representation of the configuration space, effectively "thermalizing" the learned distribution. Samples of the original and generated configurations near the critical temperature are shown in Figure 4.7. In this section, we evaluate the RBM's ability to capture the thermodynamic behavior of the Ising model.

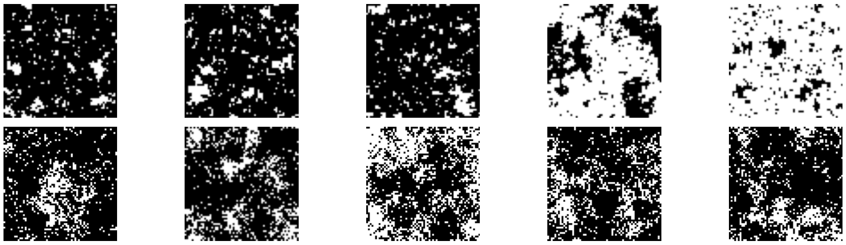

Figure 4.7 Samples of original Ising configurations (top row) and RBM-generated configurations (bottom row) at $T = 2.25$.

4.3.1 Wasserstein Distance

As introduced in Section 2.3.2, the Wasserstein distance is an effective metric for similarity between two probability distributions. In this section, we analyze how the Wasserstein distance varies with temperature.

Figure 4.8 shows the Wasserstein distance as a function of temperature. The Wasserstein distance is computed between the generated configurations and the true Ising model configurations at each temperature point. We notice the Wasserstein distance exhibits a distinct behavior near the critical temperature near the critical temperature.

At temperatures far from the critical temperature, both below and above, the Wasserstein distance remains relatively high, which means the generated Ising configurations at these temperatures are less similar to the true configurations. This suggests that the RBM faces challenges in capturing the long-range order in the ferromagnetic phase ($T < T_c$) and the disordered paramagnetic phase ($T > T_c$).

However, as the temperature approaches T_c, the Wasserstein distance experiences a sharp decrease, reaching its minimum at $T = 2.0$ and increasing again at $T = 2.25$. This behavior indicates the RBM can effectively capture the critical fluctuations of the Ising model near the phase transition and generate configurations that closely resemble the true critical behavior. However, the model's performance may be less optimal at temperatures far from criticality, where the Ising model exhibits simpler patterns and shorter-range correlations.

4.3.2 Other Evaluation Metrics

To further quantify the similarity between the original and generated Ising model configurations, we computed the Kullback-Leibler (KL) divergence and mean energy distance across different temperatures.

Like the Wasserstein distance, the KL divergence measures the difference between two probability distributions, say, P and Q. In the context of comparing the original and generated Ising model configurations, we consider P as the probability distribution of the original samples and Q as the probability distribution of the generated samples. The KL divergence is defined as:

$$\mathrm{KL}(P \parallel Q) = \sum_{x} P(x) \ln \left(\frac{P(x)}{Q(x)} \right) \tag{4.1}$$

where x represents the configuration, and the summation is taken over all

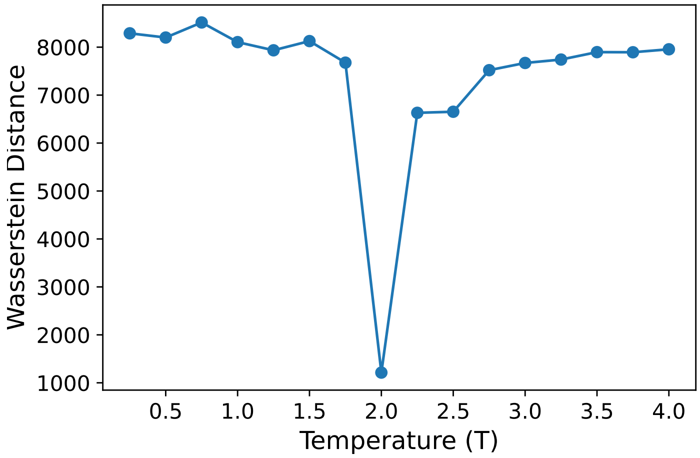

Figure 4.8 Wasserstein distance as a function of temperature for RBMs with 900 hidden units.

possible configurations. $P(x)$ and $Q(x)$ denote the probabilities of observing configuration x in the original and generated distributions, respectively.

The KL divergence quantifies the amount of information lost when using the generated distribution Q to approximate the original distribution P. A lower KL divergence indicates better agreement between the two distributions.

The mean energy distance measures the average absolute difference between the energies of the original and generated configurations. Let $E_o(x)$ and $E_g(x)$ represent the energy functions of the original and generated Ising models, respectively. The mean energy distance is defined as:

$$\mathrm{MED} = \frac{1}{M} \sum_{i=1}^{M} \left| E_o(x_i) - E_g(x_i) \right| \tag{4.2}$$

where M is the total number of configurations, and x_i denotes the i-th configuration.

The mean energy distance quantifies the average deviation between the energies of the original and generated configurations. A lower mean energy distance indicates that the generated samples more closely resemble the energy profile of the original Ising configurations.

Figure 4.9 shows the KL divergence and the mean energy distance as functions of temperature. Notably, although exhibiting a more gradual increase when the temperature exceeds T_c, these metrics both exhibit similar trends as the Wasserstein distance, with higher values at both low and high temperatures and a dip around the critical temperature. The consistent trends observed across these metrics underscore the challenges in generating accurate samples that are far from the critical point.

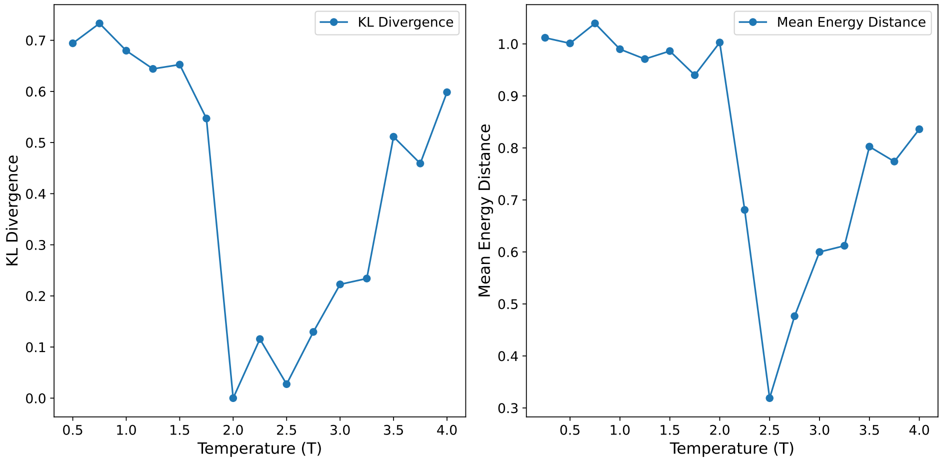

Figure 4.9 KL divergence and mean energy distance as functions of temperature for RBMs with 900 hidden units.

Chapter 5

Conclusion

In this thesis, we have demonstrated the effectiveness of Restricted Boltzmann Machines (RBMs) in capturing the critical behavior and phase transition properties of the 2D Ising model. By training RBMs on equilibrium configurations generated using Monte Carlo simulations across a range of temperatures, we have shown that these generative models can learn the patterns, long-range correlations, and scale-invariant fluctuations that emerge near the critical temperature. Our analysis of the learned weights, generated samples, and performance metrics, such as the Wasserstein distance, reveals the RBM's ability to compress the thermodynamic information and generate new configurations that resemble the true Ising model. While acknowledging the limitations of our study, our work highlights the potential of RBMs as a robust tool for studying critical phenomena and phase transitions in complex systems.

5.1 Discussion of Key Results

We trained RBMs using equilibrium 2D Ising configurations collected from Monte Carlo simulations at various temperatures. Our results have shown that RBMs can effectively capture the essential thermodynamic behavior and phase transition properties of the Ising model.

One of the key aspects of our work is the successful generation of 2D Ising model configurations using Monte Carlo simulations. Based on single-spin flips, the Metropolis algorithm is efficient for simulating the Ising model at temperatures far from the critical point. However, the Metropolis algorithm suffers from critical slowing down near the critical temperature. To overcome this limitation, we employed the Wolff algorithm, a cluster-flipping method particularly effective for simulating the Ising model near criticality. Combining these two algorithms generated accurate and representative configurations of the 2D Ising model across a wide range of temperatures.

The analysis of the learned weights and representations of the RBM (Figures 4.4 and 4.5) demonstrates the model's ability to capture the statistical properties of the Ising model. The visualization of the learned weights reveals the RBM's ability to learn the short-range interactions and critical fluctuations in the Ising model.

The temperature dependence of the Wasserstein distance, as shown in Figure 4.8, reveals that the RBM achieves the lowest dissimilarity between the generated and true configurations near the critical temperature. This trend in similarity is verified by the KL divergence and mean energy distance shown in Figure 4.9. This indicates that the RBM can effectively learn the complex patterns, long-range correlations, and scale-invariant fluctuations that emerge at criticality. However, the model's performance may be less optimal at temperatures far from criticality, which requires further investigation.

These findings demonstrate the RBMs' potential as a powerful tool for studying critical phenomena and phase transitions in complex systems, such as the Ising model. RBMs can learn the intricate correlations from the dataset and provide a compressed representation of the system's behavior, allowing for the efficient generation of new samples.

5.2 Limitations

While we have demonstrated the RBM's effectiveness in capturing the critical behavior of the 2D Ising model, it is important to acknowledge the limitations of this work.

First, we saw a sharp dip in the Wasserstein distance near the critical temperature, as shown in Figure 4.8. This result suggests the RBM can capture the complex patterns near criticality more effectively than the simpler patterns far from the critical temperature. This rather counterintuitive trend in the Wasserstein distance suggests further investigations into the data generation process using the trained RBMs.

The choice of hyperparameters and architecture of the RBM can impact the results and performance as well. In our study, we mostly analyzed RBMs with 900 hidden units and selected hyperparameters based on trial and error and literature recommendations, such as Hinton (2002). However, the optimal hyperparameters and architecture may vary depending on the specific problem and dataset. More rigorous approaches, such as cross-validation or Bayesian optimization, could be employed to find the optimal hyperparameters and architecture, as discussed by Bergstra and Bengio (2012). Additionally, exploring other architectures, such as deep Restricted Boltzmann machines, may provide further insights and improvements in capturing the critical behavior of the Ising model (see Mehta et al. (2019) for examples of training deep RBMs).

Lastly, while we have visualized the learned weights and filters of the RBMs, the exact physical meaning of the learned features and their correspondence to properties of the Ising configurations still need to be studied. As summarized by Carleo et al. (2019), it is challenging to analytically study the learning process of the RBM, and relating physically interpretable information of the learned representations and relating them to the underlying physics is an ongoing effort in the field of machine learning for physics.

In conclusion, while our work has shown promising results in applying RBMs to study the critical behavior of the 2D Ising model, the approach can be extended to further ensure the quality of the training and data-generating procedures.

5.3 Future Work

The results demonstrate the potential of RBMs in capturing the critical behavior and phase transition properties of the 2D Ising model. However, there are several open questions and potential directions for future research that can further enhance our understanding of RBMs and their application to studying critical phenomena.

One of the most intriguing findings in this work is the unusual trend in the Wasserstein distance between the generated and true Ising configurations across different temperatures (Figure 4.8). The sharp dip in the Wasserstein distance near the critical temperature suggests that the RBM can more effectively capture the complex patterns and long-range correlations that occur at criticality compared to the simpler patterns far from the critical temperature. This counterintuitive behavior demands further investigation to uncover the underlying mechanisms. Future work could involve a more detailed analysis of the learned weights and representations of the RBM at different temperatures. Understanding the factors that contribute to this unusual trend could provide valuable insights into the RBM's learning process.

Another important direction for future work is to explore the features learned by the hidden layer of the RBMs and how they relate to the physical properties of the Ising model. While we have visualized the learned weights and filters (Figures 4.4 and 4.5), future work could involve developing methods to visualize and interpret the activations of the hidden units, such as using dimensionality reduction techniques to project the hidden representations onto a lower-dimensional space. Additionally, analyzing the correlations between the hidden unit activations and physical observables, such as the magnetization or energy, could provide insights into how the RBM encodes the thermodynamic properties of the system.

Lastly, the quality of the generated Ising configurations can potentially be improved by increasing the number of hidden units in the model. In our study, we primarily focused on RBMs with 900 hidden units, which demonstrated promising results. By systematically varying the number of hidden units and analyzing the resulting generated samples, we can investigate the trade-offs between model complexity and performance.

In conclusion, our work has laid the foundation for several exciting directions for future research in applying RBMs to study critical phenomena. By investigating the unusual trend in the Wasserstein distance, exploring the learned features of the hidden layer, and optimizing the model's learning capacity, we can deepen our understanding of how RBMs learn and capture the essential properties of complex systems. These efforts will not only advance our knowledge of RBMs as a tool for studying critical phenomena but also contribute to the highly active field of machine learning for physics, where interpretability, generalization, and physical insights are of paramount importance.

Appendix A

Thermodynamics of 2D Ising Model

In this appendix, we show the analytical results of the mean energy, specific heat capacity, and magnetization for the 2D Ising model derived by Kramers and Wannier (1941) and Yang (1952).

The 2D Ising model has been solved exactly by Onsager (1944). The critical temperature is given by:

$$k_B T_c = \frac{2J}{\ln(1+\sqrt{2})}. \tag{A.1}$$

The magnetization per spin below the critical temperature is given by:

$$\langle m \rangle = \left(1 - \sinh^{-4}(\frac{2J}{k_B T})\right)^{1/8}. \tag{A.2}$$

The specific heat capacity near the critical temperature behaves as:

$$C \sim -\ln|T - T_c|. \tag{A.3}$$

Appendix B

Energy and Probability of RBMs

In this appendix, we present the derivations of the energy and probability of RBMs, which can be found in Murphy (2012).

Starting from the definition of the free energy:

$$\mathcal{E}_\theta(\mathbf{v}) = -\ln \sum_{\mathbf{h}} e^{-\mathcal{E}_\theta(\mathbf{v},\mathbf{h})} = -\ln p_\theta(\mathbf{v}) - \ln Z_\theta \tag{B.1}$$

Substituting the effective energy of the RBM:

$$\begin{aligned}
\mathcal{E}_\theta(\mathbf{v}) &= -\ln \sum_{h_1=0}^{1} \sum_{h_2=0}^{1} \cdots \sum_{h_{n_h}=0}^{1} \exp\left(\sum_{j=1}^{n_v} b_j v_j + \sum_{i=1}^{n_h} c_i h_i + \sum_{j=1}^{n_v} \sum_{i=1}^{n_h} W_{ij} v_j h_i \right) \\
&= -\sum_{j=1}^{n_v} b_j v_j - \ln \prod_{i=1}^{n_h} \sum_{h_i=0}^{1} \exp\left(-\sum_{j=1}^{n_v} W_{ij} v_j - c_i + \sum_{j=1}^{n_v} W_{ij} v_j + c_i \right) \\
&= -\sum_{j=1}^{n_v} b_j v_j - \sum_{i=1}^{n_h} \ln \left(e^{-\sum_{j=1}^{n_v} W_{ij} v_j - c_i} + e^{\sum_{j=1}^{n_v} W_{ij} v_j + c_i} \right) \\
&= -\mathbf{b}^T \mathbf{v} - \sum_{i=1}^{n_h} \ln \left(e^{-\mathbf{W}_i^T \mathbf{v} - c_i} + e^{\mathbf{W}_i^T \mathbf{v} + c_i} \right).
\end{aligned}$$

The conditional probability distribution of the hidden units given the visible units can be derived as:

$$
\begin{aligned}
p_\theta(\mathbf{h}|\mathbf{v}) = \frac{p_\theta(\mathbf{v},\mathbf{h})}{p_\theta(\mathbf{v})} = \frac{e^{-\mathcal{E}_\theta(\mathbf{v},\mathbf{h})}}{e^{-\mathcal{E}_\theta(\mathbf{v})}} &= \frac{1}{\Omega_\theta(\mathbf{v})} e^{\sum_{i=1}^{n_h} c_i h_i + \sum_{i=1}^{n_h} h_i \mathbf{w} i^T \mathbf{v}} \\
&= \frac{1}{\Omega_\theta(\mathbf{v})} \prod i = 1^{n_h} h_i \left(c_i + \mathbf{w} i^T \mathbf{v} \right) = \prod i = 1^{n_h} p_\theta(h_i|\mathbf{v})
\end{aligned}
$$

where $p_\theta(h_i|\mathbf{v}) \propto e^{h_i(c_i + \mathbf{w}_i^T \mathbf{v})}$. This gives the single unit conditional probability:

$$
\begin{aligned}
p_\theta(h_i = 1|\mathbf{v}) &= \frac{p_\theta(h_i = 1|\mathbf{v})}{p_\theta(h_i = -1|\mathbf{v}) + p_\theta(h_i = 1|\mathbf{v})} \\
&= \frac{e^{c_i + \mathbf{w}_i^T \mathbf{v}}}{e^{-c_i - \mathbf{w}_i^T \mathbf{v}} + e^{c_i + \mathbf{w}_i^T \mathbf{v}}} \\
&= \frac{1}{1 + e^{-2(c_i + \mathbf{w}_i^T \mathbf{v})}} \\
&= \sigma(2(c_i + \mathbf{w}_i^T \mathbf{v})).
\end{aligned}
$$

Similar relations can be derived for $p_\theta(h_i = -1|\mathbf{v})$, $p_\theta(v_j = 1|\mathbf{h})$ and $p_\theta(v_j = -1|\mathbf{h})$.

Bibliography

Ambrosio, Luigi, Nicola Gigli, and Giuseppe Savaré. 2005. *The Wasserstein Distance and its Behaviour along Geodesics*, 151–165. Basel: Birkhäuser Basel. doi:10.1007/3-7643-7309-1_9. URL https://doi.org/10.1007/3-7643-7309-1_9.

Bergstra, James, and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *J Mach Learn Res* 13(null):281–305.

Brush, Stephen G. 1967. History of the lenz-ising model. *Rev Mod Phys* 39:883–893. doi:10.1103/RevModPhys.39.883. URL https://link.aps.org/doi/10.1103/RevModPhys.39.883.

Cardy, John. 1996. *Phase transitions in simple Systems*, 1–15. Cambridge Lecture Notes in Physics, Cambridge University Press.

Carleo, Giuseppe, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. 2019. Machine learning and the physical sciences. *Reviews of Modern Physics* 91(4). doi:10.1103/revmodphys.91.045002. URL http://dx.doi.org/10.1103/RevModPhys.91.045002.

Fisher, M E. 1967. The theory of equilibrium critical phenomena. *Reports on Progress in Physics* 30(2):615. doi:10.1088/0034-4885/30/2/306. URL https://dx.doi.org/10.1088/0034-4885/30/2/306.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Gu, Jing, and Kai Zhang. 2022. Thermodynamics of the ising model encoded in restricted boltzmann machines. *Entropy* 24(12):1701. doi:10.3390/e24121701. URL http://dx.doi.org/10.3390/e24121701.

Hinton, Geoffrey E. 2002. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation* 14(8):1771–1800. doi:10.1162/089976602760128018. URL https://doi.org/10.1162/089976602760128018. _eprint: https://direct.mit.edu/neco/article-pdf/14/8/1771/815447/089976602760128018.pdf.

———. 2012. A practical guide to training restricted boltzmann machines. In *Neural Networks*. URL https://api.semanticscholar.org/CorpusID:21145246.

Kingma, Diederik P., and Jimmy Ba. 2017. Adam: A method for stochastic optimization. 1412.6980.

Kolouri, Soheil, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K. Rohde. 2017. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine* 34(4):43–59. doi:10.1109/MSP.2017.2695801.

Kramers, H. A., and G. H. Wannier. 1941. Statistics of the two-dimensional ferromagnet. part i. *Phys Rev* 60:252–262. doi:10.1103/PhysRev.60.252. URL https://link.aps.org/doi/10.1103/PhysRev.60.252.

Krauth, Werner. 2006. *Statistical Mechanics Algorithms and Computations*. Oxford: Oxford University Press.

Mehta, Pankaj, Marin Bukov, Ching-Hao Wang, Alexandre G.R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. 2019. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports* 810:1–124. doi:10.1016/j.physrep.2019.03.001. URL http://dx.doi.org/10.1016/j.physrep.2019.03.001.

Mehta, Pankaj, and David J. Schwab. 2014. An exact mapping between the variational renormalization group and deep learning. 1410.3831.

Morningstar, Alan, and Roger G. Melko. 2017. Deep learning the ising model near criticality. 1708.04622.

Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Newman, M E J, and G T Barkema. 1999. The Ising model and the Metropolis algorithm. In *Monte Carlo Methods in Statistical Physics*. Oxford University Press. doi:10.1093/oso/9780198517962.003.0003. URL https://doi.org/10.

1093/oso/9780198517962.003.0003. https://academic.oup.com/book/0/chapter/422704903/chapter-pdf/52593551/isbn-9780198517962-book-part-3.pdf.

Oh, Sangchul, Abdelkader Baggag, and Hyunchul Nha. 2020. Entropy, free energy, and work of restricted boltzmann machines. *Entropy* 22(5). doi:10.3390/e22050538. URL https://www.mdpi.com/1099-4300/22/5/538.

Onsager, Lars. 1944. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Phys Rev* 65:117–149. doi:10.1103/PhysRev.65.117. URL https://link.aps.org/doi/10.1103/PhysRev.65.117.

Peyré, Gabriel, and Marco Cuturi. 2020. Computational optimal transport. 1803.00567.

Rubner, Yossi, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* 40(2):99–121.

Santambrogio, Filippo. 2015. *Primal and dual problems*, 1–57. Cham: Springer International Publishing. doi:10.1007/978-3-319-20828-2_1. URL https://doi.org/10.1007/978-3-319-20828-2_1.

Torlai, Giacomo, and Roger G. Melko. 2016. Learning thermodynamics with boltzmann machines. *Phys Rev B* 94:165,134. doi:10.1103/PhysRevB.94.165134. URL https://link.aps.org/doi/10.1103/PhysRevB.94.165134.

Wolff, Ulli. 1989. Collective monte carlo updating for spin systems. *Phys Rev Lett* 62:361–364. doi:10.1103/PhysRevLett.62.361. URL https://link.aps.org/doi/10.1103/PhysRevLett.62.361.

Yang, C. N. 1952. The spontaneous magnetization of a two-dimensional ising model. *Phys Rev* 85:808–816. doi:10.1103/PhysRev.85.808. URL https://link.aps.org/doi/10.1103/PhysRev.85.808.