

Deep Learning And Critical Behaviour Of The Ising Model

Uddhav Sen

Supervised by Prof. Girish Setlur

Department of Physics, IIT Guwahati

March 7, 2025

Outline

- 1 Introduction
- 2 MC Simulations
- 3 Max-Ent
- 4 MFT and Criticality
- 5 DNNs at Criticality
- 6 DNN Codes
- 7 Conclusion

Beginning Thoughts

- The main purpose of the thesis is to understand how many body systems in physics *correlate* with information propagation in machine learning architectures.
- There has been a lot of effort spent to formally get the ball rolling in this direction.
- Feynman[Feynman 1982] and Yuri Mann's ideas about how simulations and consequently information processing should be inherently quantum to better mimic the true nature of reality have revolutionalised the field.
- The Universal Approximation Theorem (UAT) on the other hand, provides us with strict laws on how increasing computational resources leads to better approximation.

Motivation

- UAT has paved the way for modern-day machine learning and intelligent information processing.
- Several influential papers[Schoenholz et al. 2017][Poole et al. 2016] from the Ganguli group at Stanford have tried to show how many body models can be used to make Deep Neural Networks(DNN) perform better.
- These papers show how physics-informed DNNs perform better than physics-ignorant ones.
- We take encouragement from these works and use the d -dimensional Ising universality class of models to see how ideas developed in the mean field theory (MFT) limit gives rise to stable information propagation in random neural architectures in the critical regime.

Qualitative Arguments

- We will show how the two-point correlator defined in a random DNN is governed by the correlation length of the mean-field model.
- The correlation function is known to diverge in the critical regime.
- The two-point correlator analogously falls off in the critical regime which leads to stable information flow between layers in the random DNN.
- This implies that even random DNNs are trainable!!!

Structure

- The thesis is a step to show that ideas from physics can optimise information processing laws and can be used for efficient simulation of otherwise intractable many-body models in physics.
- The beamer presentation starts with Monte Carlo simulations of the 2D Ising model and the critical behaviour of observables like specific heat, magnetisation, susceptibility etc.
- The main chunk of the work focuses on theoretical motive for using criticality in DNNs followed by simulation results. We conclude with the wide range of future areas of research in the highly interdisciplinary line of work.

Background

The first part of my M.Sc. thesis was carried out in the 3rd Semester.

- The report included the theoretical background of how phase transition occurs at $T = 0$ for 1D.
- The partition function was calculated by the well-known transfer matrix method.
- We moved onto the 2D case to look for a positive critical temperature T_c .

2D Ising Model: Definition

The 2D Ising Hamiltonian can be written as,

$$\mathcal{H} = -J \sum_{i,j \in \Omega: |i-j|=1} s_i s_j \quad (1)$$

Here $J > 0$ and Ω is a square lattice of dimension L . We define a parameter known as Magnetisation M as;

$$M = \frac{1}{|\Omega|} \sum_{i \in \Omega} s_i \quad (2)$$

M is zero in the disordered state and it is non-zero in the ordered state. This change occurs due to a phase transition at critical temperature T_c and the system stays ferromagnetic or ordered for $T < T_c$.

Metropolis Algorithm: Idea

- The proof of a positive T_c given by the Ising model is given by Peierl's argument.
- Although we will look at a generalised derivation of how to calculate observables of the d –dimensional Ising class later; we will be focusing on simulation techniques in this section.
- Monte Carlo methods are a class of sampling algorithms for Markov-Chain-based random processes and are extremely important in many branches of physics.
- The basic idea of the algorithm is to compute a function dependent on paths and configurations based on parameter space.

Metropolis Algorithm: Theory

- We aim to find a unique and optimal probability distribution that follows the equilibrium distribution policy of a Markov chain; $\pi^t P \pi = \mathbf{1}$.
- P is the transition matrix and π is the equilibrium distribution.
- The 2D Ising model can also be simulated using this algorithm as we try to minimize the Hamiltonian by varying the path parameters.
- More precisely, we start with the Partition function $\mathcal{Z} = \sum_k e^{-\beta \mathcal{H}(s_k)}$ where s_k is the set of all spin configurations in the system.

Metropolis Algorithm: Theory(contd.)

- Expectation of any observable $\langle O \rangle$ can be calculated as a weighted sum w.r.t. the Boltzmann weights;

$$\begin{aligned}\langle O \rangle &= \sum_k p(k) O(k) \\ &= \frac{1}{Z} \sum_k O(k) e^{-\beta \mathcal{H}(s_k)}\end{aligned}\quad (3)$$

- The algorithm is used to simulate energy, magnetization, specific heat and susceptibility of the system as it aims to reach equilibrium.

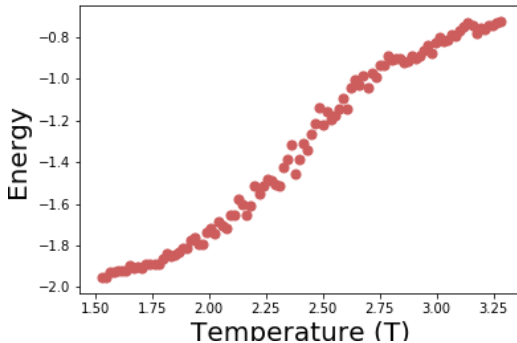
Metropolis Algorithm: Pseudocode

- 1 Construct the initial Hamiltonian \mathcal{H}_i for an initial configuration of $|\Omega|$ spins.
- 2 For any random $\omega \in \Omega$; perform $s_\omega \rightarrow -s_\omega$ and recompute the Hamiltonian \mathcal{H}_f .
- 3 Calculate $\Delta E = \mathcal{H}_f - \mathcal{H}_i$.
- 4 if $\Delta E < 0$ then change initial spin configuration i to f .
- 5 if $\Delta E > 0$, then accept the change with a probability of $e^{-\beta \Delta E}$ to satisfy the condition of detailed balance;
 $\pi_i P_{ij} = \pi_j P_{ji}$.

We repeat the process till we reach the equilibrium distribution and consequently get the expectation values of interesting observables O .

Simulations: Energy

- The energy is given by the expectation of the Hamiltonian $\mathcal{H} = -J \sum_{i,j \in \Omega: |i-j|=1} s_i s_j$ as $\langle \mathcal{H} \rangle = \frac{1}{\mathcal{Z}} \sum_k \mathcal{H}(k) e^{-\beta \mathcal{H}(s_k)}$.



Simulations: Magnetisation

- The curve in the next slide highlights a gradual decrease in magnetization and a characteristic drop at around the theoretically calculated $T_c = 2.27K$.
- The lower values of M at high temperatures imply that the system is more disordered as there exhibiting paramagnetic behavior.
- As $T \rightarrow 0$ the spins align themselves to showcase ferromagnetic behavior.

Simulations: Magnetisation(contd.)

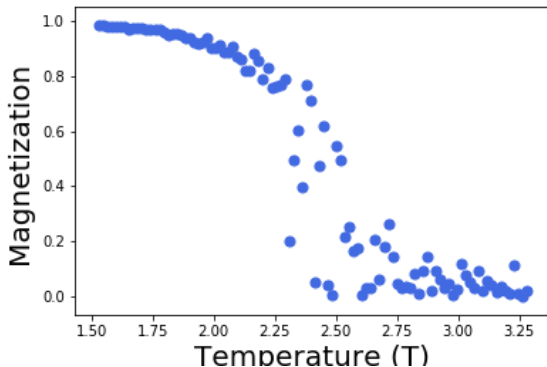


Figure: Metropolis Simulation for Magnetisation.

Simulations: Specific Heat

- Specific heat is defined as the first derivative of the expected energy of the system.
- For a more formal comparison, we note that the critical exponent α for the 2-D mean field model is related to specific heat as $C \propto \frac{1}{(T-T_c)^\alpha}$.
- The simulation plot in the next slide clearly supports the diverging nature of C at T_c .

Simulations: Specific Heat(contd.)

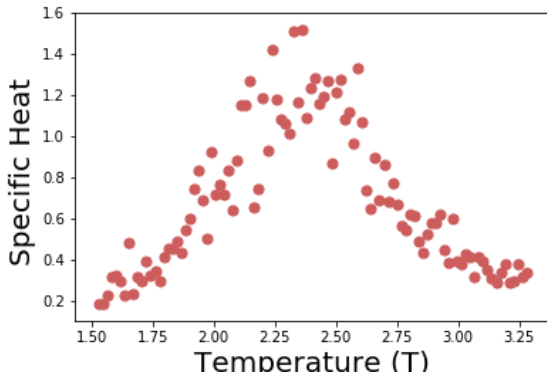


Figure: Metropolis Simulation for Specific Heat.

Simulations: Susceptibility(contd.)

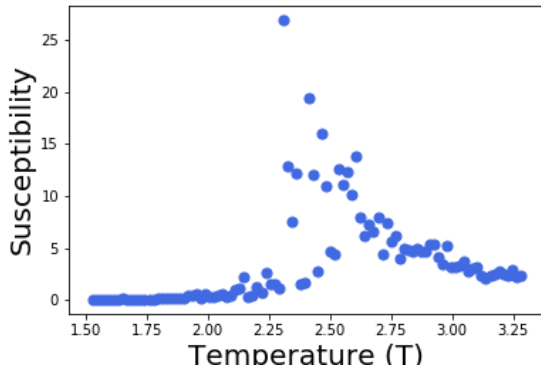


Figure: Metropolis Simulation for Susceptibility.

Simulations: System

- We can see the gradual shift in the configuration of spins by plotting snapshots of the configurations in time with two different colors highlighting the two different spins allowed in the system.
- The following plots show the system reaching its equilibrium state of least interaction energy.
- I also have simulated the 4D Ising model with and without vacancy disorder and calculated the same set of observables.
- The codes for the 4D case are all available in my **Github** repository.

Simulations: System(contd.)

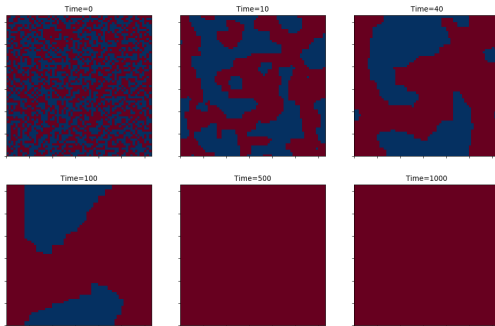


Figure: Metropolis Simulation of system configurations.

Principal Idea

- The marriage between information theory and MFT becomes 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[Jaynes 1957].
- 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[Shannon 1948], these two concepts are essentially identical and can be expressed as $S = -\sum_i p_i \ln p_i$ where p_i is the probability of the i^{th} outcome/microstate.

Entropy

- 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, \dots, x_n) representing specific outcomes or measurement values.
- Each value x_i is associated with a corresponding probability $p_i \in (p_1, \dots, p_n)$, such that $\sum_i p_i = 1$.
- Our goal is to find a function $H(p_1, \dots, p_n)$ that accurately measures the amount of uncertainty represented by this probability distribution.

Constraints on H

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 $A(n) \equiv H(1/n, \dots, 1/n)$ 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; $p_i = \frac{n_i}{\sum n_i}$, $n_i \in \mathbb{N}$.

Evaluation of H

- We fix the form of H by applying the composition law where we coarse-grain n equally likely alternatives into clusters of size n_i to get $A(\sum_i n_i) = H(p_1, \dots, p_n) + \sum_i p_i A(n_i)$.
- We consider the equally likely outcome where $n_i = m$, in which case $\sum_i n_i = nm$, and we get $A(nm) = A(n) + A(m)$.
- This has a well-known solution which is $A(n) = K \ln n$ where $K > 0$.
- After substitution, we get $H = -\sum_i p_i \ln p_i$.
- We notice that H reproduces the Boltzmann definition of entropy in physics (for statistical systems, we have $S = \ln \Omega$).

Maximum Entropy Principle

- 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 maximises entropy while meeting known constraints.
- With the connection firmly in hand, we introduce a constrained optimisation algorithm to determine the best-suited probability distribution function.

Derivation of Probabilities

Mathematically, we aim to maximise entropy provided the given constraints using the Lagrange multiplier method. The objective function is given as;

$$\mathcal{L}(p_i; \alpha, \beta) = - \sum_i p_i \ln p_i + \alpha \left(\sum_i p_i - 1 \right) + \beta \left(\sum_i p_i f(x_i) - \langle f(x) \rangle \right) \quad (4)$$

Taking the gradients to zero, we get;

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial p_j} &= - \sum_i \delta_{ij} \ln p_i - \sum_i p_i \partial_j \ln p_i + \alpha \sum_i \delta_{ij} + \beta \sum_i \delta_{ij} f(x_i) = 0 \\ \implies -\ln p_j - 1 + \alpha + \beta f(x_j) &= 0 \\ \implies \ln p_j &= -\alpha - \beta f(x_j) \\ \implies p_j &= e^{-\alpha - \beta f(x_j)} \end{aligned} \quad (5)$$

Final Points

- We have redefined $\alpha \rightarrow \alpha - 1$ 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.
- We will use this directly in our MFT calculation.

The Hamiltonian

We consider the d -dimensional Ising Hamiltonian with N spins isotropically (i.e., $\frac{N}{d}$ spins per direction):

$$\hat{\mathcal{H}} = -\frac{1}{2} \sum_{i,j} \hat{J}_{ij} s_i s_j - \sum_i \hat{h}_i s_i \quad (6)$$

where $i = 1, 2, \dots, N$ and $s_i = \pm 1$. We assume that the correlation between spins at different sites is negligible and replace s_j with its mean value $\langle s_j \rangle = s$ and neglect terms of order δs^2 , which is valid 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.

$$\begin{aligned} s_i &= (s + (s_i - s)) \\ s_j &= (s + (s_j - s)) \end{aligned} \quad (7)$$

The MF Hamiltonian

We insert (7) to get the mean-field Hamiltonian;

$$\begin{aligned}
 \hat{\mathcal{H}}_{MF} &= -\frac{1}{2} \sum_{i,j} (-s^2 + s(s_i + s_j) + \delta s_i \delta s_j) - \sum_i \hat{h}_i s_i \\
 &= \frac{s^2}{2} \sum_{j,i} \hat{J}_{ij} - \frac{s}{2} \sum_{j,i} (s_i + s_j) \hat{J}_{ij} - \sum_i \hat{h}_i s_i + O(\delta s^2) \\
 &= \frac{s^2}{2} \sum_{j,i} \hat{J}_{ij} + \sum_i s_i \left(-\sum_j \hat{J}_{ij} s - \hat{h} \right) + O(\delta s^2) \quad (8)
 \end{aligned}$$

Symmetric pairwise interactions have been assumed, where $\hat{J}_{ij} = \hat{J}_{ji}$.

Max-Ent and MFT

- If we have no a priori knowledge about the p.d.f. w.r.t. which a particular observable 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, $p_i \equiv p(x_i) = \frac{1}{Z[\beta]} e^{-\beta E_i}$ 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.
- 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.
- 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.

MF Partition Function

- We can easily calculate $Z_{MF} = 2^N e^{-NdJs^2} \cosh^N(2dJs + h)$.
- The corresponding free energy per unit particle is
$$f_{MF}(s) = -\frac{1}{N\beta} \ln Z = \frac{dJs^2}{\beta} - \frac{\ln \cosh(2dJs + h)}{\beta}.$$
- The interaction strength is represented by $J \equiv \beta \hat{J}$, while the strength of the external magnetic field is $h \equiv \beta \hat{h}$.
- Since we already know from Bogolyubov inequality that f_{MF} provides an upper bound on the true free energy, we obtain the tightest possible bound by minimizing over s : $\frac{\partial f_{MF}}{\partial s} = 0$.
- We get the condition commonly known as self-consistent: $s = \tanh(2dJs + h)$.

Beginnings of Criticality

We focus on the $h=0$ case: the critical point s_0 will always be small ($|s_0| < 1$) and independent of T (as $\lim_{x \rightarrow \pm\infty} \tanh x = \pm 1$),

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

f_{MF} near the critical point is approximately $f_{MF}(s_0) \approx \frac{r}{2}s_0^2 + \frac{g}{4!}s_0^4$
 with $r \equiv \frac{2dJ}{\beta} (1 - 2dJ) = \frac{T_c}{T} (T - T_c)$, $g \equiv \frac{32d^4 J^4}{\beta} = \frac{2T_c^4}{T^3}$.
 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.

Continuum Limit

- MFT is limited in its ability to accurately predict the critical point, which is precisely where it fails.
- This is because fluctuations at all scales become significant whereas MFT only considers fluctuations to linear order.
- It is necessary to move beyond the discrete lattice of MFT and employ a continuum field theory.
- We will discover that MFT corresponds to the dominant saddle point approximation of the quantum field theory.

The QFT Partition Function

- 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 to express ϕ_i in terms of μ_i as $\phi_i = J_{ij}^{-1}(\mu_j - h_j)$.
- Using the above equations alongwith $d\phi_i = J_{ij}^{-1}d\mu^j$; $\implies d^N\phi = \det J^{-1}d^N\mu$, we calculate Z in the local limit ($|\mathbf{y} - \mathbf{x}| \ll 1$).
- We also note that in the continuum limit $\mu_i \rightarrow \mu(\mathbf{x})$ and $\sum_i \rightarrow \int d^d\mathbf{x}$.
- We denote the minimum as μ_0 and expand the action to second order in the fluctuations $\delta\mu \equiv (\mu - \mu_0)$.

The QFT Partition Function(contd.)

- We skip a huge number of steps(all present in the thesis) and write the final form in the saddle point approximation as;

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

- The analytic functions \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.
- The unit free energy is

$$f_{MF} = \frac{1}{\beta} \left(\tilde{r} \mu_0^2 + \frac{\tilde{g}}{2} \mu_0^4 - \frac{3}{4} h \mu_0 \right) \equiv \frac{\hat{r}}{2} \hat{\mu}_0^2 + \frac{\hat{g}}{4!} \hat{\mu}_0^4 - \hat{h} \mu_0.$$
- We see that the free energy matches with lattice f_{MF} upto a source term!!

The QFT Partition Function(contd.)

- The key takeaway from this exercise in field theory is that MFT corresponds to the leading saddle point of Z .
- Moving forward, we exploit the Gaussian nature of the action and evaluate the partition function using standard techniques.
- 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 DNNs.

The Two-Point Correlator

- The calculation of the Green's function $G(\mathbf{x} - \mathbf{y})$ is closely followed from David Tong's SFT lectures.
- As before, we skip the long and tedious calculation already done in the thesis.
- There is a well known approach of transforming the system to momentum space and computing the tedious Gaussian integrals to get the correlation function in field theory.

- The result is[Kopietz, Bartosch, and Schütz 2010]

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

- 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 $r \sim \xi$ where we have defined $m^2 = \frac{1}{\xi^2}$.

The Two-Point Correlator(contd.)

$$G(r) \sim \begin{cases} \frac{1}{r^{d-2}} & r \ll \xi, \\ \frac{e^{-r/\xi}}{r^{(d-1)/2}} & r \gg \xi. \end{cases} \quad (9)$$

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

- 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., $\langle \mu^2 \rangle \ll \langle \mu \rangle^2$.

Dimensionality Dependence

- To see the dimensional dependence explicitly, we can integrate these expectation values over a ball of radius ξ and compare the ratio:

$$\begin{aligned}
 R &\equiv \frac{\int_0^\xi d^d x \langle \mu(\mathbf{x}) \mu(0) \rangle}{\int_0^\xi d^d x \langle \mu^2 \rangle} \simeq \frac{1}{\mu_0^2 \xi^d} \int_0^\xi dr \frac{r^{d-1}}{r^{d-2}} \\
 &= \frac{\xi^{2-d}}{\mu_0^2} \sim |T - T_c|^{(d-4)/2}
 \end{aligned} \tag{10}$$

- By forcing $R \ll 1$, it is observed that the MFT results are only reliable in dimensions $d \geq 4$.

Connections to a DNN?

- 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.

Random DNN: Definition

- We start by defining a Gaussian weighted random DNN where the input y_i^l of neuron i in layer l is given by $y_i^l = \sum_j W_{ij}^l x_j^{l-1} + b_i^l$ 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;

$$W_{ij}^l \sim \mathcal{N}(0, \sigma_w^2 / N_{l-1}) \implies p(w_{ij}^l) = \sqrt{\frac{N_{l-1}}{2\pi\sigma_w^2}} e^{-\frac{1}{2} \left(\frac{w_{ij}^l}{\sigma_w} \right)^2}$$

$$b_i^l \sim \mathcal{N}(0, \sigma_b^2) \implies p(b_i^l) = \frac{1}{\sqrt{2\pi}\sigma_b} e^{-\frac{1}{2} \left(\frac{b_i^l}{\sigma_b} \right)^2} \quad (11)$$

- This immediately leads to the fact that y_i^l also follows a Gaussian.

Information Propagation Between Layers

- We focus our attention on the main task: analysis of information propagation in the network by evaluating the analogous two-point correlator introduced in MFT.
- Let us introduce an additional index a, b, \dots 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 Θ_a , and $y_{i,b}^l$ is its value in response to the input Θ_b .
- The input data is fed into the network by setting $\mathbf{x}_a^0 = \Theta_a$.
- The two-point correlator between these inputs at a single neuron is given by:

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

Information Propagation Between Layers(contd.)

- The calculation requires (11) and $\langle W_{ij}^l W_{ik}^l \rangle = \delta_{ij} \sigma_w^2 / N_{l-1}$ and the weights and biases are independent with $\langle W_{ij}^l \rangle = 0 = \langle b_i^l \rangle$.
- We take the continuum limit, which corresponds to a large- N limit, where the sum becomes:

$$\frac{1}{N_{l-1}} \sum_j x_{j,a}^{l-1} x_{j,b}^{l-1} \xrightarrow{N \rightarrow \infty} \int \mathcal{D}y_a \mathcal{D}y_b \phi(y_a^{l-1}) \phi(y_b^{l-1}).$$

- The squared variance in layer l corresponding to the input Θ_a is denoted by $q_a^l := \sigma_{y_a^l}^2$.
- We define new integration variables μ_1, μ_2 inspired by [Poole et al. 2016] such that

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

- We note this happens only because the randomness is Gaussian; i.e. $\langle y_{i,a}^l \rangle = 0$.

Similarities With MFT

- 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 $l \rightarrow \infty$, where ρ^* is the value of ρ at the critical point.
- To do this, we introduce the variable $\epsilon^l = \rho^l - \rho^*$ and 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[Schoenholz et al. 2017];

$$\epsilon^{l+1} = \epsilon^l \sigma_w^2 \int \mathcal{D}\mu_1 \mathcal{D}\mu_2 \phi'(y_a^l) \phi'(y_b^l).$$
- This implies that, at least asymptotically, $\epsilon^l \sim e^{-l/\xi}$.

Divergence Finally!!

- We have defined $\xi^{-1} = -\ln [\sigma_w^2 \int \mathcal{D}\mu_1 \mathcal{D}\mu_2 \phi'(y_a) \phi'(y_b)] \equiv -\ln \chi$.
- 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 directly implies that ϵ' becomes extremely slow to perturbations as DNN depth increases!!!
- We can conclude that provided $\xi \gg 1$ these random networks are trainable!!!
- This leads to the conclusion that we must build DNNs around the critical point for better training performance.
- We use the learning rate as the hyperparameter and tune it around the critical regime in our codes.

Example 1: Learning Rate

- 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.
- We train the model for different learning rates and plot the training and validation accuracy for different learning rates as the critical hyperparameter.

Plot

- 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.

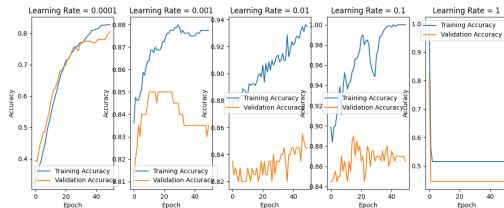
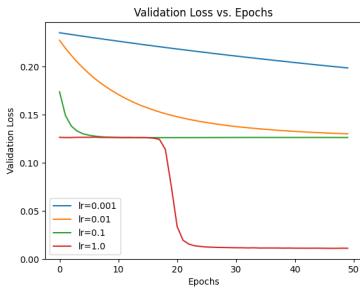


Figure: Accuracy plots for various learning rates.

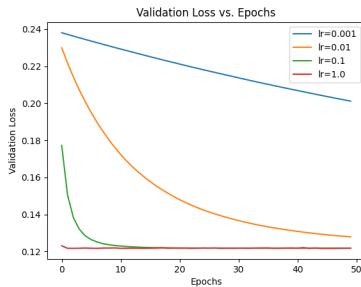
Example 2: Gaussian DNN

- We generate a random Gaussian neural network with 3 layers and 10 neurons per layer with $lr = 0.1$; 10 features for 100 samples.
- 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 rate.

Plot



(a) Depth=3 layers



(b) Depth=5 layers

Figure: Validation loss vs Epoch for different depths

Example 3: 4D Ising Model At Different Temperatures

- 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.

CNN Architecture

- 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.

Plot

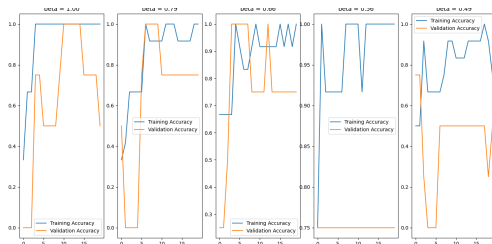


Figure: Training and Validation accuracy at different temperatures of the 4D Ising Model

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 .

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 support our claims.

Future Scope

- We hope to employ tensor networks and Quantum Cellular Automata to expand our network of Ising Universality class to open quantum many body systems.
- We have realised that local order correlations are extremely difficult to compute for high-dimensional systems.
- They are also extremely difficult to predict using DNNs as machine learning algorithms are not the best extrapolators.
- To combat these issues, we need a mathematical structure/formalism which can represent large dimensional systems to a tensor combination of small dimensional systems.
- This will be extremely important in physics because we could then find a simpler tensor product rule comprised of one-body Hamiltonians of a complex many-body Hamiltonian.

Acknowledgement

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. As already stated, all the codes are available in my **Github** repository.

References I



Feynman, Richard P (1982). “Simulating physics with computers”. In: *International journal of theoretical physics* 21.6/7, pp. 467–488.






Jaynes, E. T. (May 1957). “Information Theory and Statistical Mechanics”. In: *Phys. Rev.* 106 (4), pp. 620–630. DOI: 10.1103/PhysRev.106.620. URL:

<https://link.aps.org/doi/10.1103/PhysRev.106.620>.



Kopietz, Peter, Lorenz Bartosch, and Florian Schütz (2010). “Mean-Field Theory and the Gaussian Approximation”. In: *Introduction to the Functional Renormalization Group*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 23–52. ISBN: 978-3-642-05094-7. DOI: 10.1007/978-3-642-05094-7_2. URL: https://doi.org/10.1007/978-3-642-05094-7_2.

References II

- 
 Poole, Ben et al. (2016). *Exponential expressivity in deep neural networks through transient chaos*. [arXiv: 1606.05340 \[stat.ML\]](#).
- 
 Schoenholz, Samuel S. et al. (2017). *Deep Information Propagation*. [arXiv: 1611.01232 \[stat.ML\]](#).
- 
 Shannon, C. E. (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423. DOI: [10.1002/j.1538-7305.1948.tb01338.x](#).