DEEP LEARNING AND CRITICAL BEHAVIOUR OF THE ISING MODEL

Under the guidance of
Dr.Girish Setlur

Uddhav Sen

Submitted in partial fulfillment of the requirements
for the degree of
Master of Science
in
Physics

Department of Physics
Indian Institute of Technology-Guwahati
India

Abstract

The thesis begins with a mathematical overview of the interconnections between many body criticality and stable information flow in the mean field approximation. We also see how the principle of maximum entropy can be used to connect the notions of ignorance and statistical entropy and use it to derive the mean field criticality theory of the 4D Ising model in detail. This correlation leads to a couple of interesting observations: Deep Neural Networks(DNN) can simulate many body systems and predict macroscopical and emergent phenomena more efficiently in the critical regime. Deep learning models can also be used to classify and predict critical point behaviour of many body systems without any a priori knowledge about the underlying physics. Furthermore, designing DNNs at the critical point may provide us with better training performance and convergence criteria. Finally, we show simulations and plots of trainability and validation accuracies supporting our theoretical predictions for various DNN architectures.

Contents

1 Introduction

Many-body physics and neural networks are two fields that might seem unrelated at first glance, but they share some fundamental similarities and have been found to be interrelated in various ways. Many-body physics is the study of systems composed of a large number of interacting particles or objects, such as atoms, molecules, or subatomic particles. These systems exhibit complex and often chaotic behavior, making it challenging to predict their properties and behavior accurately. Neural networks, on the other hand, are computational models inspired by the structure and function of the human brain. They are designed to learn patterns and relationships from data and make predictions based on that learning.

One area where they intersect is in the development of machine learning algorithms that can accurately simulate the behavior of complex physical systems. Neural networks can be trained on large datasets of simulated or experimental data to learn the underlying physical laws governing the behavior of a system[1]. This approach has been used to simulate many-body quantum systems, which are notoriously difficult to model using traditional computational methods[2][3]. By using neural networks, researchers have been able to make more accurate and efficient predictions of the behavior of quantum systems.

Another area of intersection is in the development of new simulation methods. Many-body physics simulations can be computationally expensive and time-consuming, but neural networks can be used to speed up the simulation process by approximating the behavior of the system. This has led to the development of new simulation methods, such as neural network quantum Monte Carlo, which uses neural networks to speed up quantum Monte Carlo simulations. One such example is the mean field approach, where the interactions between particles are approximated using a neural network that is trained on a large dataset of simulated or experimental data. The neural network learns to approximate the mean field effect of all the other particles on a given particle, and this approximation can be used to speed up the simulation process by avoiding the costly computation of pairwise interactions between all the particles.

Finally, many-body physics and neural networks also intersect in the study of emergent phenomena. Many-body systems can exhibit emergent behavior, where the behavior of the system as a whole is greater than the sum of its individual parts. Neural networks can be used to study emergent behavior by learning from data and identifying patterns in the behavior of the system. For example, the Ising model is a many-body physics system that is used to study phase transitions and critical phenomena in condensed matter physics. It consists of a lattice of spins that interact with their nearest neighbors, and the behavior of the system depends on the temperature and external magnetic field.

For example, a Convolutional Neural Network(CNN) can be trained on large datasets of simulated or experimental data to identify patterns and relationships between the spins in the lattice and the properties of the system, such as its energy and magnetization. By analyzing these patterns, the CNN can gain insights into the emergent behavior of the Ising model, such as the emergence of magnetic domains or the critical behavior near the phase transition. This information can then be used to predict the behavior of the Ising model in different scenarios, such as in the presence of an external magnetic field or at different temperatures. Neural networks have been used in many recent studies to study the Ising model and related many-body physics systems, such as the XY model and the Heisenberg model. These studies have demonstrated the effectiveness of neural networks in identifying emergent behavior and predicting the behavior of complex systems.

The structure of the report is as follows. We start by showing how the Ising model under a mean field approximation leads to a breakdown at the critical point using the principle of maximum entropy and field theoretical techniques. Then, we show how this breakdown corresponds to the idea of stable information flow at the critical point. DNNs built at this point account for correlations at all scales and hence show improved trainability. This order-to-chaos transition property has been studied in the next section. The next section gives a brief overview of how deep learning techniques have been used to learn about critical phases of many body systems including the 4D-Ising model.

2 Mean Field Theory: Introduction

The mean field theory (MFT) is a widely used approximation method in many-body physics that simplifies the calculations by reducing the interactions between individual particles to an effective interaction with a mean field. The idea behind MFT is to replace each degree of freedom in the system, along with its interactions with other degrees of freedom, with an effective degree of freedom that interacts with an average field representing the effect of all other degrees of freedom in the system. The MFT is based on the observation that calculating the partition functions of many-body systems with explicit interactions is often extremely challenging or impossible. However, by approximating the interactions with a mean field, the partition function can be simplified and made tractable.

MFT breaks down precisely at the critical point and it becomes important to understand how it occurs and when is it valid. We consider the d-dimensional Ising Hamiltonian with N spins isotropically (i.e., $\frac{N}{d}$ spins per direction):

$$\hat{\mathcal{H}} = -\frac{1}{2}\sum_{i,j}\hat{J}_{ij}s_i s_j - \sum_i \hat{h}_i s_i \tag{1}$$

where $i = 1, 2, ..., N$ and $s_i = \pm 1$. In the case of a one-dimensional lattice, boundary effects can be eliminated by connecting the ends of the lattice, for example, by setting $s_{N+1} = s_1$. However, in the present discussion, we focus on the thermodynamic limit where $N \to \infty$. In this limit, we can consider the effect of all other spins on a given spin s_i as an external magnetic field. This is because we can rewrite the Ising Hamiltonian (1) as:

$$\hat{\mathcal{H}} = \sum_i s_i \left(-\frac{1}{2} \sum_j \hat{J}_{ij} s_j - \hat{h} \right) \tag{2}$$

We then replace s_j by the average value, $\langle s_j \rangle \equiv s$ and since no spin is special, the most likely value of s_j is mean. This is also known as the isotropy argument. This allows us to define an effective magnetic field at site i:

$$\hat{h}_i^{\text{eff}} \equiv \frac{s}{2} \sum_j \hat{J}_{ij} + \hat{h} \tag{3}$$

So that the Hamiltonian becomes;

$$\hat{\mathcal{H}} \approx -\sum_i s_i \hat{h}_i^{\text{eff}} \tag{4}$$

and we assume that the correlation between spins at different sites is negligible and replace s_j with its mean value $\langle s_j \rangle = s$. This neglects terms of order δs^2, which is valid in the limit of large system sizes or in the thermodynamic limit. This is equivalent to neglecting the fluctuation effects, as the fluctuations would be small in the thermodynamic limit. The resulting Hamiltonian is then diagonal, and we can solve it exactly by finding the eigenvalues and eigenvectors of the matrix.

$$\begin{aligned} s_i &= (s + (s_i - s)) \\ s_j &= (s + (s_j - s)) \end{aligned} \tag{5}$$

Putting (5) into the first term of (1);

$$\begin{aligned} s_i s_j &= (s + (s_i - s))(s + (s_j - s)) \\ &= (s + \delta s_i)(s + \delta s_j) \\ &= s^2 + s(\delta s_i + \delta s_j) + O(\delta s^2) \\ &= -s^2 + s(s_i + s_j) + O(\delta s^2) \end{aligned} \tag{6}$$

We insert (6) to get the mean-field Hamiltonian;

$$\begin{aligned}\hat{\mathcal{H}}_{MF} &= -\frac{1}{2}\sum_{i,j}(-s^2 + s\,(s_i + s_j) + \delta s_i \delta s_j) - \sum_i \hat{h}_i s_i \\ &= \frac{s^2}{2}\sum_{j,i}\hat{J}_{ij} - \frac{s}{2}\sum (s_i + s_j)\,\hat{J}_{ij} - \sum_i \hat{h}_i s_i + O(\delta s^2) \\ &= \frac{s^2}{2}\sum_{j,i}\hat{J}_{ij} + \sum_i s_i \left(-\sum_j \hat{J}_{ij} s - \hat{h}\right) + O(\delta s^2) \end{aligned} \tag{7}$$

Symmetric pairwise interactions have been assumed, where $\hat{J}ij = \hat{J}ji$. It is important to note that the first term in this assumption is proportional to the size of the lattice and does not contribute to the dynamics of the system. By defining an effective action in equation (3) and incorporating a factor of 2, we were able to recover the effective one-body Hamiltonian in equation (4) by working to linear order in the fluctuations. Therefore, although we were able to average over the fluctuations in our mean field approximation, we also lost some valuable information regarding the interactions between the spins.

3 The Principle of Maximum Entropy: A Constrained Optimisation Scheme

The marriage between information theory and mean field theory might look forced in the beginning but it becomes extremely obvious once we realise the similarities between ignorance and statistical counting. In this section, we explore the connection between the subjective and objective concepts of entropy, as demonstrated by Jaynes[4]. The information-theoretic notion of entropy reflects our lack of knowledge, while the thermodynamic notion of entropy measures the statistical microstates of a system. As Shannon noted[5], these two concepts are essentially identical and can be expressed as;

$$S = -\sum_i p_i \ln p_i \tag{8}$$

where p_i is the probability of the i^{th} outcome/microstate. These concepts share the same mathematical form which leads to an intuitive sense that there must be some deeper relationship lurking beneath the surface.

We will begin by discussing the connection between entropy and subjective ignorance. In his influential paper, Claude Shannon demonstrated that equation (8) is the only unambiguous measure of the "amount of uncertainty" conveyed by a discrete probability distribution. This proof established a formal link between information theory and physics. To illustrate, we consider a random variable X, which can take on any of the discrete values $(x_1, \ldots, x_n)$ representing specific outcomes or measurement values. Each value x_i is associated with

a corresponding probability $p_i \in (p_1, \ldots, p_n)$, such that $\sum_i p_i = 1$. Our goal is to find a function $H(p_1, \ldots, p_n)$ that accurately measures the amount of uncertainty represented by this probability distribution. We constrain H to satisfy the following three conditions:

- H is well defined; i.e. it is a continuous function of p_i.
- If all p_i are equal, then $p_i = 1/n$ and $A(n) \equiv H(1/n, \ldots, 1/n)$ follows $A'(n) > 0$.
- If we break down an event into its sub-events, the initial value of H must be the weighted sum of the values of H for each individual sub-event.

The second property ensures that entropy increases as system size increases and the third property includes Bayesian logic in the system. Then it becomes convenient to represent the probabilities in simpler form; $p_i = \frac{n_i}{\sum n_i}, n_i \in \mathbb{N}$. This enables us to fix the form of H by applying the composition law to the case in which we coarse-grain n equally likely alternatives into clusters of size n_i, which results in the requirement

$$A\left(\sum\nolimits_i n_i\right) = H(p_1, \ldots, p_n) + \sum\nolimits_i p_i A(n_i).$$

Finally we consider the equally likely outcome where $n_i = m$, in which case $\sum_i n_i = nm$, and we get the functional equation $A(nm) = A(n) + A(m)$. This has a well-known solution which is $A(n) = K \ln n$. Here K is positive by the second condition. This leads to;

$$\begin{aligned} K \ln \sum_i n_i &= H(p_1, \ldots, p_n) + K \sum_i p_i \ln n_i \\ &\implies H(p_1, \ldots, p_n) = -K \sum_i p_i \ln p_i \end{aligned} \tag{9}$$

(9) reproduces the form of (8) and we see how they are connected formally using the same functional representation(for statistical systems, we have $S = \ln \Omega$). The principle of maximum entropy or max-ent estimation is rooted in this observation. This statistical inference method aims to assign probabilities and make inferences based on incomplete information. We strive for our decisions to be as impartial as possible, so we must employ the probability distribution that maximizes entropy while meeting known constraints. This distribution is the most neutral with regard to unknown information. Using any other distribution would introduce arbitrary constraints and lead to biased decision-making. As such, max-ent estimation may be viewed as a measure of rationality, as it represents the most accurate guess we can make based on the available information.

With the connection firmly in hand, we introduce a constrained optimisation algorithm to determine the best-suited probability distribution function.

We start with a variable X which assumes discrete values x_i with associated probabilities p_i. The expectation value of some function f(x) becomes;

$$\langle f(x) \rangle = \sum_{i=1}^{n} p_i f(x_i) \tag{10}$$

Given this information, the question is: what is the expected value of a function $g(x)$? However, the problem is currently unsolvable as we need knowledge of p_i to calculate $\langle g(x) \rangle$. The second constraint, $\sum_i p_i = 1$, is straightforward and established by (10). The principle of max-ent enables us to solve this problem and ensures that our probability distribution is as inclusive as possible, avoiding the concentration of probability density more narrowly than the provided information allows. Mathematically, we aim to maximise entropy provided the given constraints using the Lagrange multiplier method. The objective function is given as;

$$\mathcal{L}(p_i; \alpha, \beta) = -\sum_i p_i \ln p_i + \alpha \left(\sum_i p_i - 1 \right) + \beta \left(\sum_i p_i f(x_i) - \langle f(x) \rangle \right) \tag{11}$$

Taking the gradients to zero, we get;

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_j} &= -\sum_i \delta_{ij} \ln p_i - \sum_i p_i \partial_j \ln p_i + \alpha \sum_i \delta_{ij} + \beta \sum_i \delta_{ij} f(x_i) = 0 \\ &\implies -\ln p_j - 1 + \alpha + \beta f(x_j) = 0 \\ &\implies \ln p_j = -\alpha - \beta f(x_j) \\ &\implies p_j = e^{-\alpha - \beta f(x_j)} \end{aligned} \tag{12}$$

We have redefined $\alpha \to \alpha - 1$ and see that we get the Maxwell Boltzmann distribution such that $e^{\alpha} = \sum_i e^{-\beta f(x_i)}$ and $\langle f(x) \rangle = \sum_i f(x_i) e^{-\alpha - \beta f(x_i)}$. The first represents the well-known partition function and the second represents the definition of ensemble expectation of a statistical system. This essentially completes the connection with the thermodynamic concept of entropy. If the only known information about a system is its average energy, the Boltzmann distribution can be used to describe the microstates of the system. This reflects the lack of any other physical constraints and is an objective feature of the system. Hence, both the subjective and objective notions of entropy are related to the idea that the correct probability distribution should be as dispersed as possible, given the constraints or knowledge about the system. In fact, it is possible to define thermodynamic entropy as the quantity that is maximized at thermal equilibrium.

We treat the proof as an algorithm to our previously described mean-field model to find critical points and breakdown. The criticality leads to a diverging correlation length and consequent stable information propagation schemes to implement deep neural nets. A quantum extension of the proof also exists where the p.d.f. needs to be replaced by the density matrix.

4 Mean Field Theory: Critical Point Breakdown

We draw from our connections in information theory and state that if we have no a priori knowledge about the probability distribution function with respect to which a particular expectation value is computed, the most unbiased choice is obtained by maximizing the entropy. In particular, if we know the expected energy, $\langle H \rangle = E$, the algorithm yields the Boltzmann distribution, $p_i \equiv p(x_i) = \dfrac{1}{Z[\beta]} e^{-\beta E_i}$ where $E_i \equiv \langle H(x_i) \rangle$, and we correlate the inverse temperature β as the Lagrange multiplier arising from the constraint on the energy. Now we also know that the free energy $F = E - TS = E - \beta^{-1} S$, we can safely say that maximising the entropy is equivalent to minimising the free energy of the system. This leads to an interesting conclusion. The max-ent principle yields the lower bound on the free energy of the system. The formal statement is known as Bogolyubov inequality which states that the mean field free energy is always greater than or equal to its actual free energy.

The mean-field partition function for (7) is;

$$\begin{aligned} Z_{\text{MF}} &= \sum_{\{s\}} e^{-\beta H_{\text{MF}}} = \prod_{k=1}^{N} \sum_{s_k = \pm 1} \exp\left[-NdJs^2 + (2dJs + h) \sum_i s_i \right] \\ &= e^{-NdJs^2} \prod_{k=1}^{N} \exp\left[-(2dJs + h) + (2dJs + h) \right] \\ &= 2^N e^{-NdJs^2} \cosh^N (2dJs + h) \end{aligned} \tag{13}$$

The partition function is constrained to account for only nearest-neighbor interactions with uniform coupling in a d-dimensional Ising model consisting of N spins. Each spin interacts with $2d$ neighboring spins, and the interaction strength is represented by $J \equiv \beta \hat{J}$, while the strength of the external magnetic field is represented by $h \equiv \beta \hat{h}$. The corresponding free energy per unit particle is then

$$\begin{aligned} f_{\text{MF}}(s) &= -\frac{1}{N\beta} \ln Z \\ &= \beta^{-1} dJs^2 - \beta^{-1} \ln \cosh (2dJs + h) \end{aligned} \tag{14}$$

where we have dropped the $\beta^{-1} \ln 2$ term since it will not affect any of the observables for which f serves as a generating function. Since we already know from Bogolyubov inequality that (14) provides an upper bound on the true free energy, we can obtain the tightest possible bound by minimizing over s:

$$\frac{\partial f_{\text{MF}}}{\partial s} = 0 \implies \quad s = \tanh (2dJs + h) \tag{15}$$

The condition is commonly known as self-consistent. One way to solve this equation is by asymptotic analysis; by finding the intersection points between the left-hand side and right-hand side of the equation. In summary, when $h \neq 0$, the global minimum of f_{MF} occurs at a positive s value, which is independent of the temperature. On the other hand, when $h = 0$, there is only one minimum at $s = 0$ for high temperatures, while two degenerate minima at $s = \pm s_0$ exist for low temperatures, depending on whether $\tanh\Big(2d\beta\hat{J}s\Big)$ intersects with s for $s \neq 0$. When x is small, the Taylor expansion of $\tanh(x)$ yields $\tanh(x) \approx x - x^3/3 + O(x^5)$. This implies that for sufficiently small β, the graph of $\tanh\Big(2d\beta\hat{J}s\Big)$ is approximately a straight line whose slope is less than s. The critical temperature that separates these two regimes is determined by imposing a condition;

$$s = \tanh\Big(2d\beta_c\hat{J}s\Big) = \tanh(s) \quad \Longrightarrow \quad T_c = 2d\hat{J}$$

.

For $d = 1, T_c = 0$, while for $d = 2, T_c \approx 2.269\hat{J}$; as I showed in the first part of the thesis submitted before. We focus on the $h = 0$ case: the critical point s_0 will always be small ($|s_0| < 1$) and independent of T (since $\lim_{x\to\pm\infty} \tanh x = \pm 1$),

$$\ln\cosh(2dJs_0) = \frac{1}{2}(2dJ)^2 s_0^2 - \frac{1}{12}(2dJ)^4 s_0^4 + O(s_0^6) \tag{16}$$

whence the free energy (14) near the critical point is approximately;

$$f_{\mathrm{MF}}(s_0) \approx \frac{r}{2}s_0^2 + \frac{g}{4!}s_0^4 \tag{17}$$

where we have defined

$$r \equiv \frac{2dJ}{\beta}(1 - 2dJ) = \frac{T_c}{T}(T - T_c) \tag{18}$$

$$g \equiv \frac{32d^4J^4}{\beta} = \frac{2T_c^4}{T^3} \tag{19}$$

At the critical temperature T_c, the value of r undergoes a sign change. This determines whether the global minima of f_{MF} occurs at $s_0 = 0$ ($T > T_c$) or $\pm s_0 > 0$ ($T < T_c$), indicating the presence or absence of a non-zero magnetization, respectively. This transition in magnetization is an example of an order parameter, where below the critical temperature, the spins prefer to align, while above it, thermal fluctuations prevent such alignment resulting in zero net magnetization.

The mean field theory (MFT) is limited in its ability to accurately predict the critical point, which is precisely where it fails. This is because at the critical point, fluctuations at all scales become significant, whereas MFT only considers fluctuations to linear order. To gain a deeper understanding of this limitation, it is necessary to move beyond the discrete lattice of MFT and employ a continuum field theory. By doing so, we will discover that MFT corresponds to the dominant saddle point approximation of the quantum field theory. In order to

achieve this, we can transform our square-lattice Ising model to a scalar field theory that is equivalent. Let's look at the original partition function again;

$$\begin{aligned} Z &= \sum_{\{s\}} e^{-\beta H} \\ &= \prod_{k=1}^{N} \sum_{s_k=\pm 1} \exp\left(\frac{1}{2}\sum_{i,j} J_{ij} s_i s_j + \sum_i h_i s_i\right) \end{aligned} \tag{20}$$

where as before we have absorbed β by defining $J = \beta\hat{J}$, $h = \beta\hat{h}$. The first step is to apply the Hubbard-Stratanovich transformation,

$$e^{\frac{1}{2}\sum_{i,j} J_{ij} s_i s_j} = \left[\frac{\det J}{(2\pi)^N}\right]^{1/2} \int \mathrm{d}^N\phi \exp\left(-\frac{1}{2}\sum_{i,j} J_{ij}\phi_i\phi_j + \sum_{i,j} J_{ij} s_i \phi_j\right) \forall s_i \in \mathbb{R} \tag{21}$$

The Hubbard-Stratonovich transform is used to decouple a system of interacting quantum fields by introducing auxiliary fields that mediate the interactions between the original fields. This allows the path integral to be written as a product of simpler Gaussian integrals, which can be evaluated using standard techniques. Applying this transformation to the first term in the partition function, we have

$$Z = \left[\frac{\det J}{(2\pi)^N}\right]^{1/2} \int \mathrm{d}^N\phi \sum_{\{s\}} \exp\left[-\frac{1}{2}\sum_{i,j} J_{ij}\phi_i\phi_j + \sum_i \left(\sum_j J_{ij}\phi_j + h_i\right) s_i\right] \tag{22}$$

We can now sum over the s which makes (22) in terms of the new field variables ϕ. We calculate for each spin separately as,

$$\sum_{s_i=\pm 1} \exp\left[\sum_i \left(\sum_j J_{ij}\phi_j + h_i\right) s_i\right] = 2\cosh\left(\sum_j J_{ij}\phi_j + h_i\right) \tag{23}$$

The partition function then becomes,

$$Z = \left[\left(\frac{2}{\pi}\right)^N \det J\right]^{1/2} \int \mathrm{d}^N\phi \exp\left[-\frac{1}{2}\sum_{i,j} J_{ij}\phi_i\phi_j + \sum_i \ln\cosh\left(\sum_j J_{ij}\phi_j + h_i\right)\right] \tag{24}$$

We define $\sum_j J_{ij}\phi_j + h_i \equiv \mu^i$ as the mean field $\langle\phi_i\rangle$ at site i, incorporating the interaction with all other sites as well as the external magnetic field. This

helps in expressing (24) in terms of μ_i as follows:

$$\begin{aligned}\phi_i &= J_{ij}^{-1}\left(\mu_j - h_j\right)\\ &\implies \sum_{i,j} J_{ij}\phi_i\phi_j = J_{ij}J_{in}^{-1}J_{jm}^{-1}\left(\mu_n - h_n\right)\left(\mu_m - h_m\right)\\ &= J_{ij}^{-1}\left(\mu_i\mu_j - h_i\mu_j - h_j\mu_i + h_ih_j\right) \end{aligned} \tag{25}$$

We have utilised the anti-symmetry of the wedge product ($A \wedge B = -B \wedge A$), together with the fact that $J_{ii} = 0$. We get a new measure in terms of the continuum mean field,

$$\begin{aligned}\mathrm{d}\phi_i &= J_{ij}^{-1}\mathrm{d}\mu^j\\ &\implies \mathrm{d}^N\phi = \det J^{-1}\mathrm{d}^N\mu \end{aligned} \tag{26}$$

Hence the partition function may be equivalently expressed as;

$$\begin{aligned}Z = &\left[\left(\frac{2}{\pi}\right)^N \det J^{-1}\right]^{1/2} e^{-\frac{1}{2}\sum_{i,j} J_{ij}^{-1}h_ih_j}\cdot\\ &\int \mathrm{d}^N\mu \exp\left[-\frac{1}{2}\sum_{i,j} J_{ij}^{-1}\mu_i\mu_j + \sum_{i,j} J_{ij}^{-1}h_i\mu_j + \sum_i \ln\cosh\mu_i\right] \end{aligned} \tag{27}$$

where we assume $J_{ij} = J_{ji}$ as ever. Now we want to obtain a more tractable expression, let us examine the scenario in which the external magnetic field is extremely small, as we did previously. In this situation, since the spin interactions do not establish any favored direction, we anticipate that the mean field will be close to zero, i.e., $|\mu_i| \ll 1$. The Taylor series corresponds to;

$$\ln\cosh\mu_i = \frac{1}{2}\mu_i^2 - \frac{1}{12}\mu_i^4 + O(\mu_i^6)$$

and the approximate partition function is;

$$\begin{aligned}Z \approx &\left[\left(\frac{2}{\pi}\right)^N \det J^{-1}\right]^{1/2} e^{-\frac{1}{2}\sum_{i,j} J_{ij}^{-1}h_ih_j}\\ &\cdot\int \mathrm{d}^N\mu \exp\left[-\sum_{i,j}\frac{1}{2}J_{ij}^{-1}\mu_i\mu_j + \sum_{i,j} J_{ij}^{-1}h_i\mu_j + \sum_i\left(\frac{1}{2}\mu_i^2 - \frac{1}{12}\mu_i^4\right)\right] \end{aligned} \tag{28}$$

We move on to taking the continuum limit, where we assign the d-dimensional vector $\mathbf{x}$ as the label for the field at each site (meaning that μ_i becomes $\mu(\mathbf{x})$, and $\sum_i$ becomes $\int \mathrm{d}^dx = \int \mathrm{d}\mathbf{x}$), and we end up with the path-integral measure.

$$\left[\left(\frac{2}{\pi}\right)^N \det J^{-1}\right]^{1/2} e^{-\frac{1}{2}\sum_{i,j} J_{ij}^{-1}h_ih_j}\int \mathrm{d}^N\mu \longrightarrow \mathcal{N}\int \mathcal{D}\mu \tag{29}$$

Thus the mean field approximate continuum field theory for the Ising model is

$$Z \approx \mathcal{N} \int \mathcal{D}\mu \exp\left\{ -\frac{1}{2} \int \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}\, \mu(\mathbf{x}) J^{-1}(\mathbf{x}-\mathbf{y}) \big[\mu(\mathbf{y}) - h\big] + \frac{1}{2} \int \mathrm{d}\mathbf{x} \left[\mu(\mathbf{x})^2 - \frac{1}{6}\mu(\mathbf{x})^4 \right] \right\} \tag{30}$$

where $h(\mathbf{x}) = h$, since the external magnetic field is the same for all lattice sites. We further simplify by imposing the partition function to include only nearest-neighbor interactions. We would also like to preserve this notion of locality in the field theory. To do this, we take $|\mathbf{y} - \mathbf{x}| \ll 1$ and Taylor expand the field $\phi(\mathbf{y})$ around $\mathbf{x}$:

$$\mu(\mathbf{y}) = \mu(\mathbf{x}) + \sum_i (y_i - x_i)\partial_i \mu(\mathbf{x}) + \frac{1}{2}\sum_{i,j}(y_i - x_i)(y_j - x_j)\partial_i\partial_j \mu(\mathbf{x}) + O\left((\mathbf{y}-\mathbf{x})^3\right)$$

Expanding $J^{-1}(\mathbf{y} - \mathbf{x})$ in powers of $-\nabla_y^2$ gives rise to non-local interactions, with higher-derivative terms suppressed by increasing powers of the separation between points. Substituting this expansion into (30) yields the following expression:

$$Z \approx \mathcal{N} \int \mathcal{D}\mu\, e^{-S[\mu]} \tag{31}$$

with $S[\mu] = \int \mathrm{d}^d\mathbf{x} \left[\frac{1}{2}\kappa(\nabla\mu)^2 - h\mu + \frac{1}{2}\tilde{r}\mu^2 + \frac{\tilde{g}}{4!}\mu^4 \right]$.

The analytic functions κ, $\tilde{r}$, and $\tilde{g}$, which can be expressed using the inverse coupling matrix, are coefficients in the expansion of $J^{-1}(\mathbf{y} - \mathbf{x})$ in terms of infinitesimally separated points in space. Higher-derivative terms are suppressed by increasing powers of the separation. The key takeaway from this exercise in field theory is that MFT corresponds to the leading saddle point of (31). If we denote the minimum as μ_0 and expand the action to second order in the fluctuations $\delta\mu \equiv (\mu - \mu_0)$, we get:

$$Z \approx \mathcal{N} e^{-S[\mu_0]} \int \mathcal{D}\mu\, e^{-\frac{1}{2}(\delta\mu)^2 S''[\mu_0]} \tag{32}$$

where the prime denotes variation with respect to μ. We also note that the linear term has vanished by definition, i.e., μ_0 is given by

$$\frac{\delta S}{\delta \mu} = 0$$
$$\implies \quad \kappa\nabla^2\mu_0 = -h + \tilde{r}\mu_0 + \frac{\tilde{g}}{6}\mu_0^3 \tag{33}$$

where we have assumed that the field vanishes at infinity. If we then keep only the leading-order saddle point, the partition function simply becomes the

constant mean field independent term;

$$Z \approx \mathcal{N} e^{-S[\mu_0]} \quad \text{with} \quad S[\mu_0] = \int \mathrm{d}^d\mathbf{x} \left(\tilde{r}\mu_0^2 + \frac{\tilde{g}}{8}\mu_0^4 - \frac{3}{2}h\mu_0 \right) \tag{34}$$

and the unit free energy is;

$$f_{\text{sp}} = \frac{1}{\beta}\left(\tilde{r}\mu_0^2 + \frac{\tilde{g}}{2}\mu_0^4 - \frac{3}{4}h\mu_0 \right) = \frac{\hat{r}}{2}\hat{\mu}_0^2 + \frac{\hat{g}}{4!}\hat{\mu}_0^4 - \hat{h}\mu_0 \tag{35}$$

the subscript "sp" denotes "saddle point", and the non-dynamical term $\ln \mathcal{N}$ has been dropped. In the second equation, the factor of β has been extracted from μ by defining $\mu = \beta\hat{\mu}$, and by setting $2\hat{r} \equiv \beta\tilde{r}$, $\hat{g} \equiv 12\tilde{g}\beta^3$, and $4\hat{h} \equiv 3h$. When $h = 0$, this is formally equivalent to f_{MF} in (17), but with an additional linear term.

Moving forward, we exploit the Gaussian nature of the action and evaluate the partition function using the standard technique. This allows us to obtain a simple expression for the two-point correlator, which captures the dominant behaviour near the critical point. It is worth noting that this is relevant for understanding the propagation of information, which has connections to information theory and neural networks as mentioned earlier. Therefore, comprehending the behaviour of correlation functions in the MFT and the impact of fluctuations on them is essential. The quadratic action is given by the partition function (31).

$$S[\mu] = \int \mathrm{d}^d\mathbf{x} \left[\frac{1}{2}(\nabla\mu(\mathbf{x}))^2 + \frac{m^2}{2}\mu(\mathbf{x})^2 - h\mu(\mathbf{x}) \right]$$

where we make $\kappa = 1$ and relabel the quadratic coefficient to m^2. The current form of the action is similar to that of a free massive scalar field, where the parameter h acts as a source. To simplify the analysis, we can perform a Fourier transform to momentum space, which decouples the modes, and absorb the source-independent term into the overall normalization. The only distinction is that we are working with Euclidean signature, so there are no problems related to convergence. This approach is important for understanding how correlation functions behave in the MFT approximation and how they may be affected by fluctuations. Moreover, it is relevant to the investigation of information propagation near the critical point, which has connections to information theory and neural networks as mentioned in the introduction.

$$Z \simeq \exp \frac{1}{2}\int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{\tilde{h}_{\mathbf{k}}\tilde{h}_{-\mathbf{k}}}{k^2 + m^2} = \exp \frac{1}{2}\int \mathrm{d}^d x \mathrm{d}^d y \, h(\mathbf{x})G(\mathbf{x} - \mathbf{y})h(\mathbf{y}) \tag{36}$$

The second equality involves Fourier transforming back to real space by identifying the Green function/propagator:

$$G(\mathbf{x} - \mathbf{y}) = \int \frac{\mathrm{d}^d k}{(2\pi)^2} \frac{e^{-i\mathbf{k}\mathbf{x}}}{k^2 + m^2} \tag{37}$$

Next, we define the Green function/propagator as the correlation between the field at $\mathbf{x}$ and $\mathbf{y}$. To proceed, we introduce a length scale $\xi^2 = m^{-2}$, and make use of the identity:

$$\int_0^\infty \mathrm{d}t\, e^{-t(k^2+\xi^{-2})} = \frac{1}{k^2 + 1/\xi^2}$$

to make the integral into:

$$G(r) = \frac{1}{(4\pi)^{d/2}} \int_0^\infty \mathrm{d}t\, t^{-d/2} e^{-r^2/4t - t/\xi^2} \tag{38}$$

This is obtained by completing the square in the exponential and performing the integral over $\mathrm{d}^d k$.

Exponentiating the $t^{-d/2}$ factor, we have;

$$\frac{1}{(2\pi)^{d/2}} \int \mathrm{d}^d k\, \frac{\mathrm{e}^{-\frac{1}{2}\mathbf{k}^2 t}}{t^{d/2}} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{y}-\mathbf{x})} = \frac{1}{(2\pi)^{d/2}} \int \mathrm{d}^d k\, \mathrm{e}^{-\frac{1}{2}(\mathbf{k}-\frac{\mathbf{y}-\mathbf{x}}{t})^2 t} \left(\frac{t}{2\pi}\right)^{d/2}.$$

Then, we expand the exponential to quadratic order, which is valid when $r \gg \xi$ or $r \ll \xi$. This is because when $r \gg \xi$, the correlation length is much smaller than the distance between the two points, and hence the field fluctuates rapidly. In contrast, when $r \ll \xi$, the correlation length is much larger than the distance between the two points, and hence the field is smooth and slowly varying. Expanding the exponential to quadratic order and completing the square, we get;

$$\frac{1}{(2\pi)^{d/2}} \int \mathrm{d}^d k\, \mathrm{e}^{-\frac{1}{2}(\mathbf{k}-\frac{\mathbf{y}-\mathbf{x}}{t})^2 t} \left(\frac{t}{2\pi}\right)^{d/2} \approx \left(\frac{t}{2\pi}\right)^{d/2} \mathrm{e}^{-\frac{(\mathbf{x}-\mathbf{y})^2}{2t}} \mathrm{e}^{-\frac{t}{2}\frac{(\mathbf{x}-\mathbf{y})^2}{\xi^2}},$$

where we have used the Gaussian integral identity. Here, the first exponential term is the usual Gaussian distribution, which decays rapidly with distance, while the second exponential term captures the long-range correlations and decays as a power-law.

$$G(r) = \frac{1}{(4\pi)^{d/2}} \int_0^\infty \mathrm{d}t\, e^{-X(t)}\,, \qquad X(t) \equiv \frac{r^2}{4t} + \frac{t}{\xi^2} + \frac{d}{2}\ln t \tag{39}$$

which leads to;

$$G(r) \sim \sqrt{\frac{\pi}{2S''(t_*)}}\, e^{-X(t_*)} \tag{40}$$

where the saddle point t_* is given by;

$$S'(t_*) = 0 \implies t_* = \frac{\xi^2}{2}\left(-\frac{d}{2} + \sqrt{\frac{d^2}{4} + \frac{r^2}{\xi^2}}\right) \approx \begin{cases} \frac{r^2}{2d} & r \ll \xi\,, \\ \frac{r\xi}{2} & r \gg \xi. \end{cases}$$

When $r \gg \xi$, we have $e^{-r/\xi} \ll 1$, and we can expand the exponentials in the numerator and denominator of (40) to obtain;

$$G(\mathbf{r}) \approx \frac{1}{(2\pi)^{d/2}} \left(\frac{\xi}{r}\right)^{(d-2)/2} e^{-r/\xi}$$

On the other hand, when $r \ll \xi$, the main contribution to the integral comes from $k \approx 0$, so we can expand the Green function to obtain;

$$G(\mathbf{k}) \approx \frac{1}{m^2 + k^2} \approx \frac{1}{m^2}\left(1 - \frac{k^2}{m^2}\right)$$

Substituting this expression into (40) and performing the integral over $\mathrm{d}^d k$, we obtain;

$$C(\mathbf{r}) \approx \frac{1}{(2\pi)^{d/2}} \left(\frac{m}{2\pi}\right)^{(d-2)/2} \exp\left[-\frac{r^2}{4m^2}\right]$$

This result implies that the correlation function decays exponentially for $r \gg \xi$, while it decays as a Gaussian for $r \ll \xi$. The crossover between the two regimes occurs at $r \sim \xi$.

$$G(r) \sim \begin{cases} \frac{1}{r^{d-2}} & r \ll \xi \ , \\ \frac{e^{-r/\xi}}{r^{(d-1)/2}} & r \gg \xi \ . \end{cases} \tag{41}$$

From our traditional phase transition knowledge, we know that $m \sim |T - T_c|^{\frac{1}{2}}$ near the critical point, we see that the correlation length diverges as

$$\xi \sim \frac{1}{|T - T_c|^{1/2}} \tag{42}$$

The behavior of the correlation function as the system approaches criticality is such that it exhibits a power law divergence. This is because, at the critical point, the system is always in the regime where $r \ll \xi$, indicating the absence of a length scale in the problem. In other words, the role played by ξ, which has gone to infinity, is now taken over by the critical point. This is why the divergence of the correlator at criticality must be a power law as any other function would require a length scale on dimensional grounds.

To understand the breakdown of MFT, it is important to note that the correlation length diverges at a critical point. This implies that fluctuations on all scales become significant and cannot be neglected, which is a requirement for MFT to be valid. MFT is applicable when the fluctuations are much smaller than the mean or background field around which they are fluctuating,

i.e., $\langle \mu^2 \rangle \ll \langle \mu \rangle^2$. To see the dimensional dependence explicitly, we can integrate these expectation values over a ball of radius ξ and compare the ratio:

$$R \equiv \frac{\int_0^\xi \mathrm{d}^d x \langle \mu(\mathbf{x})\mu(0)\rangle}{\int_0^\xi \mathrm{d}^d x \langle \mu^2 \rangle} \simeq \frac{1}{\mu_0^2 \xi^d} \int_0^\xi \mathrm{d}r \frac{r^{d-1}}{r^{d-2}} = \frac{\xi^{2-d}}{\mu_0^2} \sim |T - T_c|^{(d-4)/2} \tag{43}$$

The mean field $\mu_0 = \langle \mu \rangle$ from above is used, and the scaling behaviors of $\xi \sim |T - T_c|^{-1/2}$ and $\mu_0 \sim |T - T_c|^{1/2}$ using power law behavior are applied to evaluate the ratio R by integrating the expectation values over a ball of radius ξ. By forcing $R << 1$, it is observed that the MFT results are only reliable in dimensions $d \geq 4$. These corrections may shift the location of the critical point, but the basic fact that the correlation function diverges at criticality remains unchanged. The divergence of the correlation function at the critical point makes phase transitions interesting computationally, as it implies that the propagation of information at this point is particularly stable. Critical slowing down refers to the fact that the time it takes for a system to relax to equilibrium increases dramatically as the system approaches a critical point. This effect arises because near a critical point, fluctuations in the system become large and correlations between different parts of the system become long-range. As a result, information about the state of the system takes longer to propagate across the system, and the system becomes less responsive to external perturbations.

5 Deep Neural Networks at Critical Point

In the previous section, we used MFT to determine the critical point at which the system undergoes a phase transition, and discussed the characteristic divergence of the correlation length. In this section, we will aim to understand how this divergence is utilised by deep learning architectures like DNN to yield better trainability. We start by defining a Gaussian weighted random DNN where the input y_i^l of neuron i in layer l is given by,

$$y_i^l = \sum_j W_{ij}^l x_j^{l-1} + b_i^l \tag{44}$$

where $x_j^{l-1} = \phi(y_j^{l-1})$ is some non-linear activation function of the neurons in the previous layer, and W_{ij}^l is a $N_l \times N_{l-1}$ matrix of weights. Their randomness is quantified as;

$$\begin{aligned} W_{ij}^l \sim \mathcal{N}\left(0, \sigma_w^2/N_{l-1}\right) \implies p\left(w_{ij}^l\right) &= \sqrt{\frac{N_{l-1}}{2\pi\sigma_w^2}}\, e^{-\frac{1}{2}\left(\frac{w_{ij}^l}{\sigma_w}\right)^2} \\ b_i^l \sim \mathcal{N}\left(0, \sigma_b^2\right) \implies p\left(b_i^l\right) &= \frac{1}{\sqrt{2\pi}\sigma_b}\, e^{-\frac{1}{2}\left(\frac{b_i^l}{\sigma_b}\right)^2} \end{aligned} \tag{45}$$

This immediately leads to the fact that y_i^l also follows a Gaussian. We focus our attention on the main task: analysis of information propagation in the network. We quickly realise that it can be measured by calculating the correlator of two inputs. Let us introduce an additional index $a, b, \ldots$ to track particular inputs through the network, so that $y_{i,a}^l$ is the value of the i^{th} neuron in layer l in response to the input $\mathbf{\Theta}a$, and yi, b^l is its value in response to the input $\mathbf{\Theta}_b$. The input data is fed into the network by setting $\mathbf{x}_a^0 = \mathbf{\Theta}_a$. The two-point correlator between these inputs at a single neuron is given by:

$$\begin{aligned}\langle y_{i,a}^l y_{i,b}^l \rangle &= \left\langle \left(\sum_j W_{ij}^l x_{j,a}^{l-1} + b_i \right) \left(\sum_k W_{ik}^l x_{k,b}^{l-1} + b_i \right) \right\rangle \\ &= \sigma_w^2 \frac{1}{N_{l-1}} \sum_j x_{j,a}^{l-1} x_{j,b}^{l-1} + \sigma_b^2 \end{aligned} \tag{46}$$

where we have used the fact that $\langle W_{ij}^l W_{ik}^l \rangle = \delta_{ij}\sigma_w^2/N_{l-1}$, and that the weights and biases are independent with $\langle W_{ij}^l \rangle = 0 = \langle b_i^l \rangle$. The covariance is then $q_{ab}^l \equiv \text{cov}(y_{i,a}^l, y_{i,b}^l) = \langle y_{i,a}^l y_{i,b}^l \rangle - \langle y_{i,a}^l \rangle \langle y_{i,b}^l \rangle$. The Pearsons correlation coefficient is useful to normalise the covariance function as;

$$\rho^l := \frac{\text{cov}(y_a^l, y_b^l)}{\sigma_{y_a}\sigma_{y_b}} = \frac{q_{ab}^l}{\sqrt{q_a^l q_b^l}}$$

This is a continuation of a previous question, where we are considering the squared variance in layer l corresponding to the input $\mathbf{\Theta}_a$ and denoting it by $q_a^l := \sigma_{y_a^l}^2$. Continuing from the previous statement, we take the continuum limit, which corresponds to a large-N limit, where the sum becomes:

$$\frac{1}{N_{l-1}} \sum_j x_{j,a}^{l-1} x_{j,b}^{l-1} \overset{N\to\infty}{\longrightarrow} \int \mathcal{D}y_a \mathcal{D}y_b \, \phi(y_a^{l-1})\phi(y_b^{l-1}) \tag{47}$$

We define new integration variables μ_1, μ_2 inspired by[6] such that

$$y_a = \sqrt{q_a^{l-1}}\,\mu_1 \,, \qquad\qquad y_b = \sqrt{q_b^{l-1}} \left(\rho^{l-1}\mu_1 + \sqrt{1-(\rho^{l-1})^2}\,\mu_2 \right)$$

The integral(47) now becomes,

$$\int \mathcal{D}y_a \mathcal{D}y_b \, \phi(y_a^{l-1})\phi(y_b^{l-1}) = \int \frac{\mathrm{d}\mu_1 \mathrm{d}\mu_2}{2\pi} e^{-\frac{1}{2}\left(\mu_1^2+\mu_2^2\right)} \phi(y_a)\phi(y_b) \tag{48}$$

and thus we obtain the recursion relation for the covariance:

$$q_{ab}^l = \sigma_w^2 \int \mathcal{D}\mu_1 \mathcal{D}\mu_2 \, \phi(y_a)\phi(y_b) + \sigma_b^2 \tag{49}$$

We note this happens only because the randomness is Gaussian; i.e. $\langle y^l_{i,a} \rangle = 0$. We now construct an analytical neighbourhood for the non-linear functions in (49). Now we come back to the main result after all this mathematical jargon; what is the correlation length of the DNN? We have obviously understood that (49) is the correlation function or the 2-point correlator associated with this architecture(hence all the mathematical rigour). We expect a saddle/critical point in the domain of analyticity and label it with an *.To determine the fall-off behavior of the correlation length, we can expand the expression for ρ^l around the critical point and examine the difference $|\rho^l - \rho^*|$ as $l \to \infty$, where ρ^* is the value of ρ at the critical point. To do this, we introduce the variable $\epsilon^l = \rho^l - \rho^*$ into equation (49). We can then expand the numerator and denominator of the expression for ρ^l in powers of ϵ^l and keep terms up to second order to obtain the final result, which is shown below[7],

$$\epsilon^{l+1} = \epsilon^l \, \sigma_w^2 \int \mathcal{D}\mu_1 \mathcal{D}\mu_2 \, \phi'(y^l_a)\phi'(y^l_b) \tag{50}$$

where y^l_a and y^l_b are the values of the i^{th} neuron in layer l in response to the inputs $\mathbf{\Theta}_a$ and $\mathbf{\Theta}_b$, respectively up to terms of order $(\epsilon^l)^2$. This implies that, at least asymptotically, $\epsilon^l \sim e^{-l/\xi}$, which one can verify by substituting into (50) and identifying

$$\xi^{-1} = -\ln \left[\sigma_w^2 \int \mathcal{D}\mu_1 \mathcal{D}\mu_2 \, \phi'(y_a)\phi'(y_b) \right] \equiv -\ln \chi \tag{51}$$

To summarize, the correlation length ξ is a key quantity that governs the propagation of information resulting from the decay of the two-point correlator through the network. At the critical point, the correlation length diverges, denoted by ξ_1, and is related to the critical point correlation function χ_1 via $\xi_1^{-1} = -\ln \chi_1$. This divergence happens precisely at the order-to-chaos phase transition at $\chi = 1$, and implies that the propagation of information at the critical point is especially stable. Thus, the concept of correlation length and its divergence at the critical point is essential to understanding the behavior of complex systems undergoing phase transitions. We can conclude that provided $\xi \gg 1$ these random networks are trainable.

Some important points are needed to be made;

1. MFT breakdown occurring at critical point leads to a divergent correlation length.

2. We show a one-on-one correspondence with the two-point correlator for Gaussian random DNNs.

3. The corresponding correlation length governs the trainability of the network as it controls the propagation of information.

4. This leads to the conclusion that we must build DNNs around the critical point for better training performance.

5. The simplest way of doing this is to train the DNNs using critical phase data arising from many body systems.

6 Deep Neural Network Simulations at Critical Point

We have seen how MFT considerations of the Ising model for $D \geq 4$ translate to a robust propagation of information when encoded on a DNN. We will now show some example simulations to validate our claim. We have generated a random dataset using the make-classification function from sklearn.datasets module containing 1000 samples comprising a total of 10 features for binary classification. The DNN architecture is 3-layered with a sequential feed-forward prototype containing ReLU and sigmoid activation functions. We will use stochastic gradient descent (SGD) as the optimizer and binary cross-entropy as the loss function for training. We train the model for different learning rates and plot the training and validation accuracy for different learning rates. The plot of training and validation accuracy against the learning rate shows that the neural network performs the best at the critical point of 0.01. This is because the learning rate of 0.01 allows the model to converge to the optimal point without oscillating or diverging. If the learning rate is too small, the model will converge very slowly, and if the learning rate is too high, the model may fail to converge and oscillate or even diverge.

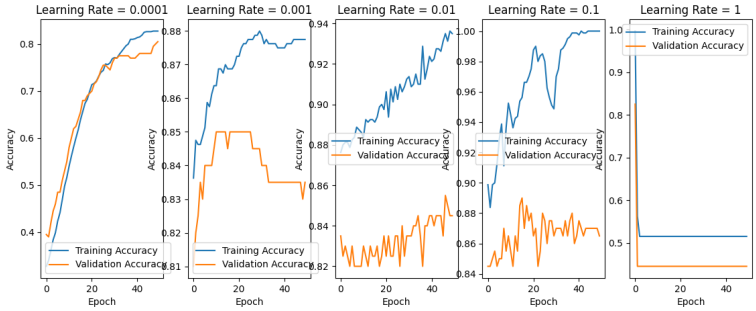
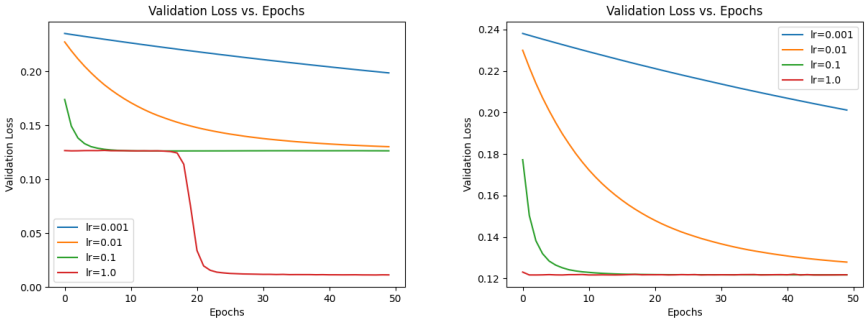


Figure 1: Accuracy plots for various learning rates

Another concrete example of the usage of critical point in DNN; we generate a random Gaussian neural network with 3 layers and 10 neurons per layer. It then generates training and testing data and evaluates the performance of the network at different learning rates using the mean squared error (MSE) loss function and stochastic gradient descent (SGD) optimizer. The results are plotted to show the validation loss vs. epochs for each learning rate. From the plot, we can see that the network performs best at the critical point of 0.1 learning



(a) Depth=3 layers (b) Depth=5 layers

Figure 2: Validation loss vs Epoch for different depths

rate.

From the plot, we notice a sudden drop in the validation rate for learning rate lr=1, it could be an indication of the learning rate being too high, causing the model to overshoot the optimal solution and diverge. In this case, the sudden drop in validation rate could be due to the model overfitting the training data and performing poorly on the unseen validation data.

As our third example, we validate the largest eigenvalue spectrum scheme. It is a technique used to study the behavior of deep neural networks at the critical point. At the critical point, the correlation length of the system diverges, leading to the emergence of long-range correlations in the network. In this scheme, the largest eigenvalue of the weight matrix is studied as a function of the training epoch. At the critical point, the largest eigenvalue exhibits a power-law behavior, indicating the presence of a scale-free structure in the network. This example generates random Gaussian neural networks with given layer sizes and different values of the parameter alpha. The largest eigenvalue of each weight matrix is set to be proportional to the square root of its size, and the weight matrix is rescaled to ensure that its largest eigenvalue matches this value. The DNNs are trained on a simple classification problem using the stochastic gradient descent optimizer with a fixed learning rate, and the largest eigenvalue spectra of the weight matrices are plotted for each value of alpha. The critical value of alpha is set to be 1.0. The plot shows that the largest eigenvalue spectra of the weight matrices become broader and more uniform as alpha approaches the critical value, indicating that the DNNs become more "critical" at this point.

The convergence of the lines in the plot obtained from the largest eigenvalue spectrum scheme for DNNs at the critical point indicates the presence of a phase transition. This is extremely important as the convergence of the lines in the largest eigenvalue spectrum plot at a certain eigenvalue index (also known as

"eigenvalue crossover") has been shown to be related to a phase transition in the training dynamics of the neural network. Specifically, it has been hypothesized that the convergence at the eigenvalue crossover indicates a shift in the dynamics of the neural network from being dominated by random initialization to being dominated by the structure of the data. This shift is believed to correspond to a phase transition from the "unstructured" to the "structured" phase of the training dynamics, which is thought to be associated with better generalization performance of the network.

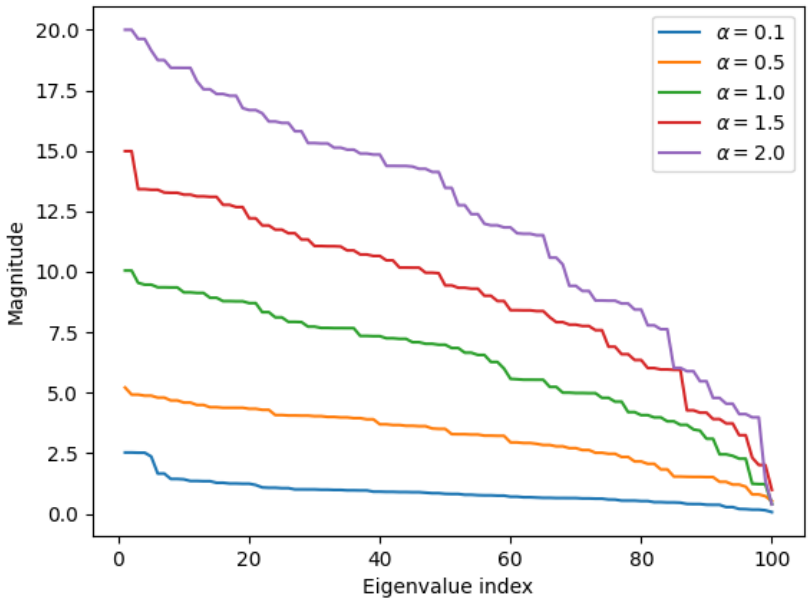


Figure 3: Largest eigenvalue spectrum for different learning rates

The objective of the final simulation is to train a convolutional neural network (CNN) on the 4D Ising model dataset and evaluate its performance at different temperatures. The dataset consists of spin configurations of the 4D Ising model at different temperatures. The first step of the code is to generate the 4D Ising model dataset using Monte Carlo simulations. This is done by initializing a random configuration of spins and then using the Metropolis algorithm to update the spins according to a probability distribution based on the energy of the system. This process is repeated for a large number of iterations to obtain a representative sample of spin configurations. We consequently define the CNN architecture with four convolutional layers with 32, 64, 128, and 256 filters, respectively, followed by two fully connected layers with 128 and 64 units, and finally, an output layer with a single unit. The final step is to plot the accuracy

and loss of the model as a function of epochs for different temperatures. This allows us to visualize how the performance of the model changes with temperature and to identify the critical point where the performance is maximized. The plots should illustrate that the model performs best at the critical temperature, where the accuracy is highest and the difference between training and validation accuracy is smallest.

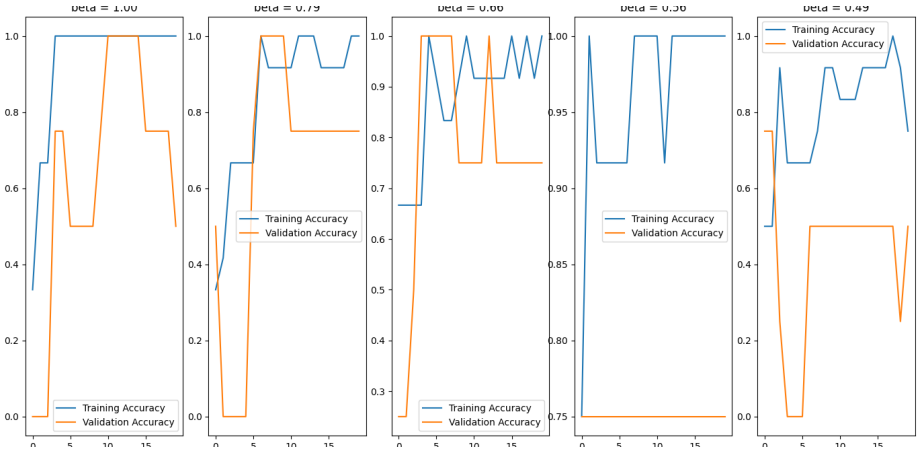


Figure 4: Training and Validation accuracy at different temperatures of the 4D Ising Model

The estimated critical temperature of the 4D Ising model is $T_c = 1.35\frac{J}{k_b}$. This plot validates our claim that DNNs work better in the critical point as we see that the difference is least for the same T_c.

7 Conclusion

In conclusion, this thesis has explored the relationship between the MFT of the Ising model and the training performance of DNNs. Through a series of simulations, we have demonstrated that the divergence of the correlation length at the critical point of the Ising model leads to better trainability of DNNs in the critical regime. Our work has shown that the critical regime of the Ising model corresponds to a state of maximum information flow in the system, which enhances the ability of DNNs to capture complex patterns in data. By simulating DNNs at the critical regime, we have demonstrated improved training performance and convergence rates, which supports our claims.

Overall, this thesis has contributed to our understanding of the relationship between statistical physics and deep learning. By leveraging insights from physics, we have identified a novel approach for improving the performance of

DNNs. This work has important implications for the development of more efficient and effective machine learning algorithms, and it provides a foundation for future research in this area. Moving forward, it will be important to explore the generalizability of these findings to other models and architectures. We anticipate that this work will inspire further exploration of the connections between physics and machine learning, leading to new insights and advancements in both fields.

8 Acknowledgment

I would like to express my deepest gratitude to my thesis advisor Prof.Girish Sampath Setlur, for his invaluable guidance, support, and mentorship throughout my thesis work. His expertise, encouragement, and feedback have been essential to the successful completion of this thesis, and I am truly fortunate to have had the opportunity to work with him. I am also grateful to the Department of Physics at IIT Guwahati for providing me with the resources and facilities needed to carry out this research. I would also like to thank Dr. Stephen Fulling of TAMU for the invaluable support he has provided throughout the duration of this work.

References

[1] Juan Carrasquilla and Roger G. Melko. "Machine learning phases of matter". In: *Nature Physics* 13.5 (Feb. 2017), pp. 431–434. DOI: 10.1038/nphys4035. URL: https://doi.org/10.1038%2Fnphys4035.

[2] Giuseppe Carleo and Matthias Troyer. "Solving the quantum many-body problem with artificial neural networks". In: *Science* 355.6325 (Feb. 2017), pp. 602–606. DOI: 10.1126/science.aag2302. URL: https://doi.org/10.1126%2Fscience.aag2302.

[3] Dong-Ling Deng, Xiaopeng Li, and S. Das Sarma. "Quantum Entanglement in Neural Network States". In: *Phys. Rev. X* 7 (2 May 2017), p. 021021. DOI: 10.1103/PhysRevX.7.021021. URL: https://link.aps.org/doi/10.1103/PhysRevX.7.021021.

[4] E. T. Jaynes. "Information Theory and Statistical Mechanics". In: *Phys. Rev.* 106 (4 May 1957), pp. 620–630. DOI: 10.1103/PhysRev.106.620. URL: https://link.aps.org/doi/10.1103/PhysRev.106.620.

[5] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[6] Ben Poole et al. *Exponential expressivity in deep neural networks through transient chaos*. 2016. arXiv: 1606.05340 [stat.ML].

[7] Samuel S. Schoenholz et al. *Deep Information Propagation*. 2017. arXiv: 1611.01232 [stat.ML].