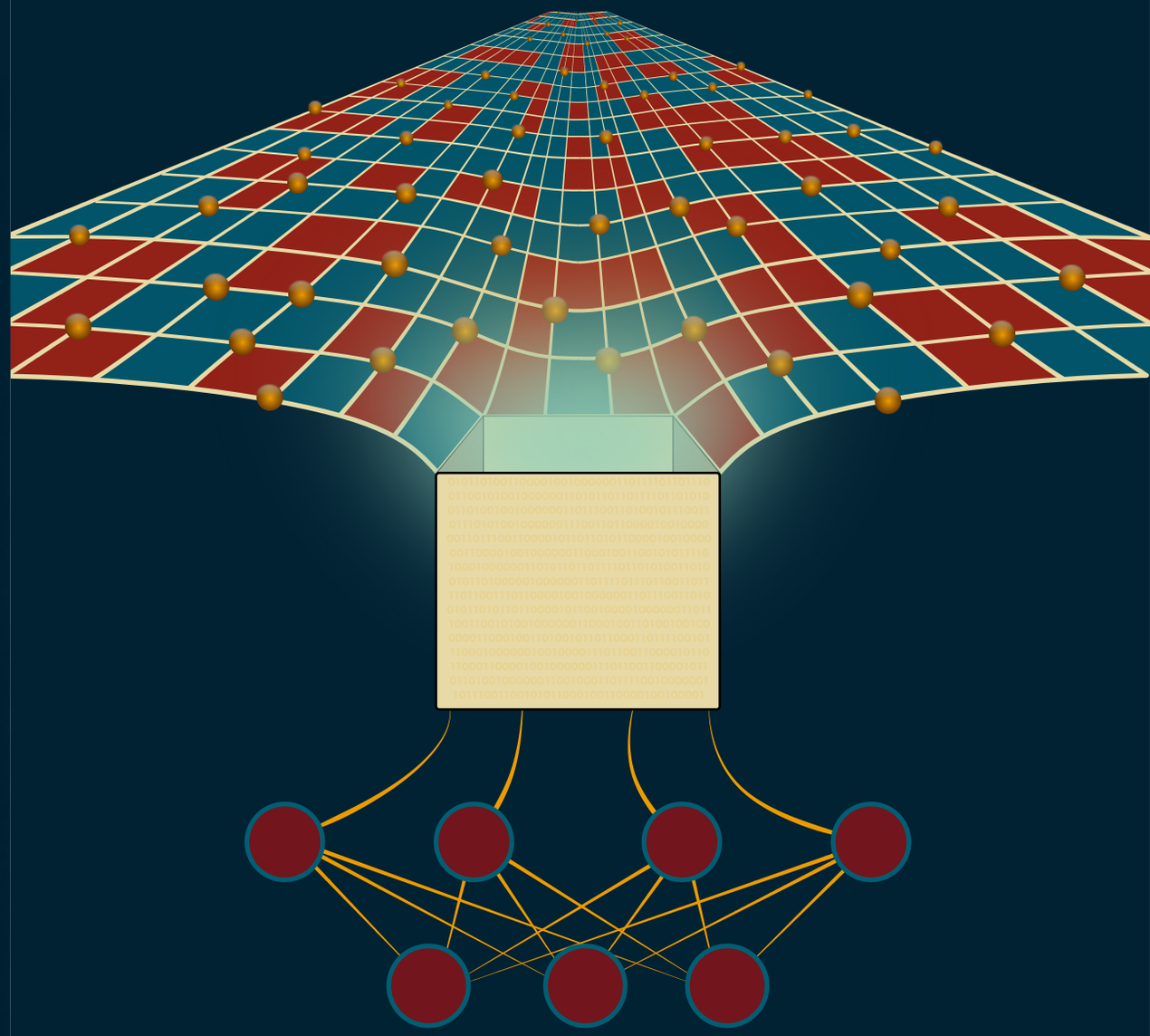
Casimir PhD Series - 2023-23

A. Bukva Numerical Exploration of Statistical Physics

Numerical Exploration of Statistical Physics



Aleksandar Bukva

Numerical Exploration of Statistical Physics

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 10 oktober 2023
klokke 13:45 uur

door

Aleksandar Bukva
Geboren te Kikinda, Servië
in 1994

Promotores:
Prof. dr. J. Zaanen
Prof. dr. K.E. Schalm

Promotiecommissie:
Prof.dr. C.W.J. Beenakker
Dr. J. Tura Brugués
Dr. J. van Wezel
Prof.dr. A. Achúcarro
Prof.dr. J.S. Caux

Casimir PhD series, Delft-Leiden 2023-23
ISBN 978-90-8593-571-1

An electronic version of this Thesis can be found at
https://openaccess.leidenuniv.nl

The cover was created by me, representing the synergy of the whole thesis in a single image.

To myself and my family.

Contents

1

Introduction

One of the recurring motifs throughout this thesis is how recent advances in computer hardware and reduced cost of computations helped us push the boundaries of physical knowledge. Our exploration starts with one of the most widely used numerical tools in any computational discipline, the Monte Carlo (MC) method. Because of the use of random numbers, these methods are believed to carry the name after the most famous casino in the world, Monte Carlo Casino in Monaco. The term *Monte Carlo* was first used in modern literature by S. Ulam, E. Fermi, J. von Neumann, and N. Metropolis while they were working on the Manhattan Project in Los Alamos National Laboratory during the second world war. MC methods will also play a significant role in this thesis as one of the primary tools for analyzing physical systems. Our journey will start with an exploration of different realizations of gauge lattice models; we will introduce simple Wegner Ising gauge theory, the fundamental block, and later build upon it by adding matter fields to the lattice. MC played a significant role in the early days of lattice gauge theories helping us advance our understanding of the theory of elementary particles. We will use it similarly to explore phase diagrams and phase transitions in our lattice gauge models. On the next stop, we will explore how closed, unitary quantum systems can (appear to) thermalize, a question that has been at the heart of statistical physics for a long time. The final stop will be one governed by the recent advances of machine learning in areas of physics. We will try to quantify the multipartite entanglement in quantum systems by combining MC simulations and newly proposed variational wave functions in the form of restricted Boltzmann machines (RBM). Some of the problems, or quirks, of MC methods can also appear in the modern training of neural networks. Most prominent is the choice of initial distribution for parameters we are trying to optimize.

Using statistical physics and field theory tools, we will find optimal initial conditions to improve neural networks' training speed and accuracy.

1.1. Statistical physics

The primary goal of statistical physics is an exploration of macroscopic quantities and the calculation thereof. Often, the systems we explore are made up of many degrees of freedom, and solving them exactly is impossible. In order to do this, we will assume that the statistical average over all possible states can replace the time average.

One of the main assumptions we make when resorting to statistical calculations instead of fully dynamically solving the system is the principle of *ergodicity*. Ergodicity states that if the system is left to evolve, all accessible states will eventually be realized. This assumption helps us often turn insolvable time integrals into relatively easy and, more importantly, simulation-friendly integrals over the probability distributions of those states. For example, let us say we want to study some volume of gas in a container. At standard temperature and pressure, one liter of oxygen contains around $3 \cdot 10^{22}$ oxygen molecules moving around the container. Just writing down equations of motion for all molecules would take a very long time, but no practical conclusion can be drawn even if we manage to do it. Hence we turn to the methods of statistical physics.

These statistical integrals are averaging over many system realizations while keeping certain parameters fixed. These sets of systems with specific parameters fixed are called *ensembles*. For example, a glass of water sitting on a counter at a fixed temperature will have many different configurations of water molecules that fluctuate while having a fixed *average* energy.

There are several well-known ensembles according to corresponding fixed quantities. If the fixed quantity is energy, the ensemble is called *microcanonical*, we can denote the total number of states with that energy as $\mathcal{N}(E)$, the probability that any of those instances are realized is $\frac{1}{\mathcal{N}(E)}$ and the probability that system is in some other energy $E' \neq E$ is simply zero.

Now we can define the *entropy* of a system as:

$$S(E) = k_B \log \mathcal{N}(E) \,. \tag{1.1}$$

Because the number of states in a system of N bits/particles is of order $\mathcal{N}(E) \sim 2^N$, it follows that the entropy is proportional to a number of particles in the system $S(E) \sim N$. We can see that this makes entropy an extensive quantity. For two non-interacting systems at energies E_1 and E_2, we get the total number of states in both systems as:

$$\mathcal{N}(E_1, E_2) = \mathcal{N}(E_1)\mathcal{N}(E_2) \,,$$

and the entropy will be:

$$S(E_1, E_2) = S(E_1) + S(E_2) \,.$$

1

Often systems are not isolated but actually in contact with some reservoir at a specific temperature T. One way we can remedy this is by including the reservoir in our calculation and using a microcanonical ensemble, but doing this results in very complex calculations and, more importantly, properties of the reservoir that are often not of interest to us. Because of this, we will introduce an ensemble at a fixed temperature called *canonical*. Now the system can be in states with the different energy that will fluctuate around some well-defined average value. The average energy of the system is defined as a sum of energies of all the states weighted with their respective probabilities:

$$\langle E \rangle = \sum_i p_i E_i \, . \tag{1.2}$$

Here p_i is a probability that the system is in the state with energy E_i. One can derive, either by entropy maximization or placing the system in contact with a heat bath, that a system with only energy as conserved quantity has a probability, also known as *Boltzmann distribution*:

$$p_i = \frac{\exp\left(-\beta E_i\right)}{\sum_i \exp\left(-\beta E_i\right)} \, . \tag{1.3}$$

We have introduced a factor, inverse temperature, $\beta = \frac{1}{k_B T}$ where k_B is Boltzman constant = $1.380649 \cdot 10^{-23} \mathrm{J} \cdot \mathrm{K}^{-1}$, and T temperature of the ensemble. Exponential factors ensure that states with energies $E_i \gg k_B T$ get suppressed while states with $E_i \leq k_B T$ have a chance of being populated. We can check that this makes sense. Let $T \to 0$, and the system is forced to a ground state energy with all higher states being suppressed. We can also define the entropy for the canonical ensemble as follows:

$$S = -k_B \sum_i p_i \log p_i \, . \tag{1.4}$$

The normalization factor is an essential quantity in Eq.(1.3). This sum over all accessible states that ensure that probability sums to 1 is called *partition function*:

$$Z = \sum_i \exp\left(-\beta E_i\right) . \tag{1.5}$$

Once we know the partition function, we can compute any quantity about the system; we have complete knowledge. Let us explore what information we can extract from the partition function. We can define the free energy at fixed temperature as:

$$F = \langle E \rangle - TS \, , \tag{1.6}$$

where T is temperature and S is entropy. We can think of free energy as energy available to the system, an intricate interplay between internal energy and entropy. At the fixed temperature, we can get free energy directly from the partition function:

$$F = -k_B T \log Z \, . \tag{1.7}$$

To connect average energy and a partition function, we can plug Eq.(1.3) into Eq.(1.2):

$$\langle E\rangle = \sum_i p_i E_i = \sum_i \frac{E_i \exp\left(-\beta E_i\right)}{Z} .$$

However, this can be rewritten in the form of a partial derivative of the partition function:

$$\langle E\rangle = \frac{\partial}{\partial \beta} \log Z . \tag{1.8}$$

As stated earlier, energy is not constant in the canonical ensemble but fluctuates. The exponential factor in Boltzmann distribution ensures that these fluctuations are around $k_B T$; we can compute them as a variance of energy:

$$\sigma_E^2 = \langle (E - \langle E\rangle)^2\rangle = \langle E^2\rangle - \langle E\rangle^2 = \frac{\partial^2}{\partial \beta^2} \log Z = -\frac{\partial \langle E\rangle}{\partial \beta} .$$

We can make a connection with specific heat that will give us a different interpretation of these fluctuations. Specific heat is a thermodynamic quantity defined as $C = \frac{\partial E}{\partial T}$ or how much the energy of a system changes as we vary the temperature. Specific heat can be measured while we keep other parameters fixed, like volume or pressure. So, for example, specific heat at fixed volume will be $C_V = \left.\frac{\partial E}{\partial T}\right|_{V=\text{const}}$. Now we can connect specific heat and the variance of energy:

$$\sigma_E^2 = k_B T^2 C_V .$$

Interesting conclusions can be drawn from here; if C_V of the system is large, it can easily absorb bigger energy fluctuations without changing its temperature. Another conclusion is that for macroscopic systems ($N \gg 1$) away from the critical point, $E \propto N$ and $C_V \propto N$, so we have that energy fluctuations scale as:

$$\frac{\sigma_E}{E} \sim \frac{1}{\sqrt{N}} . \tag{1.9}$$

In the limit, when $N \to \infty$, the energy becomes exactly $\langle E\rangle$. This is also known as a *thermodynamic limit*. These fluctuations are too small for real systems to be detected, but they are often how we calculate thermodynamic quantities in simulations.

The question now is what happens with the fluctuations of other quantities. If there are terms in the Hamiltonian $\mathscr{H}$, of form $-XY$, where Y is some field and X conjugate variable to which field Y is coupled then we have:

$$\langle X\rangle = \frac{1}{\beta Z} \frac{\partial}{\partial Y} \sum_i \exp\left(-\beta \left\langle i\middle|\mathscr{H}\middle|i\right\rangle\right) = -\frac{\partial F}{\partial Y} . \tag{1.10}$$

Now $\left\langle i\middle|\mathscr{H}\middle|i\right\rangle$ contains the term of form $-XY$ on which partial derivative acts. So if we desire to find an average value of any conjugate variable, we need to differentiate

free energy with respect to the appropriate coupled field. If there is no term coupling field to desired quantity, we can add it, find the derivative, and later set the same field to zero, restoring an original Hamiltonian. Doing further differentiation, we can see that:

$$-\frac{1}{\beta}\frac{\partial^2 F}{\partial Y^2} = \frac{1}{\beta}\frac{\partial \langle X\rangle}{\partial Y} = \langle X^2\rangle - \langle X\rangle^2 . \tag{1.11}$$

We used this variance to connect energy fluctuations to the specific heat. Now we have a general way of finding these variances from the second derivatives of free energy with respect to coupled fields. The derivative in Eq.(1.11) measures the magnitude of the response of X to changes happening in Y. This is called *susceptibility* of X to Y and is usually written as χ. This gives us a general way of calculating susceptibility directly from the simulations by measuring fluctuations in desired quantities. A practical example of susceptibility that will be important to us is the magnetic susceptibility or how total magnetization of the system changes as we vary the external magnetic field.

Some systems can go through a *phase transition*, a sudden change in a system's specific property, becoming discontinuous in the large N limit. The prominent feature of a phase transition is the appearance of a non-vanishing value of *order parameter*, some quantity that is non-zero in the ordered phase and then identically zero in the disordered phase. Different order parameters are identified in different systems; for example, in a ferromagnet, it is a spontaneous magnetization, and in the liquid-gas phase, it is the difference in the density between the liquid and gas phases at the transition line. They can be scalar or multi-component quantities as well.

The classification of phase transitions can be done by the order of discontinuous derivative of a free energy. If the first derivative of the free energy shows discontinuous behavior, we call those transitions *first order*. If, on the other hand, the first order is continuous but the second derivative is not, we label those transitions *second order*.

We can deduce some qualitative characteristics of a model using *Landau theory of phase transitions*. The main idea revolves around expanding the free energy in the powers of the order parameter. For this expansion to converge, we have to keep the order parameter small, so near the critical point. The Ising model in an external magnetic field h is a popular toy system to illustrate the main concepts. The energy of the Ising model can be expressed as:

$$E = -J\sum_{\langle ij\rangle} s_i s_j - h\sum_i s_i , \tag{1.12}$$

where spins $s_i \in \{-1, 1\}$ and we sum over all neighboring sites $\langle ij\rangle$ in a d-dimensional hyper-cubic lattice.

Let us start with the description of a second-order phase transition. Consider a gen-

eral model (like the Ising case defined above) and denote an order parameter m (in our case, total magnetization). Then we can expand the free energy as:

$$F(T,m) = F_0(T) + a(T)m^2 + b(T)m^4 + \cdots . \tag{1.13}$$

We do not have odd terms because the theory is invariant under the change $m \to -m$, which forbids odd terms in the free energy (in our case, when $h = 0$ Ising has this symmetry). The system's state can be found by extremizing the free energy, $\frac{\partial F}{\partial m} = 0$. We can immediately see that result will depend on the signs of the temperature-dependent terms. These terms can change sign with varying temperature as well. For simplicity, we will assume that $b(T) > 0$ for all T. If we do not impose this condition and let $b(T)$ have any sign, then we need to include higher terms of expansion (m^6 in our case) which can lead to the formation of tri-critical points, a phenomenon explored in the later chapters. In order to illustrate this further, we will use the free energy of the Ising model that we can calculate from mean-field theory[1]:

$$F = -\frac{1}{\beta}\log Z = \frac{1}{2}JNqm^2 - \frac{N}{\beta}\log\left(2\cosh\beta h_{\text{eff}}\right), \tag{1.14}$$

where $q = 2d$ is the number of nearest neighbors and $h_{\text{eff}} = h + Jqm$ is an effective (mean-field) magnetic field. We can expand this mean-field expression in order to compare coefficients with our free energy expansion:

$$F(T,m) = -Nk_BT\log 2 + \left(\frac{NJq}{2}(1 - Jq\beta)\right)m^2 + \left(\frac{N\beta^3J^4q^4}{12}\right)m^4 + \cdots . \tag{1.15}$$

The first term is just an additive constant, which will vanish once we take a derivative with respect to m. Consider then the quadratic term with coefficient $a(T) = \frac{NJq}{2}(1 - Jq\beta)$ there are two different cases $a(T) > 0$ and $a(T) < 0$. From the expression of $a(T) = 0$, we can also determine the critical temperature $T_c = \frac{Jq}{k_B}$. In the case of $a(T) > 0; T > T_c$, and thus when temperatures are high, we see that the only equilibrium solution is for $m = 0$. On the other hand, when $T < T_c$, for low temperature, there are three solutions, $0, \pm\sqrt{-\frac{a(T)}{2b(T)}}$. Substituting these values in Eq.(1.15), we see that the free energy of a state with $m = 0$ is higher than the other two and represents an unstable solution, Fig.(1.1). We thus find that the free energy equals:

$$F(T) = \begin{cases} F_0(T) & T > T_c \\ F_0(T) - \frac{a(T)^2}{4b(T)^2} & T < T_c \end{cases} . \tag{1.16}$$

The important point is that if $a(T)$ is a smooth function, then the equilibrium value of m also changes continuously from $m = 0$ in the $T > T_c$ regime to a $m \neq 0$ in a $T < T_c$ regime. We can also see that the free energy is continuous at the transition. However, suppose we differentiate free energy twice to get a specific heat. In that case, we will have a term $\frac{a(T)'^2}{b(T)}$ which is usually not equal to zero, and hence specific

1

heat has a discontinuous change at $T = T_c$, defining characteristics of a second order phase transition.

When we were expanding the free energy in Eq.(1.13), we said that odd terms are forbidden by the symmetry $m \to -m$, so-called $\mathbf{Z}_2$ symmetry. Then when we found the equilibrium values of m for $T < T_c$, we saw that the system had to choose one ground state of the two, $\pm\sqrt{-\frac{a(T)}{2b(T)}}$, this is the well-known phenomena of *spontaneous symmetry breaking*. The Ising model discussion above illustrates how it is tied to the second-order phase transition.

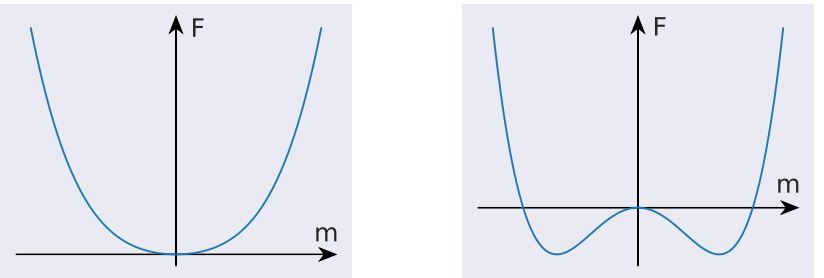


Figure 1.1: Free energy curves in the case of second-order phase transition. In the left graph, when $a(T) > 0$, we see only one equilibrium solution at $m = 0$. On the other hand, when $a(T) < 0$ (right graph), two more solutions appear. The previously stable solution at $m = 0$ now becomes unstable. We can see that the system can choose which stable solution to be in, which is what we call spontaneous symmetry breaking.

Let us briefly explore what happens when we include odd power terms of the order parameter in our expansion of the free energy F.

$$F(T,m) = F_0(T) + \alpha(T)m + a(T)m^2 + \gamma(T)m^3 + b(T)m^4 + \cdots . \tag{1.17}$$

If we look back at our definition of Ising model Eq.(1.12) we see that for $h \neq 0$ the system doesn't have $\mathbf{Z}_2$ symmetry in $m \to -m$, so we can use the expansion above:

$$F(T,m) = -Nk_BT\log 2 + \frac{JNq}{2}m^2 - \frac{N}{2k_BT}(B+Jqm)^2 + \frac{N}{24(k_BT)^3}(B+Jqm)^4 + \cdots . \tag{1.18}$$

We will use the same assumption again that $b(T) > 0$ for all T. In the regime where $T < T_c$, we again have three solutions, but now the curve is skewed, and more importantly, two minima are no longer degenerate. An introduction of the external magnetic field h broke the degeneracy that was initially there, and the other solution is called a meta-stable state. When odd terms in the expression for the free energy, $\alpha(T)$ and $\gamma(T)$, change signs, the curve goes from the left to the right picture in Fig.(1.2). This changes the true ground state, and the transition from this new meta-stable state to the true ground state with lower energy is the first-order phase transition. In the example of the Ising model, this transition occurs when the external magnetic

field h changes sign. If we increase the temperature above the critical $T > T_c$, the free energy curve looks similar to the second-order transition. However, it is shifted because of the linear term $\alpha(T)$.

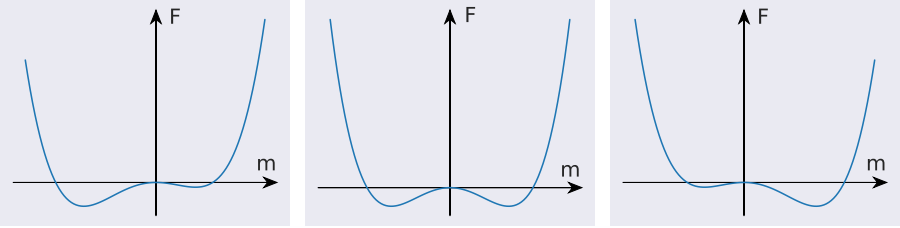


Figure 1.2: Curves of free energy in case of a first-order phase transition. Again we have three solutions, but only one is the true ground state and has lower energy than the other. A state with higher energy is a meta-stable state. Curves change shape as we change the sign of $\alpha(T)$ and $\gamma(T)$, corresponding to a phase order transition.

The toy model presented here can be solved exactly using the mean-field technique. Nevertheless, the question remains: How to compute the free energy for more complicated systems where a direct analytical approach does not work? One solution is to do it numerically or more precisely using the statistical techniques of Monte Carlo simulations that we will explore in the next chapter.

1

1.2. Monte Carlo techniques

The Monte Carlo methods are probably one of our most important numerical methods. The main idea is to obtain the approximate value of an integral by randomly sampling the value of the integrand at the expense of introducing statistical errors. Traditional grid methods that subdivide total volume become increasingly computationally expensive as we try to evaluate integrals in higher dimensions. For large systems, the number of configurations i used in Eq.(1.2) is the integral of primary interest to physicists. The expectation value of some quantity $\langle \xi \rangle$, like the total energy of a system or a magnetization, is computed by summing over all possible states weighted by their respective probabilities.

$$\langle \xi \rangle = \frac{\sum_\alpha \xi_\alpha e^{-\beta E_\alpha}}{\sum_\alpha e^{-\beta E_\alpha}} , \tag{1.19}$$

where $\beta = 1/kT$ and E_i energy of a state. Sums like these are only feasible for small systems and generally cannot be computed exactly. One possible solution that we can take is to evaluate this sum only on a small appropriate subset of states and use this as an estimate. Suppose we take L states $\{\alpha_1, \ldots, \alpha_L\}$, then our estimate will be given by:

$$\xi_L = \frac{\sum_{i=1}^{L} \xi_{\alpha_i} p_{\alpha_i}^{-1} e^{-\beta E_{\alpha_i}}}{\sum_{j=1}^{L} p_{\alpha_j}^{-1} e^{-\beta E_{\alpha_j}}} . \tag{1.20}$$

We call ξ_L the *estimator* of ξ. It is clear from the procedure that as we take more samples, the estimator becomes more accurate, and in the limit $\lim_{L\to\infty} \xi_L = \langle \xi \rangle$.

The question then is how to choose p_α. The most straightforward choice is to take all the states with an equal probability. Immediately we can see that for specific cases, this will be a terrible choice. Imagine evaluating an integral that is highly peaked around a particular value. Choosing points with equal probability will sample parts of the integral that do not contribute much toward the final result. Another way we can look at this from the physics point of view is when evaluating these sums at a very low temperature. Usually, only a handful of states will effectively contribute to the sum, so sampling over all (improbable) states will be a waste of computational time. The technique for selecting these most appropriate states is called *importance sampling*.

Return to a physics example; we know that at a specific temperature, only the states in a small energy window will have measurable contributions to the sum, while an exponential factor will suppress the others. The natural choice for the probability distribution then presents itself; we will choose states according to their Boltzmann distribution. Looking at the equation Eq.(1.20), we can see that if we pick the states according to their Boltzmann weight $p_\alpha = Z^{-1} e^{-\beta E_\alpha}$, this expression simplifies con-

siderably:

$$\xi_L = \frac{1}{L}\sum_{i=1}^{L} \xi_{\alpha_i} \,. \tag{1.21}$$

We see that all the factors have canceled out. We are left with a simple sum, or more precisely, in order to get an estimate of a quantity ξ, we are going to sample states according to their Boltzmann distribution, measure ξ in those states and then find an average over all of the measurements. When we settle on the desired distribution, the question remains how to realize it. One of the most widely used methods to achieve the desired distribution is through a *Markov process*.

In order to understand Markov processes, let us define a stochastic process at discrete times $t_1, t_2, t_3, \ldots$ and a system with a finite set of states $\alpha_1, \alpha_2, \alpha_3, \ldots$. Let us label with X_t the state of a system at time t. The defining property of a Markov process is that:

$$P(X_{t_n} = \alpha_n | X_{t_{n-1}} = \alpha_{n-1}, X_{t_{n-2}} = \alpha_{n-2}, \ldots X_{t_1} = \alpha_1) = P(X_{t_n} = \alpha_n | X_{t_{n-1}} = \alpha_{n-1}) \,, \tag{1.22}$$

or in words, Markov process is *memory-less*, the state of a system in the next time step only depends on the system's current state. Probability of generating state α_n given that system is in α_{n-1} is called *transition probability* $P(\alpha_{n-1} \to \alpha_n)$. Another condition that these probabilities should satisfy is that they do not change over time. As with all probabilities, these should satisfy the fundamental requirement:

$$\sum_{\alpha_n} P(\alpha_{n-1} \to \alpha_n) = 1 \,, \tag{1.23}$$

this means the process must generate some state, even the one the system already is in. We will end up with a *Markov chain* of states by repeatedly generating new states. Suppose transition probabilities are chosen correctly after a sufficiently long time. In that case, the chain will come to an equilibrium, and its new states will satisfy the desired Boltzmann distribution.

With chosen correctly, the following is meant. For the Markov process to achieve the desired equilibrium distribution, we must satisfy two additional conditions, *ergodicity* and *detailed balance*. Ergodicity, as we stated before, means that the Markov process should be able to reach any valid state from any valid state of the system if we let it run long enough. We know that every state in the Boltzmann distribution has a non-zero probability of appearing, so if this condition is violated, that would mean that there would be a pair of states α and γ such that if we start from a state α we would never be able to reach state γ and our goal of reaching the Boltzmann distribution would not be achievable. This does not mean that all transition probabilities should be non-zero, just that there should always be a path from two states that the process can follow.

The second condition that we need to satisfy is the one of a detailed balance. This condition ensures that the final distribution we have reached is the desired Boltzmann distribution. What this exactly means is that the process has reached an equilibrium state. Let us take not a single state α, but an ensemble with distribution p_α. Then this ensemble is in equilibrium if the rate of states transforming into any state α is the same as the rate of states transforming from α:

$$\sum_\beta p_\alpha P(\alpha \to \beta) = \sum_\beta p_\beta P(\beta \to \alpha) \,. \tag{1.24}$$

Then using Eq.(1.23) we get:

$$p_\alpha = \sum_\beta p_\beta P(\beta \to \alpha) \,. \tag{1.25}$$

If our transition probabilities satisfy this condition, the probability distribution p_α will be the equilibrium distribution. Only imposing the condition in Eq.(1.24) is not enough to guarantee that the probability distribution generated from any state will eventually settle to the desired distribution p_α. In order to understand this, let us consider transition probabilities $P(\beta \to \alpha)$. We can form a large matrix where each entry in column α and row β would be a transition probability from α to β; we call this a Markov matrix. If we label the probability that our system is in a state α at time t with $p_\alpha(t)$, then we can say that a probability that the system is in state β in a subsequent time step is:

$$p_\beta(t+1) = \sum_\alpha P(\alpha \to \beta) p_\alpha(t) \,,$$

or in a more compact, matrix notation:

$$\boldsymbol{p}(t+1) = \boldsymbol{P} \cdot \boldsymbol{p}(t) \,.$$

Now if we let this run for some time at the certain point the Markov chain will reach an equilibrium and we will have:

$$\boldsymbol{p}(\infty) = \boldsymbol{P} \cdot \boldsymbol{p}(\infty) \,.$$

which is just a form of standard eigenvector equation. However, it is also possible for the Markov process to reach *dynamic equilibrium*, which means that $\boldsymbol{p}$ rotates around several different values. Such rotation is called a *limit cycle*. If this were the case, then the eigen equation would be:

$$\boldsymbol{p}(\infty) = \boldsymbol{P}^n \cdot \boldsymbol{p}(\infty) \,,$$

where we call n the length of the limit cycle. So how can we ensure limit cycles do not happen? The easiest way is to impose the condition of *detailed balance*:

$$p_\alpha P(\alpha \to \beta) = p_\beta P(\beta \to \alpha) \,. \tag{1.26}$$

We can trivially see that if this condition is satisfied, then Eq.(1.24) is also satisfied, but more importantly, we are eliminating the possibility of cycles. In order to understand why, look at the left side of Eq.(1.26); this is an overall rate at which transition from α to β happens, while the right side is the reverse. So this tells us that, on average, the system goes from $\alpha \to \beta$ as often as $\beta \to \alpha$. This would inherently be violated in the limit cycle as the states need to go from one to the other in a cyclic way. Another way we can justify the detailed balance condition is that most physical systems satisfy it in some way. The majority of the physical systems exhibit time-reversal symmetry. If we had a cycle in such a system upon reversing the arrow of time, the cycle would also reverse its direction. However, this reversal of cycle direction would completely change the system's dynamics, and any already established equilibrium would not be the same. In order to represent physical systems without this behavior, we would also like to remove it from our models.

We have established how to generate the desired probability, and from the detailed balance, our transition probabilities should satisfy the following:

$$\frac{P(\alpha \to \beta)}{P(\beta \to \alpha)} = \frac{p_\beta}{p_\alpha} = e^{-\beta(E_\beta - E_\alpha)} . \tag{1.27}$$

Here in the last step, we substituted the Boltzmann distribution. If we choose transition probabilities that satisfy this and the ergodicity condition, our Markov process should converge to the Boltzmann distribution. The procedure described so far looks straightforward; we need to choose the correct transition probability $P(\alpha \to \beta)$, and we are set. This is not always such an easy task, and there can be many different ways of creating a state β from α that still, in the end, does not produce the desired distribution. There is a trick that can help us construct the desired process using any algorithm that we come up with, called the *acceptance ratio*.

The main observation that lies at the heart of the acceptance ratio is that we can modify the probability of a state staying at the same state as much as we want (still with the constraint that probability should be positive and less than one) and automatically satisfy Eq.(1.27). This allows us to tune our transition probabilities as we like and then, by adjusting the "staying" probability, satisfy the required condition. Let us see how we can do this. First, we will break the transition probability into two parts:

$$P(\alpha \to \beta) = g(\alpha \to \beta) A(\alpha \to \beta) .$$

The first quantity on the right side, $g(\alpha \to \beta)$, is called the *selection probability*, the probability given an initial state α that we will end up in the state β. The second quantity, $A(\alpha \to \beta)$, is the acceptance ratio, the probability that we accept the generated state. Now if we expand the Eq.(1.27) in terms of selection probability and acceptance ratio, we will get:

$$\frac{P(\alpha \to \beta)}{P(\beta \to \alpha)} = \frac{g(\alpha \to \beta) A(\alpha \to \beta)}{g(\beta \to \alpha) A(\beta \to \alpha)} . \tag{1.28}$$

From the equation above, we see that the ratio $\frac{A(\alpha\to\beta)}{A(\beta\to\alpha)}$ can take any value between 0 and ∞. This means that selection probabilities can be tuned to anything we like, and the acceptance ratio will absorb any difference. Partitioning transition probabilities into acceptance ratio and selection probability enables us to create a Monte Carlo algorithm that will generate random state ν given state μ and then adjust selection probabilities for these states so they satisfy the condition in Eq.(1.28) and reach the desired Boltzmann distribution. So we finally reached a theoretically complete algorithm to generate the desired Boltzmann distribution.

Algorithm 1 Markov process algorithm

1: design an algorithm that will generate a random new state β given that we are at α with some probabilities $g(\alpha \to \beta)$
2: accept that state with the probability $A(\alpha \to \beta)$
3: adjust acceptance ratios in such a way that we satisfy Eq.(1.28)
4: after reaching it's equilibrium state, Markov process will have the desired Boltzmann probability distribution.

The algorithm outlined above looks nice and easy, but we still need to pay attention to how we choose our acceptance ratios. If they are small, we will usually stay in the same state and not move at all. We need to make a delicate balance so that our algorithm explores the phase space without slowly crawling around it. This can be ensured by making the acceptance ratio close to one. The good thing is that Eq.(1.28) only fixes ratio $\frac{A(\alpha\to\beta)}{A(\beta\to\alpha)}$, which means that we have the freedom to multiply both of them with some constant. In practice, this often means we fix the larger to one and then multiply the other with an appropriate constant to keep the ratio fixed. Another thing that we can try to do while designing a new algorithm is to put as much of $P(\alpha \to \beta)$ into the selection probability because the perfect algorithm would be the one where the states are only selected by the transition probability, which means acceptance ratio is always one. In the absence of a perfect, we can try to keep our acceptance ratio as close to one as possible.

We have covered the theoretical part of designing an algorithm to ensure that the resulting chain will have the Boltzmann distribution. Let us look into one of the most common algorithms, *Metropolis algorithm*.

In order to show all the details of how this algorithm works, we will again use the Ising model, where "spins" s_i live on the sites of a lattice and can take values $\{-1,+1\}$. If we have d dimensional lattice of size N, our system can be in 2^{N^d} possible states. These numbers grow exponentially as we go to higher dimensions and bigger lattices. This type of problem is the perfect playground for Monte Carlo methods. To recall from

Eq.(1.12) the Ising Hamiltonian is:

$$E = -J \sum_{\langle i\, j \rangle} s_i s_j - h \sum_i s_i \,, \tag{1.29}$$

where J is the interaction energy between the nearest neighbors and h is the external magnetic field. The most common questions regard the values of magnetization m and specific heat C_V at the fixed temperature, which we computed earlier in the mean-field limit. For simplicity, we will consider the Ising model with no external magnetic field $h = 0$

As we showed in the section 1.1 in the thermal equilibrium, energies of the system stay in a tiny window around the mean value; they do not fluctuate much. We want to replicate this behavior in our algorithm as well. The easiest way we can do this is by flipping a single spin at a time. This type of algorithm has a *single spin-flip dynamics*. In the d dimensional lattice, the maximum energy difference would be $2zJ$, where z is *lattice coordination number*, or simply the number of neighbors a site has. Another advantage of flipping a single spin is that we also satisfy the ergodicity condition.

The Metropolis algorithm now sets the selection probabilities $g(\alpha \to \beta)$ equal for all possible states β while setting it to zero for all the other ones. Let us say that we have N^d spins in our system; that means there are N^d different spins that we could flip and also N^d different states that we could end up in. This means that we have N^d different selection probabilities $g(\alpha \to \beta)$ that are non-zero, and all of them have the same value:

$$g(\alpha \to \beta) = \frac{1}{N^d} \,.$$

In practice, we randomly pick a site with equal probability and flip its spin. Now our condition of the detailed balance (1.28) is:

$$\frac{P(\alpha \to \beta)}{P(\beta \to \alpha)} = \frac{g(\alpha \to \beta) A(\alpha \to \beta)}{g(\beta \to \alpha) A(\beta \to \alpha)} = \frac{A(\alpha \to \beta)}{A(\beta \to \alpha)} = e^{-\beta(E_\beta - E_\alpha)} \,. \tag{1.30}$$

If we remember the algorithm design steps, we will set the largest of the acceptance ratios to 1 and adjust the other one to make this the most efficient algorithm. Assume that state α has the lower energy than β, $E_\alpha \le E_\beta$. Then the larger acceptance ratio would be where we go from the state β to the state α, $A(\beta \to \alpha)$, we will set it to 1. In order to satisfy Eq.(1.30) we must then set $A(\alpha \to \beta) = e^{-\beta(E_\beta - E_\alpha)}$. Now our final algorithm is:

$$A(\alpha \to \beta) = \begin{cases} e^{-\beta(E_\beta - E_\alpha)} & \text{if } E_\beta - E_\alpha \ge 0 \\ 1 & \text{otherwise} \end{cases} \,. \tag{1.31}$$

Put in words, this means that if we select a state with lower energy, we always accept the move, and if we select the state with higher energy, we will accept it with the probability $e^{-\beta(E_\beta - E_\alpha)}$.

1

So far, we only said that we must wait long before our system reaches equilibrium. Now we will specify what this exactly means. First, let us define this period. The time we need to wait for the system to reach the thermal equilibrium is called the *thermalization time*, τ_T. How do we know how long is enough? We could look at the state of a system as in the Fig.(1.3) and then gauge when it has reached the equilibrium, but the better and simpler way to do this is to plot some quantity, like energy or magnetization, over "time" and see when it settles around the average value. The usual way to measure time in Monte Carlo simulations is in terms of *sweeps*. Sweep is the number of updates we do to flip all the spins on the lattice. So if we sequentially try to flip spins in the lattice after N^d attempts, we will complete one sweep. In Fig.(1.4), we see that after 300 sweeps, the average magnetization and energy have thermalized. However, in practice, this is not always enough. Sometimes if the energy landscape is particularly rough, we could end up in one of the *local minima* and oscillate around it. In order to avoid this, we can start our simulation with different initial conditions, like at $T = 0$ (all spins aligned) and $T = \infty$ (completely randomly oriented spins), and let them run. When two runs settle around the same average value, we have good reason to believe that thermalization has completed and we have reached equilibrium Fig.(1.5). If we are still worried that we might have ended up in some local minimum, we can do a third run with a different randomly oriented configuration and repeat the same procedure.

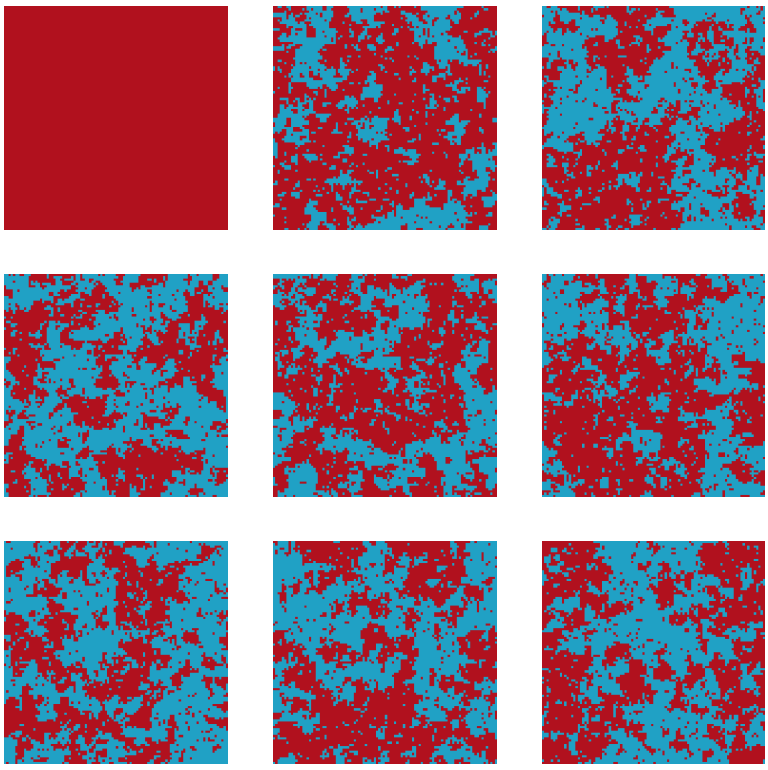


Figure 1.3: Time slices of the spin configuration on a 100 × 100 lattice throughout the simulation. Red are spins −1, and blue spins are +1. We have started at $T = 0$ configuration with all the spins aligned and then simulated $J = 1$ and $T = 2.4$ for 1000 sweeps. Slices were taken every 100 sweeps. The graph shows that the system reached equilibrium configuration around the 400th sweep.

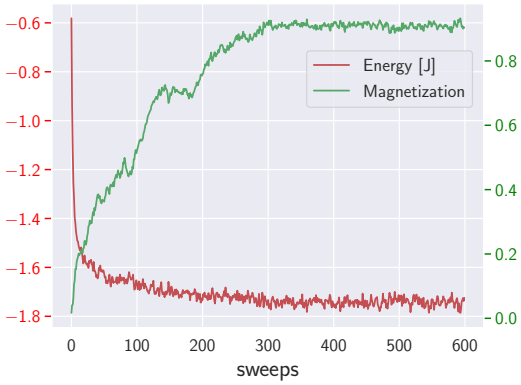


Figure 1.4: Graph of magnetization per site (green squares) and energy per site (red circles) for a 2D Ising model with no external magnetic field ($h = 0$) on a lattice $N = 100$, $J = 1$ near the critical temperature ($T_c \sim 2.26$) $T = 2.0$. We can see that energy and magnetization reach their thermal values around 300-400 sweeps.

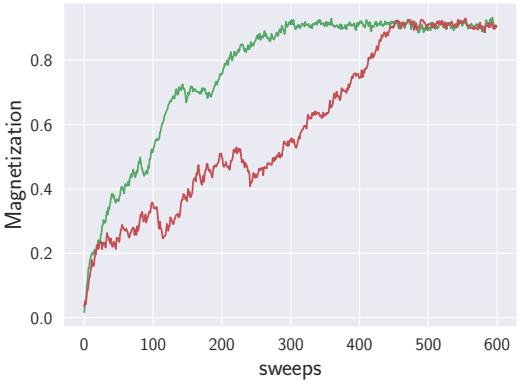


Figure 1.5: Graph of magnetization per site for two 2D Ising models with different initial configurations (at $T = \infty$), no external magnetic field ($h = 0$) on a lattice $N = 100$, $J = 1$ and equilibrium temperature $T = 2.0$. We can see that two different initial configurations had different paths through the phase space, but both have settled on the same value for the magnetization per site. Waiting for around 500 sweeps would guarantee that our Monte Carlo chain has thermalized for these values of J and T.

We have established how to get an adequately converged Monte Carlo simulation. The next step is measuring some quantity, such as energy or magnetization. If we recall our algorithm's exact dynamics, we will realize that the single flip dynamics states are actually highly correlated. This implies that two or more successive measurements will be correlated. In order to solve this, we will wait for a certain amount of sweeps between successive measurements, called *correlation time*, τ.

In order to estimate autocorrelation time, we will use *time-displaced autocorrelation*

function. We give here an example of it for magnetization:

$$\chi(t) = \int \mathrm{d}t'(m(t')m(t'+t) - \langle m\rangle^2) \,. \tag{1.32}$$

This is just the two-point function we have seen in the first part of the introduction, but now instead of being between two sites i and j, it is between two time-steps t and t'. One assumption that we have to make in order to make this work is that the autocorrelation function falls off exponentially at long times (which is a very reasonable assumption to make):

$$\chi(t) = e^{-\frac{t}{\tau}} \,. \tag{1.33}$$

We can see from this that waiting time τ between two measurements will only reduce the autocorrelation function by $1/e$, so the most common practice is to take measurements every 2τ sweeps. Given the exponential form of the autocorrelation function Eq.(1.33) we can define what the *integrated correlation time*:

$$\int_0^\infty \frac{\chi(t)}{\chi(0)}\mathrm{d}t = \int_0^\infty e^{-\frac{t}{\tau}}\mathrm{d}t = \tau \,. \tag{1.34}$$

This equation gives us a direct way of estimating the autocorrelation time. To complete the procedure, we need a discrete version of the Eq.(1.32) that we can use in order to estimate the autocorrelation time:

$$\chi(t) = \frac{1}{t_{\max}-t}\sum_{t'=0}^{t_{\max}-t} m(t')m(t'+t) - \left(\frac{1}{t_{\max}-t}\right)^2 \sum_{t'=0}^{t_{\max}-t} m(t') \sum_{t''=0}^{t_{\max}-t} m(t''+t) \,. \tag{1.35}$$

When using this equation to calculate χ, we have to take care that at long times when t gets close to $t_{\max}$, the upper limit in the sums gets very small, and because of the statistical nature of the measurements, errors can get very large. In order to avoid this, we can run a simulation long enough (hopefully for several autocorrelation times), and then we would not need to worry about the long tails of $\chi(t)$.

We have established how to generate new states from the old one, measure quantities of interest and calculate autocorrelation time so our sample end up uncorrelated. The final step is to quantify the errors of those measurements in order to get the complete results. Our errors can be divided into two categories, *systematic errors* and *statistical errors*. We have introduced systematic errors due to the procedures we used to measure observables. Another source of systematic errors can be if we do not wait long enough for the system to thermalize. Also, if we do not collect enough samples after the thermalization and let our samples be correlated. Systematic errors are hard to measure, and we can mitigate them by following the correct procedure for reaching the equilibrium and collecting enough uncorrelated samples after. On the other hand, statistical errors are more easily measured; they arise from the random nature of the Monte Carlo simulations and the system's thermal fluctuations. The simplest way to minimize statistical errors is to collect more uncorrelated samples.

1

Statistical errors of a quantity are relatively easy to estimate. As we stated, due to the thermal fluctuations from one state to the other, measurements (let us take magnetization m) vary around some average value. The natural way is to take the mean of all of these samples as the actual average value and the error of the mean as an error to that estimate. Let us take N measurements of magnetization m; our best estimator for the true average is the mean value:

$$\langle m \rangle = \frac{1}{N} \sum_{i=0}^{N} m_i \,, \tag{1.36}$$

and the estimate of the standard deviation of the mean is:

$$\sigma^2 = \frac{1}{N(N-1)} \sum_{i=0}^{N} (m_i - \langle m \rangle)^2 = \frac{1}{N-1} \left(\left\langle m^2 \right\rangle - \langle m \rangle^2 \right) . \tag{1.37}$$

This might look similar to the susceptibility we defined in section 1.1. There we calculated susceptibility as a variance of the energy over the given ensemble of states, while here, the σ^2 is an unbiased estimator for the true variance of the mean. An essential fact about Eq.(1.37) is that it assumes that our samples are statistically independent. If, on the other hand, we have samples that were sampled every Δt sweeps and the autocorrelation time is τ, then our estimate for the standard deviation is[2]:

$$\sigma^2 = \frac{1 + \frac{2\tau}{\Delta t}}{N-1} \left(\left\langle m^2 \right\rangle - \langle m \rangle^2 \right) . \tag{1.38}$$

We can immediately see that if $\Delta t \gg \tau$, this expression would reduce to the Eq.(1.37), but often, because of practical reasons, we have $\Delta t \ll \tau$. If that is the case, then we can practically ignore the 1 and note that the number of samples is the total amount of time divided by the sampling interval:

$$N = \frac{t_{\max}}{\Delta t} \,.$$

For the large number of samples we have:

$$\sigma^2 = \frac{2\tau}{t_{\max}} \left(\left\langle m^2 \right\rangle - \langle m \rangle^2 \right) . \tag{1.39}$$

A neat property of this estimation is that our standard deviation does not depend on Δt, which means we can sample with any frequency we want.

In some cases, it is not feasible to calculate the errors this way, for example, in any derived quantities that depend on the average values, like specific heat. We only have access to the average values after we finish running and do not have access to specific heat during every time step. In principle, we could do a detailed analysis of error propagation taking into account how $\langle E \rangle$ and $\left\langle E^2 \right\rangle$ errors are correlated, but this would be a complex and error-prone process, also the resulting equation would

be highly dependent on the initial equation we used to calculate the observable. This would mean that every time we have a different form of an equation, we would have to do the same analysis again. Luckily, we can use robust methods to estimate an error of these quantities. The most known and well-established is *bootstrap method*.

Bootstrap belongs to a class of the *resampling methods*. Let us first sample N measurements of the energy. Next, we can consider these N measurements as a state space for the energy, forming some probability distribution. In essence, we want to infer the information about this probability and functions of it. For energy, our goal would be to calculate the mean and the error. The idea is to draw N new samples with equal probability and repetition from the original set of N samples, repeating the procedure M times. Calculating the mean for each set of M repetitions would get us M values of the mean. These new M values represent the probability distribution of the mean, and we can use them to calculate the standard deviation. This method is advantageous when, for example, we have stored some observable values but want to estimate some function of it.

An example of a function of an observable would be the specific heat $C_V(\langle E\rangle, \left\langle E^2 \right\rangle)$. The procedure would go like this, from initial N measurements sample new N samples with repetition M times. For each of the M samples, we compute C_V, leaving us with a list of specific heats that we call a. Now it turns out that the standard deviation of the specific heat is given by[2]:

$$\sigma^2 = \langle a^2\rangle - \langle a\rangle^2 \,. \tag{1.40}$$

This means we take the standard deviation of the newly constructed list of C_V values, and the mean would simply be the average of a. It can also be shown that even if we draw correlated samples, bootstrap will estimate standard deviation equally well.

So far, we have covered the basic steps and algorithms for constructing the Markov process with the desired distribution. These steps are the same for multiple fields where Monte Carlo simulation might be used. Now we focus on the applications in physics. We will stick with the $2D$ Ising model. At the mean-field level, Ising in $2D$ dimensions goes through the phase transition as we change the temperature. We can think of it this way, at a very high temperature $T = \infty$, all the spins are randomly oriented, with plenty of energy at their disposal and no order among the spins. This is a paramagnetic phase, where the total magnetization of the system is zero.

As we decrease the energy, spins try to align more and more, as this is now energetically more favorable, and at a certain point, the total magnetization becomes finite. This phase, where the total system has non-zero magnetization, is the ferromagnetic phase. As discussed in the previous section, this phase transition belongs to the second-order phase transitions class. Second order means the second derivative of the free energy is discontinuous. The standard second derivative is specific heat

1

$C_V = -T\left(\frac{\partial^2 F}{\partial T^2}\right)\Big|_{V=\text{const}}$. The straightforward way to see if the system goes through a phase transition is to plot specific heat as a function of the changing variable. Thus, specific heat should be diverging at the transition point for the second-order phase transition. Strictly speaking, diverging behavior is only seen in the thermodynamic limit (for infinite systems). Since we can only simulate systems of finite sizes, we will see that specific heat will develop a sharper peak as we increase the size. This can be seen in Fig.(1.6). We can compare the value of T with the exact known critical temperature $T_c = \frac{2J}{\ln\left(1+\sqrt{2}\right)} \approx 2.27J$, as well as the meanfield result $T_c = 4J$.

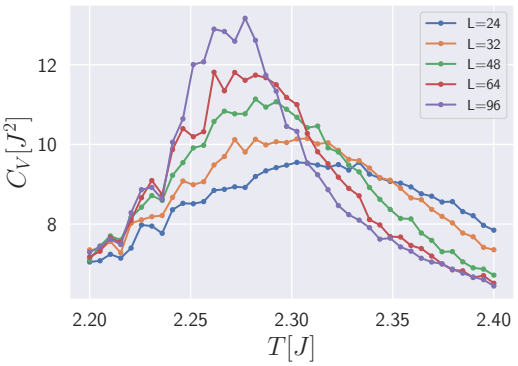


Figure 1.6: Specific heat curves C_V of 2D Ising for $J = 1$ on a square $L \times L$ lattice around the transition point for different system sizes. Strictly speaking, specific heat only shows diverging behavior for infinite system sizes. Because of this, determining the exact point of transition through these curves is not very precise. The exact value of critical temperature is $T_c = \approx 2.27$, while the mean-field result is $T_c = 4$ (way off on the above graph).

Now that we have identified the existence of a continuous phase transition, we would like to describe it in more detail for two reasons: both related to the numerical implementations. The first is critical slowing down, and the second is finite size determination of free phase transition. Landau's theory of phase transitions shows that the free energy has a singularity in the thermodynamic limit, usually described by a power law of the observables near the critical point. Underlying this divergence is a divergent *correlation length* ξ. It gives a sense of how correlated or ordered the spins in our system are, and it is diverging near the transition according to the:

$$\xi \sim |T - T_c|^{-\nu}\,, \tag{1.41}$$

whit ν the *critical exponent*. The specific heat has a similar shape near the criticality:

$$C_V \sim |T - T_c|^{-\alpha}\,, \tag{1.42}$$

with the difference that α can now be positive and negative. From the renormalization group theory, we know that these critical exponents are related by scaling

relations. In most cases, only two exponents are independent, and the rest can be derived from them[2]. These two exponents fully characterize the critical behavior of the model. Also, it can be shown that models often have the same set of these critical exponents and can be further categorized in *universality* classes. This means that models with different microscopic details exhibit the same behavior near the critical point, and details get washed away. Everything is controlled by ν.

The issue is that these divergences only occur in the thermodynamic limit. For finite-size systems, the critical behavior is smeared out, and estimating the location of the peak will introduce the error. Due to a related problem, critical slowing down, one cannot simply sample more[2].

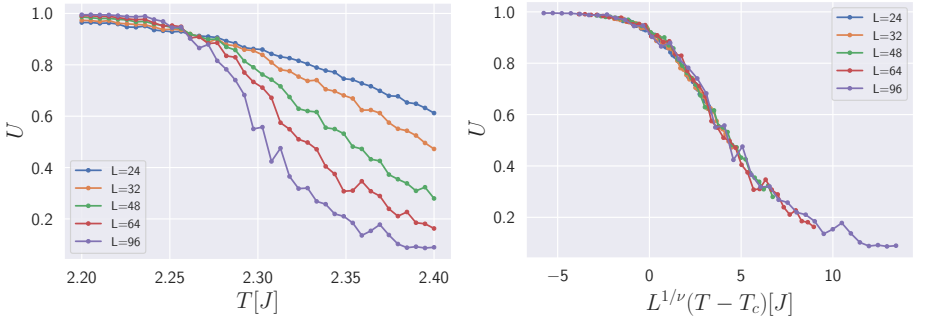
One better way to determine the transition point and critical exponents is through the technique of *finite size scaling*. Using the results from the renormalization group, we know that the magnetization scale, close to the transition, with system size L as:

$$\langle m_L \rangle \sim L^{\frac{\beta}{\nu}} \tilde{M}\left[L^{\frac{1}{\nu}}(T - T_c)\right] , \tag{1.43}$$

where $\tilde{M}(x)$ in unknown scaling function. We can see from the Eq.(1.43) that $\langle m_L \rangle / L^{\beta/\nu}$ at $T = T_c$ will be independent of the system size, as the argument of $\tilde{M}$ will be zero. So this means that scaled magnetizations will all have the same value precisely at the critical temperature. On the plot where these different scaled magnetization curves would be drawn together at the point of critical temperature, they would intersect each other. If we want to be completely precise, we are ignoring some non-analytic corrections to the scaling so the curves intersect only at the same point in the sense $L \to \infty$. Doing finite-size scaling using this method is not helpful in practice because neither β nor γ is known a priori. A better approach is to compute a combined quantity that is dimensionless. One such quantity is the *Binder cumulant*[3]:

$$U = \frac{1}{2}\left[3 - \frac{\langle m^4 \rangle}{\langle m^2 \rangle^2}\right] \sim \tilde{G}\left[L^{\frac{1}{\nu}}(T - T_c)\right] . \tag{1.44}$$

In the limit, $T \to 0$, Binder cumulant goes to 1; in the opposite limit, when the temperature $T \to \infty$ Binder goes to 0. Because Binder cumulant is a dimensionless quantity, different L curves should approximately intersect at the same point, given that the non-analytic corrections are small. We can use this to determine ν, see Fig.(1.7a). On the other hand, if we plot all the curves with the appropriate critical scaling ν on the plot U vs. $L^{1/\nu}(T - T_c)$ they all fall on top of each other as in Fig.(1.7b).



(a) (b)

Figure 1.7: **(a)** Binder cumulant curves for different system sizes. Because the Binder cumulant is a dimensionless quantity, all the curves cross at the same point, given that the non-analytic corrections are small. This method is more robust and reliable for determining the exact critical temperature $T_c \sim 2.26$. **(b)** After doing the finite size scaling for the Binder cumulant data in Fig.(1.7a), all the curves have collapsed on a single line. We can use Binder cumulant curves to determine the exact position of T_c and then use that to fit all the curves to determine the critical exponent $\nu = 1$.

In this thesis, we shall use the Binder cumulant differently to determine the order of a phase transition. In the case of continuous, second-order phase transition, the Binder cumulant has a smooth transition. In the case of a first-order phase transition, instead of being continuous, the Binder cumulant has a dip at the point of the critical transition. The dip occurs because, at the point of transition order, the parameter experiences a discontinuity which is also present in any functions that depend on it. This dip diverges with the system size[3]. We can use Binder cumulant to see whether the actual phase order exists and determine its order. Another possibility is the case of a weak first-order transition, where the Binder cumulant still exhibits a dip, but this dip is not diverging with the system size.

1.3. Machine learning

Since discovering quantum mechanics, physicists have sought better, more accurate approximation techniques to solve real-world quantum problems. As we know from the introductory quantum mechanics courses, only a handful of toy models Schrödinger equation can be solved exactly. If we start considering systems with many interacting particles, our analytical methods can only serve up to a certain point. Some well-known numerical techniques, such as exact diagonalization and tensor networks, also start having problems when we apply them to systems of many highly interacting particles. The Monte Carlo techniques reviewed in the previous section can provide approximate answers. Recently, one of the pioneering methods that used machine learning was introduced by Carleo and Troyer [4], where they consider an RBM network as an ansatz in variational quantum Monte Carlo.

The term *machine learning* (ML) has become an everyday expression in recent times. It can be defined as a sub-field of Artificial intelligence primarily tasked with developing algorithms for learning from experience (or data). The backbone of ML is based on applied statistics while also drawing some ideas from statistical physics. The main goal of ML is to estimate unknown, often very complicated, functions that depend on a large number of unknowns and give useful predictions on new (unknown) data. This program of pushing ever bigger and more complex algorithms led to the development of *deep learning* (DL), where massive networks with millions of parameters are trained on more extensive data sets. The development of highly efficient and parallel GPUs led to a boom in developing new ML and DL algorithms. The main appeal of ML and DL is that some tasks, hard to put in an algorithmic form, can now be tackled with ease. A typical example would be recognizing a cat, dog, or human being in a picture or detecting a new phase state of a quantum matter. ML and DL have grown tremendously in recent years; hence a complete introduction is simply impossible here, and we will try to introduce the essential elements of both fields used in the later chapters.

Before actually doing any computation, we need to establish the main ingredients we need to make the computer learn [5]:

- a *task* that we need to solve, like regression, classification, generating pictures, or learning how to play a game
- *data* in the case of ML and DL, we can look at the data as an experience that our algorithm lives and learns from. Some forms of data can be given in a table, like pictures of handwritten digits and their labels, or it can be given in the form of a reward, for example, how long the robot managed to walk
- a *model* that learns from the data.

Usually, more than these are needed to specify the learning algorithm fully. We need to define a measure that will compare the performance of our algorithm with the actual known data. Our task is to minimize this difference between the predicted and known data.

- *Regression* can be viewed as one of the most common forms of ML. We assume that there is some (potentially unknown) relation between two variables $\mathbf{x}$ and $\mathbf{y}$, and we try to learn what that relation is. We will call variable $\mathbf{x}$ an input and variable $\mathbf{y}$ an output. These variables are not restricted to a single dimension; they are often multi-dimensional. As stated, our task is to find some function f that will satisfy the relation $\mathbf{y} = f(\mathbf{x})$. The simplest example would be a linear regression, where we assume that our function has the form $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, and we have free parameters $\mathbf{A}$ and $\mathbf{b}$ to optimize.

- *Classification* is a task where our output variables do not have a continuous output but are rather discrete and categorized in different *classes (labels)*. In comparison with the regression, we are now trying to find a function that will map our input $\mathbf{x}$ to a target $\mathbf{y}$ but also encode a representation for multiple different classes. The simplest example would be a binary classification, a case when we need to distinguish between two different labels, for example, is this a picture of a cat or not. One of the most famous datasets that we will use later is MNIST (Modified National Institute of Standards and Technology), which consists of 70.000 handwritten digits distributed over 10 "classes", the digits $0, 1, 2, \ldots, 9$.

The data, the main ingredient in many machine learning tasks, comes in datasets $\mathscr{D}$, containing data points $\mathbf{x}_i$, $\mathscr{D} = \{\mathbf{x}_i\}$. Depending on the information available in our data set, we can divide the types of learning we can do:

- *Supervised learning* is a class of problems where our data points are *labeled*. It can either be a regression or a classification. Methods used to do supervised learning can be classical ML or more modern and complex DL.

- *Unsupervised learning* is applied when data does not have accompanying labels that we can use. Some examples of unsupervised learning methods are the initial pre-processing of data when we are trying to select the best features to use, for trying to find some order in the given data set by grouping points in clusters according to some features or increasing the dimensionality by adding features through generative models. An example from physics application is trying to distinguish two phases of a matter when we do not understand the underlying process governing it.

- *Reinforcement learning*, in this case, we usually do not have a data set, but rather some environment that our model explores. Through the actions of a model and feedback received from the environment model chooses the next best step in order to maximize some in advance predetermined metric.

In this thesis, we will explore some topics from the supervised classification problems through the DL and reinforcement learning methods to find the ground state of a quantum system.

The last ingredient we need is a model. In general, this is some function $f_{\boldsymbol{\theta}}(\mathbf{x})$ of the input data. There are various forms of functions, each suitable for a specific job. The model function is specified by giving its mathematical form, in ML the network, and a set of variable parameters $\boldsymbol{\theta}$, in ML the weights. Our job then is to train a model to find an optimal set of parameters $\hat{\boldsymbol{\theta}}$ that will minimize some target loss function $\mathscr{L}$ or maximize a model performance. In essence, we are modeling the true relation between $\mathbf{x}$ and ansatz $f_{\boldsymbol{\theta}}(\mathbf{x})$. In physics, we know this from variational wave function approximation. We will review this later. One of the most common loss functions to minimize is the mean-squared error (MSE) used in regression problems:

$$\mathscr{L}_{\text{MSE}} = \sum_{i=1}^{N} (y_i - f_{\boldsymbol{\theta}}(x_i))^2 \,, \tag{1.45}$$

where N is the number of data inputs. For classification problems, the most widely used loss function is a cross-entropy (CE). The simplest form of CE is when we only have two categories, binary cross-entropy (BCE). In the case of multiple different categories, we use categorical cross-entropy (CCE):

$$\mathscr{L}_{\text{BCE}} = -\sum_{i=1}^{N} y_i \log\big(f_{\boldsymbol{\theta}}(x_i)\big) + (1 - y_i) \log\big(1 - f_{\boldsymbol{\theta}}(x_i)\big) \,, \tag{1.46}$$

$$\mathscr{L}_{\text{CCE}} = -\sum_{i=1}^{N} \sum_{j=1}^{K} y_{i,j} \log\big(f_{\boldsymbol{\theta}}(x_i)\big) \,, \tag{1.47}$$

where now index j goes over K different categories and the constraint $\sum_i f_{\boldsymbol{\theta}}(x_i) = 1$. The last formula is written in a form where labels $y_{i,j}$ are given in a *one-hot encoding*.

$$y_{i,j} = \begin{cases} 1, & \text{if } y_i = j \\ 0, & \text{otherwise} \end{cases} \tag{1.48}$$

For example, if we have 5 categories, label $y_3 = (0,0,1,0,0)$ is a vector with all zeros and one at the i-th (3rd) class position.

We need to specify the minimization procedure after choosing a model and loss function. This procedure can either be gradient-based or gradient-free. A standard widely used is *gradient descent* (GD). In short, we start our learning process by initializing

a random set of parameters $\boldsymbol{\theta}_0$ and compute the loss function. After that, we compute the gradient of a loss function with respect to the model parameters. Finally, we update our parameters by subtracting previously computed gradients along the steepest direction:

$$\theta_{j+1} = \theta_j - \eta \frac{\partial \mathscr{L}}{\partial \theta_j} , \tag{1.49}$$

where η is a learning rate, it controls the size of steps we take. Every time we complete one full update, we say that one *epoch* has passed. Choosing an appropriate learning rate is a delicate process; if the learning rate is too small, our training will take forever, and we will waste computing resources. If the learning rate is too big, our algorithm might never converge. The most common way of choosing a learning rate is through trial and error.

Gradient descent is only sometimes the best algorithm to use. The algorithm will converge if the learning rate is small enough, but there is no guarantee that the minimum reached is the actual global minimum, not a local one. A clever way to deal with this problem is to introduce some stochasticity. There are various ways to do so, but one way is to incorporate it directly into GD. A standard, modified version of GD is called stochastic gradient descent (SGD). One computes the gradients not on a whole input data set but only on a small set of *batches*. The initial data set is divided into batches of equal size, and in each step, the neural network is trained only on a single batch. We say that one epoch has passed after we trained the neural network on all the batches and went through the data set once. Doing this has two benefits; one is that now we do not need to compute gradients with respect to all the inputs hence reducing the computational cost, and second by only updating parameters after computing gradients with respect to one batch, we are introducing randomness and avoiding saddle points and narrow local minima.

A further improvement to an SGD is in the form of momentum or inertia. This serves as a memory of the direction in which we are moving and helps us move in a direction with consistent but small gradients while avoiding oscillations in high curvature directions:

$$v_j = \gamma v_{j-1} + \eta \frac{\partial \mathscr{L}}{\partial \theta_j} , \tag{1.50}$$

$$\theta_{j+1} = \theta_j - v_j , \tag{1.51}$$

where v_j is a running average of the previously computed gradients and $(1-\gamma)^{-1}$ sets a time scale of how far back we want to look. If we set $\gamma = 0$, we revert to a previous case of (S)GD without momentum.

Parameters like η, γ, the number of batches we use, and the total number of epochs are called *hyperparameters*. Besides these, all parameters that control the learning process can be put in this category. In order to find a good set of hyperparameters,

the established method is to split the input data into two sets, *training set* and *validation set*. Doing this is a good practice besides finding an optimal set of hyperparameters. We want an independent data set that was not seen before, on which we will evaluate our model. In practice, we might have several choices for our models, and a consistent way to compare them is to see how they perform on previously unseen data. Being able to make generalizations and predict results from unknown data is the primary goal of ML and DL. We can get in a situation whereby choosing a very complex model will fit every possible feature in our data set (*overfitting* the data), but then when presented with a new, previously unseen data model, will perform terribly.

On the other hand, if we choose a simpler model, our loss might be higher for the training data. However, when presented with new data, this simpler model will outperform a more complicated one, see Fig.(1.8). This is called *bias-variance trade-off*.

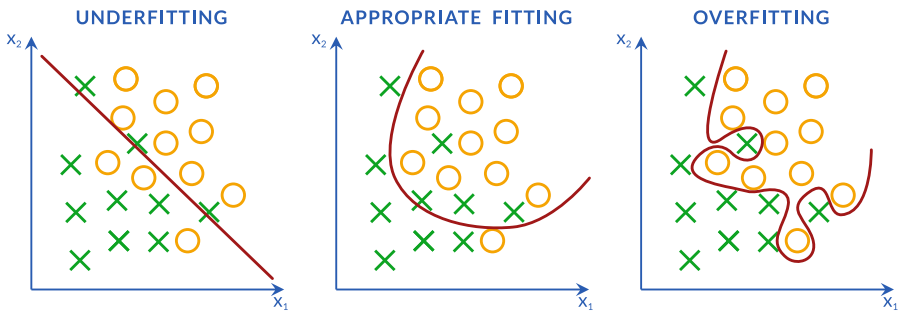


Figure 1.8: Three different scenarios of model selection. On the left, the selected model does not have enough representational power to capture desired features in our data set. The center image model is complex enough to balance bias and variance while still being able to generalize on new data. We have a too-complex model for the given data set on the right. Too many parameters and the model will be able to capture all the features of the data, even the noise, leading to an inferior generalization.

This principle can be nicely illustrated on an example of the MSE loss function and linear relation between input and target:

$$y = f_{\boldsymbol{\theta}}(\mathbf{x}) + \epsilon \,, \tag{1.52}$$

where ϵ is a noise distributed according to a Gaussian distribution $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. In this case, the MSE loss function would be:

$$\mathscr{L}_{\text{MSE}} = \sum_{i=1}^{N} (y_i - f(\mathbf{x_i}; \boldsymbol{\theta}))^2 \,. \tag{1.53}$$

Following the simple exercise in algebra and probability [5], we can factor this loss into three individual components:

1

$$\text{error} = \text{bias}^2 + \text{var} + \text{noise}\,, \tag{1.54}$$

where *bias* is:

$$\text{bias}^2 = \sum_{i=1}^{N} \left(f(\mathbf{x}_i) - \left\langle f(\mathbf{x_i}; \hat{\boldsymbol{\theta}}_{\mathscr{D}}) \right\rangle_{\mathscr{D}} \right)^2 , \tag{1.55}$$

and *variance*:

$$\text{var} = \sum_{i=1}^{N} \left\langle \left(f(\mathbf{x_i}; \hat{\boldsymbol{\theta}}_{\mathscr{D}}) - \left\langle f(\mathbf{x_i}; \hat{\boldsymbol{\theta}}_{\mathscr{D}}) \right\rangle_{\mathscr{D}} \right)^2 \right\rangle_{\mathscr{D}} . \tag{1.56}$$

Bias represents how well our model would perform if we had an infinite amount of data, and the variance tells us how much our model fluctuates because of the finite amount of samples we have. Increasing the number of parameters will reduce the bias, but at some moment, the variance will increase. This is the central concept in ML, the trade-off between the complexity of a model and the amount of data we have to train it on. Because, in practice, we are usually presented with a limited amount of data. Choosing a less complex model with a higher bias will often lead to less variance and better generalization on new data points.

Now that we have covered the basic intuition behind ML and DL, we will specify one simple model we will use later in the thesis, *logistic regression*. Let us consider a case where our target variable y can take a value $m = 0, \ldots, M-1$ from one of the M classes. We want to define a function that, given an input, returns a probability that it belongs to one of the M classes. One such function is the *sigmoid* function:

$$\sigma(s) = \frac{1}{1 + \exp(-s)} \tag{1.57}$$

Let us start first with a simple case of two classes $y_i = \{0, 1\}$. Then a probability that given data point $\mathbf{x}_i$ belongs to a category is:

$$p(y_i = 1 | \mathbf{x}_i, \boldsymbol{\theta}) = \frac{1}{1 + \exp\left(-\mathbf{x}_i^T \boldsymbol{\theta}\right)} \tag{1.58}$$

$$p(y_i = 0 | \mathbf{x}_i, \boldsymbol{\theta}) = 1 - p(y_i = 1 | \mathbf{x}_i, \boldsymbol{\theta}) \tag{1.59}$$

The most appropriate loss function would be cross-entropy:

$$\mathscr{L}_{\text{BCE}} = -\sum_{i=1}^{N} y_i \log\left(\sigma(\mathbf{x}_i^T \boldsymbol{\theta})\right) + (1 - y_i) \log\left(1 - \sigma(\mathbf{x}_i^T \boldsymbol{\theta})\right) \tag{1.60}$$

We can minimize this loss function, but there is no simple closed-form solution and some of the minimization methods we have discussed need to be applied. Moving to a general case when instead of 2, we have M classes, and following one-hot encoding, we can write individual probabilities for each class as:

$$p(y_{i,m'} = 1|\mathbf{x}_i, \{\boldsymbol{\theta}\}_{k=0}^{M-1}) = \frac{\exp\left(-\mathbf{x}_i^T \boldsymbol{\theta}_{m'}\right)}{\sum_{m=0}^{M-1} \exp\left(-\mathbf{x}_i^T \boldsymbol{\theta}_m\right)} \tag{1.61}$$

This is known as *SoftMax* function, and the appropriate cross-entropy is:

$$\begin{aligned}\mathscr{L}_{\text{CCE}} = -\sum_{i=1}^{N}\sum_{m=0}^{M-1} y_{i,m} \log\big(p(y_{i,m} = 1|\mathbf{x}_i, \boldsymbol{\theta}_m)\big) \\ + (1 - y_{i,m}) \log\big(1 - (p(y_{i,m} = 1|\mathbf{x}_i, \boldsymbol{\theta}_m))\big)\end{aligned} \tag{1.62}$$

1.3.1. Neural networks for ML

Models that we have considered so far are simple ML models that form a basis for any further improvements. The introduction of neural networks catapulted the field of ML and DL to the heights of today.

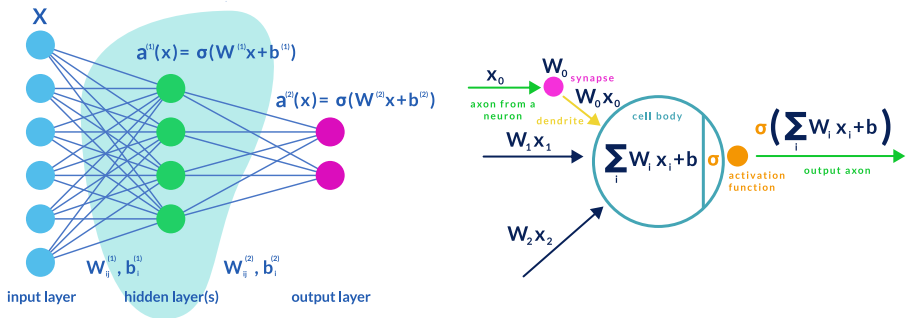


Figure 1.9: On the left is an illustration of a fully connected neural network with a single hidden layer and two layers in total. On the right is a neuron, an elementary building block of neural networks.

The fundamental building element of a neural net is a neuron i, Fig.(1.9) that takes as an input d dimensional vector of features $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ and produces a scalar output $a_i(\mathbf{x})$. A neural network is formed by stacking many neurons on top of each other to form a single layer. Then multiple layers are stitched together to form the whole network. Stacking multiple layers enhances our neural networks' expressivity, which helps us approximate complex functions. The existence of a universal approximation theorem states that a neural network with a single layer can approximate

any "nice" and continuous function with arbitrary accuracy. The first layer is called an input layer, the last layer is called an output layer, and any layer(s) in-between is(are) called hidden layer(s). What makes neural networks work is an introduction of some non-linearity between layers, *activation function*. We usually take the same activation function for all the neurons in a single layer. Going from one to the other layer, first, the linear transformation of the form is computed:

$$z^i = \boldsymbol{\omega}^i \cdot \mathbf{x} + \mathbf{b}^i = \mathbf{x}^T \cdot \mathbf{w}^i \,, \tag{1.63}$$

where weight vector $\boldsymbol{\omega}^i = (\omega_1^1, \omega_2^1, \ldots, \omega_d^1)$, $\mathbf{x} = (1, \boldsymbol{x})$ and $\mathbf{w}^i = (b^i, \boldsymbol{\omega}^i)$. After the linear transformation, the activation function is applied to get a post-activation value of a layer:

$$a_i(\mathbf{x}) = \sigma_i(z^i) \tag{1.64}$$

There are many possibilities for activation functions; some of the popular ones are step-function, sigmoid, hyperbolic tangent, rectified linear units (ReLUs), leaky rectified linear units (leaky ReLUs), and exponential linear units (ELUs)[5]. What we choose as an activation function will impact the performance of our network and greatly depend on a specific task we are trying to solve, which we will later see in Chapter 5. Another consideration should be taking a derivative of activation functions as they are required for any gradient-based function minimization methods.

If we remember from our previous sections how to train a network, we will need to find a derivative of a loss function with respect to all the weights for any gradient-based method. At first, this seems like a daunting task. However, luckily, there is a very nice and elegant algorithm that can help us find all of the desired derivatives easily, *backpropagation*. Behind the fancy name is nothing more complicated than a simple chain rule for differentiation. Because this is crucial background knowledge to the chapter, we discuss it here. In order to fully understand this backbone of an algorithm, let us set the stage first. We will take a network with L layers labeled with $l = 1, \ldots, L$. Weights $\omega_{j,k}^l$ connect k-th neuron from $l-1$ layer to a j-th neuron in layer l. The bias associated with this neuron is b_j^l. Then we can write a post-activation value a_j^l of this neuron as:

$$a_j^l = \sigma\left(\sum_k \omega_{j,k}^l a_k^{l-1} + b_j^l\right) = \sigma(z_j^l) \tag{1.65}$$

where the linear combination part, the pre-activation value, is defined as:

$$z_j^l = \sum_k \omega_{j,k}^l a_k^{l-1} + b_j^l \tag{1.66}$$

Let us think about how we compute the loss function. Directly the value of a loss function depends on the post-activation values a_j^L from L-th layer, but these values indirectly depend on the post-activations from the previous layers. This is the crucial observation to set up a backpropagation algorithm. Define the error Δ_j^L of j-th

neuron in the final L-th layer as a partial derivative of the loss function with respect to the weighted input z_j^L:

$$\Delta_j^L = \frac{\partial \mathscr{L}}{\partial z_j^L} \tag{1.67}$$

Analogous to this, we can define the error for any neuron j in arbitrary layer l as:

$$\Delta_j^l = \frac{\partial \mathscr{L}}{\partial z_j^l} = \frac{\partial \mathscr{L}}{\partial a_j^l}\sigma'(z_j^l) \tag{1.68}$$

we can also make the chain rule different in order to get the following:

$$\Delta_j^l = \frac{\partial \mathscr{L}}{\partial z_j^l} = \frac{\partial \mathscr{L}}{\partial b_j^l}\frac{\partial b_j^l}{\partial z_j^l} = \frac{\partial \mathscr{L}}{\partial b_j^l} \tag{1.69}$$

where we have used Eq.(1.66) to find $\frac{\partial b_j^l}{\partial z_j^l} = 1$. As we stated before, the layered structure of the network ensures that the layer error $l+1$ depends on the post activations from layer l, and we can use the chain rule to expand:

$$\begin{aligned}\Delta_j^l = \frac{\partial \mathscr{L}}{\partial z_j^l} &= \sum_k \frac{\partial \mathscr{L}}{\partial z_k^{l+1}}\frac{\partial z_k^{l+1}}{\partial z_j^l} \\ &= \sum_k \Delta_k^{l+1}\frac{\partial z_k^{l+1}}{\partial z_j^l} \\ &= \left(\sum_k \Delta_k^{l+1}\omega_{k,j}^{l+1}\right)\sigma'(z_j^l)\end{aligned} \tag{1.70}$$

The final equation is:

$$\frac{\partial \mathscr{L}}{\partial \omega_{j,k}^l} = \frac{\partial \mathscr{L}}{\partial z_j^k}\frac{\partial z_j^k}{\partial \omega_{j,k}^l} = \Delta_j^l a_k^{l-1} \tag{1.71}$$

Now we have all the necessary ingredients to state the entire backpropagation algorithm:

Algorithm 2 The backpropagation algorithm

1: Calculate the post-activation values of all neurons in the input layer a_j^1
2: Now using Eq.(1.65) compute all the z^l and a^l values until the last layer
3: Calculate the error of the final layer using Eq.(1.68); for this, we will need to compute the analytical form of loss and activation functions manually
4: Using Eq.(1.70) we can propagate error backwards in order to calculate Δ_j^l
5: The final step is using Eqs.(1.69, 1.71) in order to calculate the desired derivatives $\frac{\partial \mathscr{L}}{\partial w_{j,k}^l}$ and $\frac{\partial \mathscr{L}}{\partial b_j^l}$

The name backpropagation now makes sense; we are using a single forward pass through the network to compute linear combinations and post-activations, and then by backtracking through the network, we compute all of the derivatives. This specific nature of backpropagation makes it highly efficient when implemented on modern GPU units. The immediate relevance to the work in this thesis is the appearance of σ' in Eq.(5.2). A core part of chapter 5 is the exploration of the effects of saturation, domain regions where $\sigma' = 0$ and therefore negligible gradients, on the final training efficiency.

1.3.2. Neural quantum states

In order to motivate this next class of neural networks, we use some topics from quantum mechanics to guide us. Let us say that we have an isolated quantum system of spins 1/2 in a chain of length N. Then we can expand any wave function in some arbitrary basis that will have 2^N coefficients. For example, consider a $1D$ system with 40 spins. Directly writing down all the coefficients alone would be an infeasible task; there are $2^40 \sim 10^12$ different coefficients, and just writing them down would take up more than 40TB of space. If we want to study real-life systems with many more particles and in many more dimensions like $2D$ or $3D$, we need to find some other methods of doing it. This should not be the end; usually, only a small part of a Hilbert space is of relevance, and this fact can help us a lot. We can apply some of the *variational methods* that try to find the optimal representation of the quantum state, one that will encompass all the necessary features and be computationally friendly. We can write the basis expansion as:

$$\left|\Psi_{\boldsymbol{\theta}}\right\rangle = \sum_{s=1}^{2^N} \Psi_{\boldsymbol{\theta}}(\mathbf{s}) \left|s\right\rangle \tag{1.72}$$

where $\Psi_{\boldsymbol{\theta}}(\mathbf{s}) = \left\langle s\middle|\Psi_{\boldsymbol{\theta}}\right\rangle$ and our goal would be to find the best $\Psi_{\boldsymbol{\theta}}(\mathbf{s})$ that approximate a ground state but $\dim[\boldsymbol{\theta}] \ll \dim[s]$. We can then use this proposed model for the wave function to compute the expectation values of desired operators in polynomial time. To cast the usual complex wave function computations in terms of probabilities, we use the following local estimator method to compute the expectation value of some arbitrary operator $\hat{O}$ [5]:

$$\begin{aligned}\left\langle \hat{O} \right\rangle &= \frac{\left\langle \Psi_{\boldsymbol{\theta}}\middle|\hat{O}\middle|\Psi_{\boldsymbol{\theta}}\right\rangle}{\left\langle \Psi_{\boldsymbol{\theta}}\middle|\Psi_{\boldsymbol{\theta}}\right\rangle} \\ &= \frac{\sum_{s,s'} \left\langle \Psi_{\boldsymbol{\theta}}\middle|s\right\rangle \left\langle s\middle|\hat{O}\middle|s'\right\rangle \left\langle s'\middle|\Psi_{\boldsymbol{\theta}}\right\rangle}{\sum_s \left|\left\langle \Psi_{\boldsymbol{\theta}}\middle|s\right\rangle\right|^2} \\ &= \frac{\sum_s \left|\left\langle \Psi_{\boldsymbol{\theta}}\middle|s\right\rangle\right|^2 \sum_{'s} \left\langle s\middle|\hat{O}\middle|s'\right\rangle \frac{\left\langle s'\middle|\Psi_{\boldsymbol{\theta}}\right\rangle}{\left\langle s\middle|\Psi_{\boldsymbol{\theta}}\right\rangle}}{\sum_s \left|\left\langle \Psi_{\boldsymbol{\theta}}\middle|s\right\rangle\right|^2}\end{aligned} \tag{1.73}$$

We can identify two terms here:

$$p(\mathbf{s}) = \frac{|\langle \Psi_{\boldsymbol{\theta}} | s \rangle|^2}{\sum_s |\langle \Psi_{\boldsymbol{\theta}} | s \rangle|^2} \tag{1.74}$$

$$O_{\text{loc}}(s) = \sum_{s'} \langle s | \hat{O} | s' \rangle \frac{\langle s' | \Psi_{\boldsymbol{\theta}} \rangle}{\langle s | \Psi_{\boldsymbol{\theta}} \rangle} \tag{1.75}$$

The first is the usual quantum mechanics probability density. The second one $O_{\text{loc}}(\mathbf{s})$ is called a *local estimator* of $\hat{O}$. Therefore we can write a quantum mechanical expectation as a classical expectation value:

$$\left\langle \hat{O} \right\rangle = \sum_s p(\mathbf{s}) O_{\text{loc}}(\mathbf{s}) = \left\langle O_{\text{loc}}(\mathbf{s}) \right\rangle_p \tag{1.76}$$

The procedure of how to estimate the expectation value of any operators in the form of Eq.(1.76) is very reminiscent of Eq.(1.21) so we can write it as:

$$\left\langle \hat{O} \right\rangle \approx \frac{1}{N_{\text{samp}}} \sum_{i=1}^{N_{\text{samp}}} O_{\text{loc}}(\mathbf{s}^i) \tag{1.77}$$

Now we are back to the conventional statistical physics that we reviewed earlier. We now know how to compute this by constructing a Markov Chain and using Monte Carlo simulation to compute the expectation value.

What remains is to make a suitable variational ansatz $\Psi_{\boldsymbol{\theta}}(\mathbf{s})$. This is where we combine ML with Monte Carlo, an insight from Carleo and Troyer [4]. We will use the *restricted Boltzmann machines* (RBM) for this job. Neural network architecture is called the *neural quantum state* (NQS).

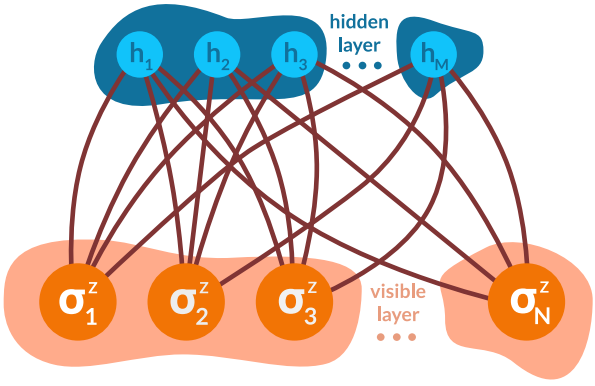


Figure 1.10: Illustration of a restricted Boltzmann machine (RBM) neural network. We use RBM as a variational ansatz to represent a wave function of N spins with $s = (\sigma_1^z, \sigma_2^z, \ldots, \sigma_N^z)$ and M hidden units $h = (h_1, h_2, \ldots, h_M)$. There is general proof that a sufficiently dense ($\alpha \gg 1$) RBM can approximate any probability distribution [6].

RBM networks are fully connected two-layer networks with one visible and one hidden layer. The visible layer has N spins, and the hidden layer has M spins. We will label the ratio of the number of hidden vs. visible neurons as $\alpha = \frac{M}{N}$. The insight of Carleo and Troyer was that the intrinsically non-local correlations of RBM could lead to a significantly more compact representation of many-body quantum states. Then the ansatz wave function is given by:

$$\Psi_{\boldsymbol{\theta}}(\mathbf{s}) = \sum_{h} \exp\left(\mathbf{b}_v^\dagger \cdot \mathbf{s} + \mathbf{b}_h^\dagger \cdot \mathbf{h} + \mathbf{h}^\dagger \mathbf{W} \mathbf{s}\right). \tag{1.78}$$

Because of the network architecture, we can trace out hidden variables and get the following:

$$\Psi_{\boldsymbol{\theta}}(\mathbf{s}) = \exp\left(\mathbf{b}_v^\dagger \cdot \mathbf{s}\right) \prod_{i=1}^{M} 2\cosh\left(\mathbf{b}_{h,i} + \mathbf{W}_i \cdot \mathbf{s}\right), \tag{1.79}$$

where $\mathbf{b}_{h,i}$ and $\mathbf{W}_i$ are i-th hidden bias and weight matrix row. Now we have a model wave function that will convert our spin configuration into a desired probability.

With this ansatz, we are back in familiar territory to find the ground state energy of some Hamiltonian:

$$E(\boldsymbol{\theta}) = \left\langle \Psi_{\boldsymbol{\theta}} \middle| \hat{H} \middle| \Psi_{\boldsymbol{\theta}} \right\rangle, \tag{1.80}$$

in terms of ML, this will be our loss function. We must also remember that this variationally obtained ground state energy is $E(\boldsymbol{\theta}) \geq E_0$, but this is strictly true when expectation values are calculated exactly.
We can exchange operator $\hat{O}$ for the energy in Eq.(1.77) in order to get stochastically approximated energy:

$$E(\boldsymbol{\theta}) \approx \frac{1}{N_{\text{samp}}} \sum_{i=1}^{N_{\text{samp}}} E_{\text{loc}}(\mathbf{s}^i) \tag{1.81}$$

where we have defined our local energy as $E_{\text{loc}}(\mathbf{s}) = \sum_{s'} \langle s|\hat{H}|s\rangle \frac{\langle s'|\Psi\rangle}{\langle s|\Psi\rangle}$. We will use gradient-based optimization methods to minimize our loss function. The energy gradient with respect to parameters is:

$$\frac{\partial E(\boldsymbol{\theta})}{\partial \theta_i} = 2\,\text{Re}\left[\left\langle E_{\text{loc}}(\mathbf{s}) O_i^\star(\mathbf{s})\right\rangle - \left\langle E_{\text{loc}}(\mathbf{s})\right\rangle \left\langle O_i^\star(\mathbf{s})\right\rangle\right] \tag{1.82}$$

where θ_i is the i-th parameter and operator $\hat{O}_i$ is defined as:

$$O_i(\mathbf{s}) = \frac{\partial}{\partial \theta_i} \log\left\langle s \middle| \Psi_{\boldsymbol{\theta}} \right\rangle = \langle s|\hat{O}_i|s\rangle \tag{1.83}$$

Now we have all the necessary ingredients to formulate the final algorithm for finding the ground state of a Hamiltonian $\hat{H}$ using RBM and variational method:

Compared to the current standard numerical methods, NQS can have the same or better accuracy[4]. Compared with MPS, PEPS, and DMRG for transverse-field Ising

Algorithm 3 Finding ground state with NQS

Randomly initialize all the parameters of a neural network $\boldsymbol{\theta}$
for $i = 1$ to N_{steps} **do**
  Generate N_{samp} samples using Markov chain
  Calculate the gradient of the energy $\frac{\partial E(\boldsymbol{\theta})}{\partial \theta_i}$ using Eq.(1.82)
  Update parameters using SGD or some other minimization technique
end for
return Optimized parameters $\hat{\boldsymbol{\theta}}$

(TFI) and anti-ferromagnetic Heisenberg models (AFH), NQS achieved better accuracy for sufficiently large α ration. AFH, when compared to DMRG, NQS with $\alpha = 4$ managed to outperform DMRG with the bond dimension of ~ 160, pointing towards a much more compact representation of a many-body wave function. This also has practical implications, as fewer variational parameters are easier to optimize. Even when applied at the critical point, NQS managed to get the accuracy of the state-of-the-art methods or even better, albeit with slowed down converged, which is expected near the criticality. The compact nature and ability to express the wave function near the critical point prompted us to use it to analyze entanglement entropy for lattice gauge fields explored in chapter 4 of the thesis.

1.4. This thesis

In the introduction, we have covered the basic ideas used later in this thesis. We started with introductory topics in thermodynamics and statistical physics, then moved to a basic introduction to Monte Carlo methods and all the required knowledge to understand our physical system's simulation design and analyze the results. The proceeding section was dedicated to the basics of machine learning, deep learning, and appropriate selection of model, loss function, and minimization method. The last section culminated in a synergy of the previously mentioned topics by combining quantum physics, Monte Carlo methods, and neural networks in neural quantum states that we used to find the ground state and its energy of lattice gauge theories.

1.4.1. Chapter 1 - Thermalization in quantum systems

The properties of closed unitary quantum systems and how they exactly thermalize have been one of the leading research questions in statistical physics for a long time. The puzzle is that a thermal ensemble is formally a mixed state, but a mixed state can never arise from unitary evolution from a pure state. The usual answer to how they thermalize is the eigenstate thermalization hypothesis ("ETH") [7]. The hypothesis is that in generic quantum theory with many degrees of freedom, most observables will have a particular form of matrix elements after averaging, and observable will *appear* to thermalize. However, recently [8] showed that ETH has to be taken with care. Even in free field theory, there are operators that *appear* to relax, called operator thermalization hypothesis (OTH). Given a particular no-go condition, the retarded Green's function will typically decay exponentially unless the condition is met. Finding an operator that will satisfy this condition in a general non-integrable theory is challenging but possible. On the other hand, this job is more straightforward in integrable theories due to the extensive number of conserved quantities. We work in the transverse field Ising (TFI) model where we compare a specially designed operator Γ that will satisfy the no-go condition with the Pauli σ^z operator that does not satisfy it. Through the examples, we show the differences and similarities of ETH and OTH and how, despite TFI being an integrable theory, σ^z will relax after the perturbation. Also, we have demonstrated how the no-go condition is a feature of integrability, and any minor deviation from integrability will cause Γ to relax. Our results were later confirmed in [9].

1.4.2. Chapter 2 - Symmetry restoration through "registry"

Starting from the simple Wegner gauge theory [10, 11], Fradkin and Shenker [12] discovered that when an added matter field is "in the fundamental", meaning that

there is an additional Higgs field that is also governed by Z_2 symmetry as the gauge fields, the Higgs phase and confining phase become one, without a phase transition. In this chapter, we propose a straightforward generalization of their lattice gauge theory that could serve as a candidate for a highly entangled state of matter. We will consider adding multiple Z_2 and $O(N)$ matter fields on the lattice and gauging them with a common Z_2 field. It will be shown how, in such a case, the Higgs phase becomes separate from the confining phase. It will be characterized by the "registry" order parameter, which turns out to be gauge invariant $p = 2^{N_{\text{rep}-1}}$ Potts type symmetry, where N_{rep} is the number of matter field copies. Interpretation for this type of symmetry is that different matter copies align their vectors locally in strictly parallel or anti-parallel ways, even in the case of continuous $O(N)$ matter fields. These theories will be studied using Monte-Carlo simulation on a $3D$ grid using the Metropolis-Hastings algorithm and annealing techniques to improve the convergence near the critical point. From the simulation results, we can discover some unidentified "pseudo-universality" associated with the form of the phase diagram for various numbers of matter field theories.

1.4.3. Chapter 3 - Entanglement entropy of lattice gauge theories

Building further on the work from the previous chapter, we will study entanglement entropy in the neural network representation of the above lattice gauge theories, now considered as quantum theories in one lower dimension. Following the seminal work of Carleo and Troeyer [4], we will construct neural quantum states as the representation of our theory using a variational wavefunction based on Restricted Boltzmann Machines used in Machine Learning. Using ideas from tensor networks that the bond dimension represents the upper bound on the amount of entanglement a state can have, we will postulate that by increasing the number of matter fields, ground state entanglement entropy of our lattice gauge theory will increase as the ratio of hidden vs. visible nodes. We have tested our hypothesis in the case of $2,3$, and 4 matter fields. Within the achievable computational limits, the results are puzzling. Even though increasing the number of variational parameters improved the energy of the ground state, the impact on the entanglement entropy is less than obvious. Curves of entanglement entropy for different system sizes look the same up to the statistical errors.

1

1.4.4. Chapter 4 - Phase space and efficient learning of deep neural networks

This chapter combines some statistical physics insights into machine learning with the computational mechanics of deep random feedforward neural networks. In recent times with the ever-growing amount of available data, neural networks have become one of the de-facto methods for analyzing and processing vast amounts of data [13]. One of the reasons why these methods became so popular is their ability to express any function with a relatively small number of parameters [5] and the ease with which this *expressivity* can be increased by adding more depth. This easy fix does not come for free; deep neural networks generally require more training computations. Specifically, they suffer from exploding or vanishing derivatives in optimizing the parameters. The phase space of deep random feedforward neural networks is characterized by the variance of initial weights and the variance of initial bias. Following previous work [14, 15] that demonstrated the existence of order-to-chaos regime change in this phase space, we will examine the behavior of the pre- and post-activations in terms of their distributions and also final accuracy on classification task such are MNIST and CIFAR10. The phase boundary dividing these two regimes is called the edge of chaos (EOC). We demonstrate that for the tanh activation function, not all points along the EOC yield the same learning efficiency. In the case of shallow and narrow neural networks, we define the line of uniformity (LOU), a set of points for which the final layer post-activation values are distributed uniformly, i.e., with maximal entropy. We show that moving away to the right from LOU and drastically increasing initial variances means that gradient saturation will start impeding optimization over parameters, i.e., the learning process.

2

Operator Thermalization vs Eigenstate Thermalization

Attribution

This paper has been previously published as a preprint on arXiv under the title ***Operator thermalization vs eigenstate thermalization***, together with Philippe Sabella-Garnier and Koenraad Schalm.[16]

Abstract

We study the characteristics of thermalizing and non-thermalizing operators in integrable theories as we turn on a non-integrable deformation. Specifically, we show that σ^z, an operator that thermalizes in the integrable transverse field Ising model, has mean matrix elements that resemble ETH, but with fluctuations around the mean that are sharply suppressed. This suppression rapidly dwindles as the Ising model becomes non-integrable by the turning on of a longitudinal field. We also construct a non-thermalizing operator in the integrable regime, which slowly approaches the ETH form as the theory becomes non-integrable. At intermediate values of the non-integrable deformation, one distinguishes a perturbatively long relaxation time for this operator.

2.1. Introduction

The question of how closed, unitary quantum systems can (appear to) thermalize has long been at the heart of statistical mechanics. Recently, it has become more pressing because of its implications for real-life experiments [17–22] and, through holography for the black hole information paradox [23–29]. The usual answer to the puzzle is the eigenstate thermalization hypothesis ("ETH")[30–33]. At its core, the statement is the following. Suppose that there is a regime where the matrix elements of an observable $\mathcal{O}$ in the basis of energy eigenstates closely approximate the following form:

$$\begin{aligned} \langle m|\mathcal{O}|n\rangle &\approx O(E)\delta_{mn} + e^{-S(E)/2} f(E,\omega) R_{mn} \,, \\ E &= \frac{E_m + E_n}{2} \;\;,\;\; \omega = E_n - E_m \,, \end{aligned} \tag{2.1}$$

where $O(E)$ and $S(E)$ are the microcanonical expectation value of $\mathcal{O}$ and entropy at energy E, $f(E,\omega)$ is a smooth function and R_{mn} is a random matrix with zero mean and unit variance. Then, it can be shown that the long time average of the expectation value of $\mathcal{O}$ in a superposition of energy eigenstates (such as a state produced by a quench in the Hamiltonian of the system) will approach its thermal expectation value, with the temperature set by the average energy of the initial state. The hypothesis is that in generic quantum theories with a large number of degrees of freedom, "most" observables have matrix elements approximately of this form and therefore the system will appear to thermalize.

Nevertheless, while this is a sufficient condition for apparent thermalization, it is not *necessary*. It is often argued that the validity of equation (2.1) for generic operators is a symptom of quantum chaos. However, it has been noticed in various contexts that even in *free* systems certain (usually composite) operators can relax to a thermal state, at least at the level of linear response [34–40]. In [40], it was in fact shown that the thermal retarded Green's function of an operator in a free theory will generically decay exponentially in time unless the operator satisfies a particular no-go condition. That condition is:

$$|\langle m|\mathcal{O}|n\rangle|^2 = 0 \;\text{ unless }\; E_n - E_m = F_i^{(\mathcal{O})}(P_n - P_m) \,, \tag{2.2}$$

where $E_{m,n}$ and $P_{m,n}$ are the energy and momentum of the states, $F_i^{(\mathcal{O})}(P)$ are (not-necessarily continuous) functions that depend on $\mathcal{O}$, with i an index that runs over a finite (system size-independent) range. By extension, such a statement should hold in any integrable theory, with a combination of the extensive set of conserved charges playing the role of momentum in the above expression.

In a generic non-integrable theory, finding an operator satisfying this no-go condition is hard, if not impossible, since the momentum difference and energy difference between two states are a priori independent quantities. In integrable theories, the

extensive number of conserved quantities makes finding operators that satisfy this condition easier. For example, in a free field theory the field itself obeys it, and one easily constructs others. However, it is also not hard to evade the no-go condition: any operator that involves two uncorrelated momentum modes will do so (for example, the square of a free field). As explained in [40], even free and integrable theories therefore have many operators that thermalize in linear response.

Of course, integrable field theories cannot be said to be chaotic for any reasonable definition of the word. In fact, their spectrum is highly regular. The fact that, at least at the level of linear response, many operators are sufficiently blind to this structure (as expressed by violating this no-go condition) and appear to thermalize is the idea that we have called *operator thermalization*. This is in contrast with eigenstate thermalization in which it is the (lack of) structure of the spectrum itself that is responsible for thermalization.

In this note, we aim to determine the difference between these two ideas more concretely by studying thermalization in a one-dimensional quantum Ising chain. When only a transverse field is present, the model is integrable, whereas it is chaotic for a certain regime with both transverse and longitudinal fields. As an example of OTH, the local magnetization σ^z, which violates the no-go condition relaxes even in the integrable regime. We show that its matrix elements in the integrable theory take a form that is *also* consistent with equation (2.1) provided we average over small energy windows. However, a detailed examination shows that the integrable structure of the spectrum is reflected in a non-Gaussian spectrum for R_{mn}. As we turn on the non-integrable deformation and transition to the chaotic regime, R_{mn} becomes smoother and one observes a classic example of ETH.

We then compare this with the behaviour of a non-thermalizing operator Γ in the integrable regime. By construction, this operator satisfies the no-go condition and does not relax. It therefore manifestly does not obey the ETH ansatz, even after averaging. Deforming the theory to the non-integrable regime slowly induces a violation of the no-go condition, and the operator then approaches a form compatible with equation (2.1). In contrast to σ^z, one clearly sees the onset of a long perturbative slowest-relaxation timescale ("mean free path") in the system, that shortens as the degree of non-integrability is increased.

We summarize our results in table 2.1. For other work on ETH in the context of integrable theories, see [41–46].

Theory	Operator	Satisifies no-go condition?	Relaxes?	Obeys ETH ansatz?
Integrable	σ^z	No	Yes	Yes, but with R_{mn} more sharply peaked than a Gaussian
	Γ	Yes	No	No
Non-integrable	σ^z	No	Yes	Yes
	Γ	No	Yes, but with a long relaxation time	Yes, with $f(E,\omega)$ flat as a function of ω

Table 2.1: Summary of results

2.2. Model details

The one-dimensional Ising model with transverse and longitudinal fields has the following Hamiltonian:

$$H = -J\sum_{n=1}^{N}\left(\sigma_i^z\sigma_{i+1}^z + h\sigma_i^x + g\sigma_i^z\right), \tag{2.3}$$

where σ_i^a are the usual Pauli matrices, obeying

$$[\sigma_i^a, \sigma_j^b] = 2i\epsilon^{ab}{}_c\sigma_i^c\delta^{ij}. \tag{2.4}$$

We impose periodic boundary conditions, $\sigma_{i+N}^a \equiv \sigma_i^a$. When $g = 0$, the transverse field Ising model is integrable and can be mapped to a model of free spinless fermions through a series of textbook transformations. First, a Jordan-Wigner transformation and Fourier transform will make the Hamiltonian quadratic:

$$\begin{aligned}
&\sigma_i^x = 1 - 2c_i^\dagger c_i \ , \ \sigma_i^z = -\prod_{j<i}(1-2c_j^\dagger c_j)(c_i + c_i^\dagger)\,, \\
&c_j = \frac{1}{\sqrt{N}}\sum_{k\in\mathscr{K}} c_k e^{ikr_j}\,, \\
&\{c_k, c_{k'}^\dagger\} = \delta_{kk'}\,.
\end{aligned} \tag{2.5}$$

We think of the system as being on a lattice with lattice spacing a and total size $L = Na$, so that $r_j = ja \in [a, Na]$. The periodic boundary conditions on σ_i^a impose either periodic or anti-periodic boundary conditions on the fermionic operators (depending on the total number of fermions), leading to

$$\mathscr{K} = \left\{\frac{2\pi}{L}n \,\middle|\, n\in\mathbb{Z} \text{ or } \left(\mathbb{Z}+\frac{1}{2}\right)\right\}. \tag{2.6}$$

Of course, the momenta must lie in the first Brillouin zone, leading to $-\frac{\pi}{a} < k \leq \frac{\pi}{a}$, so that $n \in \left(-\frac{N}{2}, \frac{N}{2}\right]$. In practice, we will work with $a = 1$, so that N measures system size. This transformation is followed by a Bogoliubov transformation:

$$\begin{aligned} c_k &= u_k \gamma_k + i v_k \gamma^\dagger_{-k} , \\ u_k &= \cos(\theta_k/2) , \; v_k = \sin(\theta_k/2) , \; \tan\theta_k \equiv \frac{\sin(ka)}{h - \cos(ka)} . \end{aligned} \tag{2.7}$$

In terms of these fermions, the Hamiltonian is diagonal:

$$H|_{g=0} = \sum_{k \in \mathscr{K}} \epsilon_k \left(\gamma^\dagger_k \gamma_k - \frac{1}{2} \right) , \; \epsilon_k \equiv 2J\sqrt{1 + h^2 - 2h\cos(ka)} . \tag{2.8}$$

The momentum operator is then

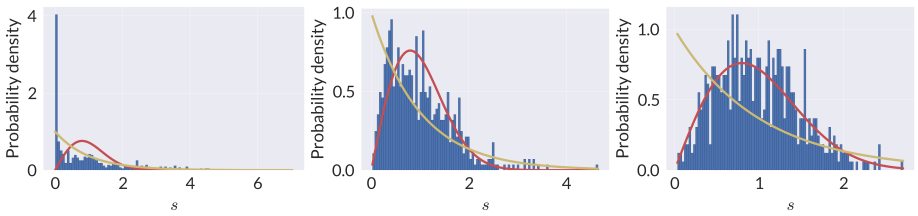
$$P = \sum_{k \in \mathscr{K}} k \gamma^\dagger_k \gamma_k . \tag{2.9}$$

Numerically, we work in this fermion basis, labelling states by occupation number of each of the momenta in $\mathscr{K}$ with the appropriate boundary conditions. To construct operators in the non-integrable regime, we first construct them in the basis of eigenstates of the integrable Hamiltonian. We then diagonalize the non-integrable Hamiltonian and numerically find the transformation between the eigenvectors. For the integrable model, we work in a basis of joint eigenvectors of the occupation number of each of the momentum modes. Away from integrability, the Hamiltonian is still translationally-invariant. We therefore work in a basis of joint eigenvectors of H and the translation operator. Throughout this paper, we set the value of the transverse field to $h = -1.05$, following [47] which studied thermalization in the mixed field Ising chain. We will mostly focus on three values of the parallel field: $g = 0$ (integrable), $g = 0.1$ (which we label simply "non-integrable") and $g = 0.5$, which (following [47]) we label "far from integrable".

In figure 2.1, we show the level statistics for these three values of the transverse field in one particular sector (i.e. for states with one particular eigenvalue of the translation operator), confirming that the far-from-integrable case follows a Wigner distribution while the integrable case is Poisson-distributed.

2.3. Operators: thermalizing and non-thermalizing

As discussed in the introduction, we will be considering two different operators in both the integrable and non-integrable regimes. The first operator is σ^z_i. This operator clearly violates the no-go condition (2.2) in the integrable regime: its matrix elements are non-zero for a two-dimensional subregion of the $(\Delta E, \Delta P)$ plane, as opposed to a discrete set of lines. This is seen explicitly in figure 2.2a This confirms the analysis made in [40] by analytical methods at the critical point ($h = 1$).



(a) Integrable ($g = 0$) (b) Non-integrable ($g = 0.1$) (c) Far from integrable ($g = 0.5$)

Figure 2.1: Distribution of unfolded level spacings in units of average level spacing s for states with one eigenvalue of the translation operator ($e^{\frac{(N-2)\pi i}{N}}$). The red line corresponds to the Wigner sumrise for the Gaussian Orthogonal Ensemble. The green line corresponds to a Poisson distribution. Level repulsion is very clearly visible for the far from integrable case. Note that the intermediate $g = 0.1$ case can be seen to approach the Wigner sumrise overall but still shows an excess of approximate degeneracies. $N = 13$.

By contrast, we can use the free fermion basis to construct an operator that obeys the no-go condition (2.2). Take

$$\Gamma = \frac{2\pi}{L} \sum_{k \in \mathscr{K}} \left(\gamma_k \gamma_{k+\delta} + \gamma^\dagger_{k+\delta} \gamma^\dagger_k \right), \tag{2.10}$$

where δ is an arbitrary (fixed) shift in momentum space. We will take it to be as small as possible, that is to say $\delta = \frac{2\pi}{N}$. This is the simplest operator that satisfies the no-go condition without being a conserved current. It creates pairs of particles with correlated momenta. It is easy to see that, in the integrable theory, such an operator has non-zero matrix elements only between states where

$$\Delta P = \pm(2k + \delta) \tag{2.11}$$

$$\Delta E = \pm(\epsilon_k + \epsilon_{k+\delta}) = \pm\left(\epsilon_{\frac{\Delta P - \delta}{2}} + \epsilon_{\frac{\Delta P + \delta}{2}} \right). \tag{2.12}$$

This is confirmed by figure 2.2b.

In the integrable regime, $\Gamma(x,t)$ is easily obtained by Fourier transforming:

$$\Gamma(x,t) = \frac{2\pi}{L} \sum_{k \in \mathscr{K}} \left(e^{i(2k+\delta)x} e^{-i(\epsilon_k + \epsilon_{k+\delta})t} \gamma_k \gamma_{k+\delta} + e^{-i(2k+\delta)x} e^{i(\epsilon_k + \epsilon_{k+\delta})t} \gamma^\dagger_{k+\delta} \gamma^\dagger_k \right), \tag{2.13}$$

In the non-integrable regime, we can construct $\Gamma(x,t)$ by evolving Γ in time explicitly with the non-integrable Hamiltonian and translation operator.

In figure 2.3a, we examine the finite-temperature retarded two-point function of σ^z, $-i\Theta(t)\langle[\sigma^z_i(t), \sigma^z_i(0)]\rangle_\beta$, as a function of time in both the integrable and non-integrable regimes. We can clearly see that it relaxes in both cases. To confirm that Γ does not relax in the integrable theory, but does as we move away from integrability, we study its retarded Green's function as a function of the parameter g while holding

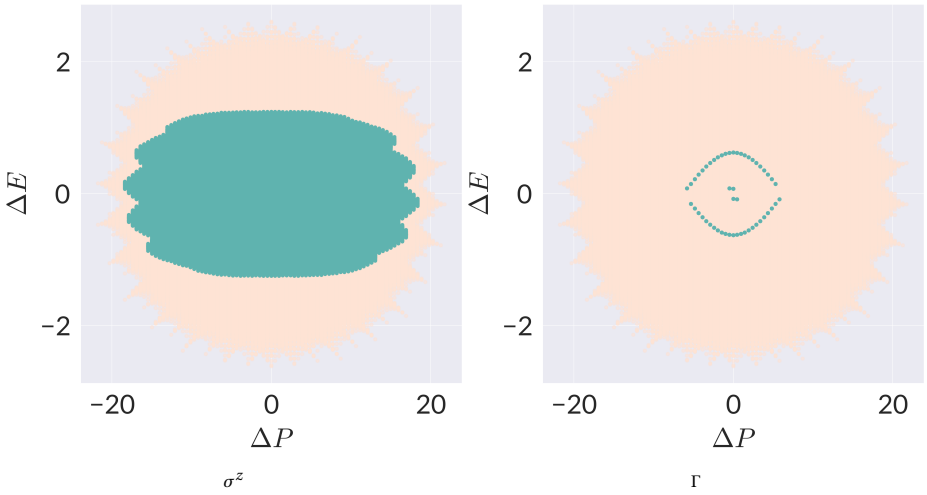
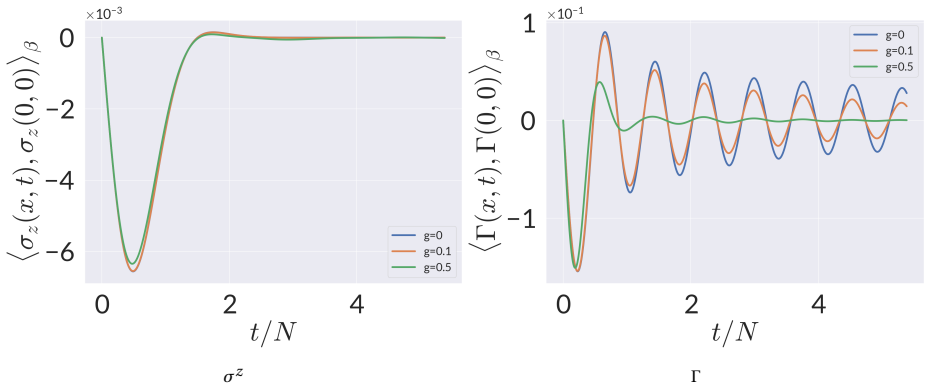


Figure 2.2: The no-go condition (2.2) for a thermalizing (2.2a) and non-thermalizing (2.2a) operator. Peach points correspond to zero matrix elements while teal points correspond to non-zero matrix elements. We can clearly see that for the thermalizing operator, the matrix elements are generically non-zero whereas for the non-thermalizing operator they are only non-zero when ΔE is given by a finite number of functions of ΔP. $N = 13$. Note that the fact that there is a non-zero matrix element with a particular ΔE and ΔP does not exclude that some other matrix element with those same values is zero.

h fixed. This is shown in figure 2.3b. There are two characteristic timescales present in this response. We see that at $g = 0$, the two-point function for Γ does not relax, but as we increase g it does. At $g = 0.5$, it relaxes in a comparable manner to σ^z. We can Fourier transform $G_R(t)$ to better study the two timescales involved: the resulting frequency distribution can be fit to Lorentzian distributions, consistent with a signal of the form $e^{-\Omega t}\sin(\omega_0 t)$. The position of the peaks of the Lorentzian gives ω_0 and their width gives Ω. The lifetime of the excitation, Ω^{-1}, and the damping ratio $\zeta = \sqrt{\frac{\Omega^2}{\omega_0^2+\Omega^2}}$ are shown in figure 2.4 as a function of the magnitude of the longitudinal field.

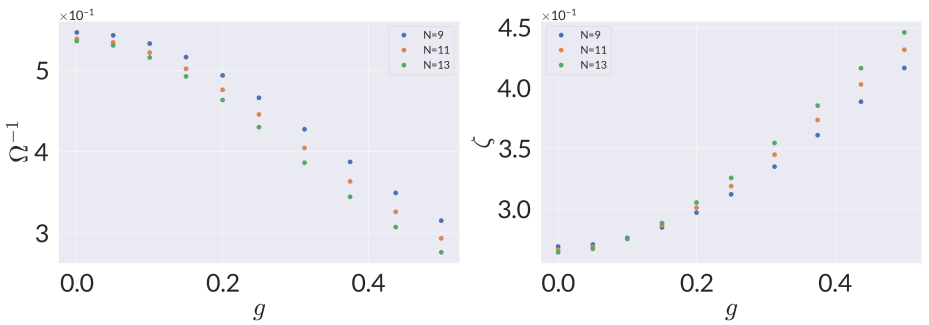
2.4. OTH vs ETH

Both OTH and ETH are fundamentally formulated in terms of the matrix elements of an operator in a basis of energy eigenstates, $\langle m|\mathcal{O}|n\rangle$. OTH specifically is a corollary to the no-go condition expressed for $\langle m|\mathcal{O}|n\rangle$. To study the relation between—and transition from— OTH to ETH, we study the matrix elements of both σ^z and Γ as we turn on the non-integrable longitudinal field. Note that in the integrable theory, the energy eigenvalues are degenerate, whereas they are not so in the non-integrable theory (up to momentum). For a proper comparison, we will therefore at various stages take an average in both cases over a small energy window ($\delta E = 0.01$) while



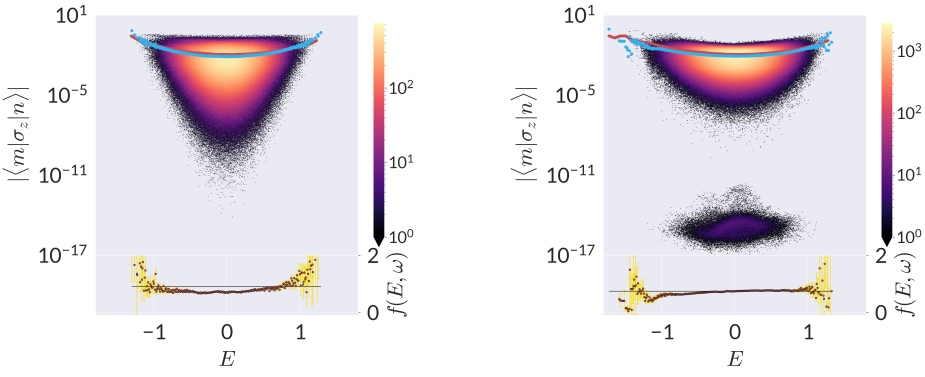
σ^z Γ

Figure 2.3: Finite-temperature retarded Green's function for σ^z and Γ for various integrability-breaking parameters g, with $h = -1.05$ We can see that σ^z always relaxes whereas in the integrable regime Γ does not, and becomes more damped as g is increased. $N = 13$



(a) Lifetime (b) Damping ratio

Figure 2.4: Lifetime (in units of system size) and associated damping ratio of an excitation of the thermal state by Γ as a function of longitudinal field g for different system sizes. We can see that for small g the lifetime scales with system size. As g is increased, the lifetime drops (and does so more steeply as we approach a continuum limit. We single out three values of g that are of interest: at $g = 0$ the theory is exactly integrable. $g = 0.1$ displays a measurable break from integrability. Finally, $g = 0.5$ is chaotic



(a) Integrable

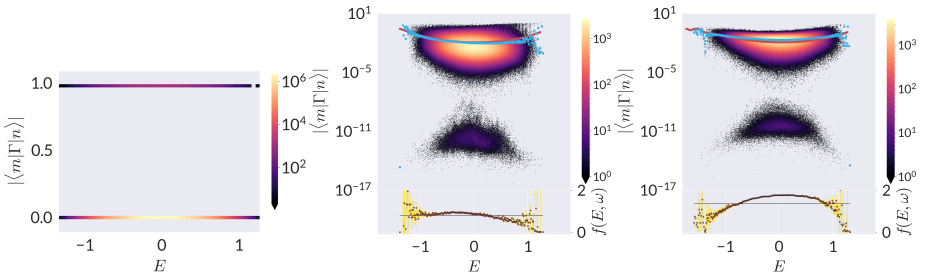
(b) Far from integrable

Figure 2.5: Dependence of the absolute value of matrix elements of σ^z on the average energy of the states at fixed ω. The blue points correspond to a running average over a small energy window and the red line is $e^{-S(E)/2}$. The bottom of the figure shows the running average divided by $e^{-S(E)/2}$, which gives the E dependence of $|f(E,\omega)|$. In both cases, the result is consistent with that function not depending on E. The error bars correspond to a 95% confidence interval if the underlying distribution is normal.

holding ω fixed or vice-versa ($\delta\omega = 0.8$).

2.4.1. Thermalizing operator, E dependence

In figure 2.5, we show the dependence of matrix elements of the thermalizing operator on the average energy of the states, $E = \frac{E_m+E_n}{2}$. In the integrable regime, half of the matrix elements $\langle m|\sigma^z|n\rangle$ are exactly zero because of parity symmetry (i.e. because the Hamiltonian is invariant under $\sigma_i^z \to -\sigma_i^z$). We exclude these points from our analysis. The behaviour of the remaining matrix elements in the integrable theory is strikingly similar to those in the non-integrable one. Taking a running average over a small (but finite) energy window allows us to extract a smooth function. In the non-integrable case, that should correspond to $e^{-S(E)/2} f(E,\omega)$. The same also happens—perhaps surprisingly— in the integrable case: the average over an energy window also scales predominantly as $e^{-S(E)/2}$. This need not have been, but shows explicitly the similarity between OTH and ETH at the level of averages. It explains in particular why many studies in 2D CFTs, which have an extensive number of conserved quantities, nevertheless find ETH-like behaviour, even though it is usually a case of OTH (see, eg [24–26, 29]). We can see that most of the dependence on E comes from this exponential factor of entropy, as expected. In the insert, we extract the function $f(E,\omega)$.



(a) Integrable (b) Non-integrable (c) Chaotic

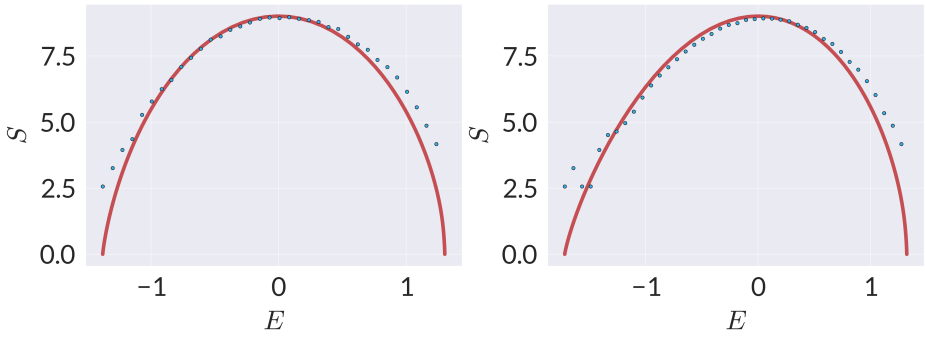
Figure 2.6: Dependence of the absolute value of matrix elements of Γ on the average energy of the states at fixed ω. Note that in the integrable case, we do not perform an average or use a log scale, since the large majority of points are exactly zero. In the non-integrable and chaotic cases, the blue line corresponds to a running average over a small energy window and the red line is $e^{-S(E)/2}$. The bottom of the figure shows the running average divided by $e^{-S(E)/2}$, which gives the E dependence of $|f(E,\omega)|$. Unlike for σ^z, there appears to be a non-trivial E dependence. The error bars correspond to a 95% confidence interval if the underlying distribution is normal.

2.4.2. Non-thermalizing operator, E dependence

By contrast, the matrix elements of the non-thermalizing operator Γ clearly (by construction) do not follow an ETH-like distribution as a function of average energy or entropy in the integrable case. This is seen in figure 2.6. Indeed, they are very sensitive to the fine-grained structure of the states, as opposed to coarse-grained features like the average energy. However, immediately upon turning on the non-integrable deformation the matrix elements of the operator start to look ETH-like. From a microscopic point of view, this is not surprising, since there are no more details of the state for it to depend on: the additional conserved charges coming from integrability are at this point completely meaningless. We see here the effects of true ETH, which is able to overcome the fact that the operator was constructed in the integrable theory explicitly to evade OTH. There is one subtle distinction with the thermalizing operator σ^z. There is now a small remnant dependence on E in addition to the entropic suppression $e^{-S(E)/2}$, i.e. the function $f(E,\omega)$ is not flat as a function of E. This dependence becomes more pronounced as the system becomes more chaotic.

2.4.3. Statistics

We now probe a bit deeper into the meaning of the running average. As mentioned above, we extract $|f(E,\omega)|$ by averaging the magnitude of the matrix elements over a small window (where we assume this smooth function to be constant) and dividing the result by $e^{-S(E)/2}$, where $e^{S(E)}$ is the number of states in the window. The method to determine this entropy turns out to be irrelevant. Figure 2.7 shows the agreement between this microcanonical entropy and the usual canonical entropy calculated at



(a) Integrable (b) Chaotic

Figure 2.7: Agreement between microcanonical and canonical entropy. The microcanonical entropy (blue points) is obtained from the logarithm of the number of states in a small but finite energy window (of size $\delta E = 0.01$). The canonical entropy (solid red line) is obtained from the usual expression evaluated at a temperature where the average energy corresponds to the energy in question.

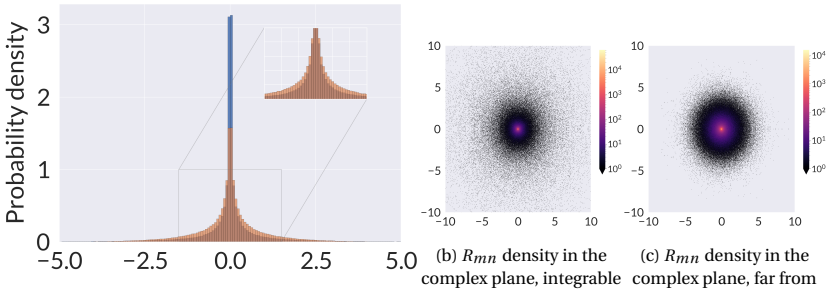
a temperature set by the average energy. Note these two match best where the spectrum is densest (i.e. around $E = 0$), and that in the integrable case it is essential to take a finite window. This is crucial. In order for the resemblance between OTH and ETH to become apparent, we have found that we must average over several energy levels in the integrable theory. A naive guess could have been that one only needed to sum over the degeneracy of a single energy level, and correspondingly in the non-integrable theory a window that just captured the splitting of these levels as the symmetry protecting the degeneracy is broken by the non-integrable deformation. This turned out to be insufficient and too narrow a window to see the resemblance between OTH and ETH. The resemblance is there for the larger window presented in figure 2.5.

Insightful results follow from looking not at the average, but at the full statistical distribution. Following ETH, parametrize the matrix elements as

$$O_{mn} = e^{-S(E)/2} f(E,\omega) R_{mn} \,. \tag{2.14}$$

We now extract R_{mn}. In figure 2.8, we show a typical example of the distribution within a window of $\frac{O_{mn}}{|O_{mn}|}$ for σ_z. Without loss of generality, we can take $\overline{|R_{mn}|} = 1$ since $\overline{O_{mn}} = 0$, so that $\frac{O_{mn}}{|O_{mn}|} = R_{mn}$.

We then see that the statistical distribution of values of σ^z around the mean reveals a distinction between the integrable and non-integrable theories. In the non-integrable regime, where ETH should hold, the matrix elements in our energy window have a standard deviation of order 1. However, in the integrable regime the distribution is distinctly more peaked (although of comparable variance). To better un-

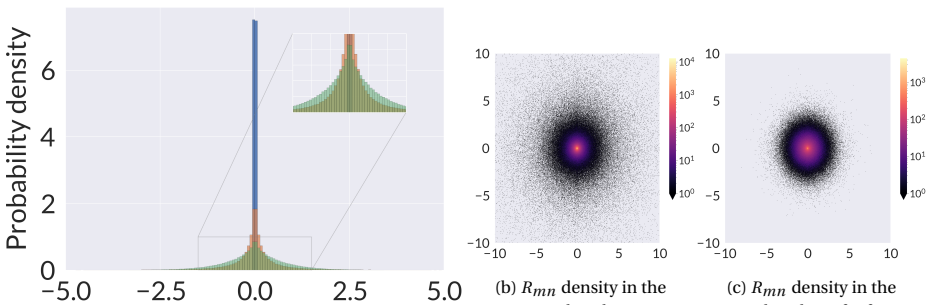


(a) Re(R_{mn})

(b) R_{mn} density in the complex plane, integrable ($h = 0$)

(c) R_{mn} density in the complex plane, far from integrable ($h = 0.5$)

Figure 2.8: (Colour online) Statistical distribution of the matrix elements $e^{-S(E)/2} f(E,\omega) R_{mn}$ of σ^z in a small energy window around $E = 0$, $\omega = 0$. $N = 13$.



(a) Re(R_{mn})

(b) R_{mn} density in the complex plane, non-integrable ($h = 0.1$)

(c) R_{mn} density in the complex plane, far from integrable ($h = 0.5$)

Figure 2.9: Statistical distribution of the matrix elements $e^{-S(E)/2} f(E,\omega) R_{mn}$ of Γ in a small energy window around $E = 0$, $\omega = 0$. $N = 13$.

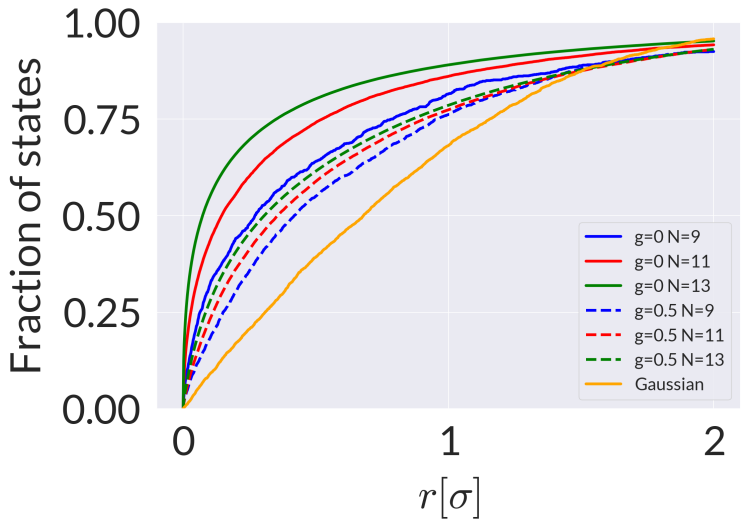


Figure 2.10: The cumulative distribution function for $|R_{mn}|/\sigma$ (as defined in the text). A faster early growth corresponds to a more peaked distribution. A Gaussian is shown for reference. Notice that σ^z in the integrable regime is not only more peaked than in the chaotic regime but becomes more so as N increases.

derstand this, we can plot the cumulative distribution of $|R_{mn}|$, that is to say

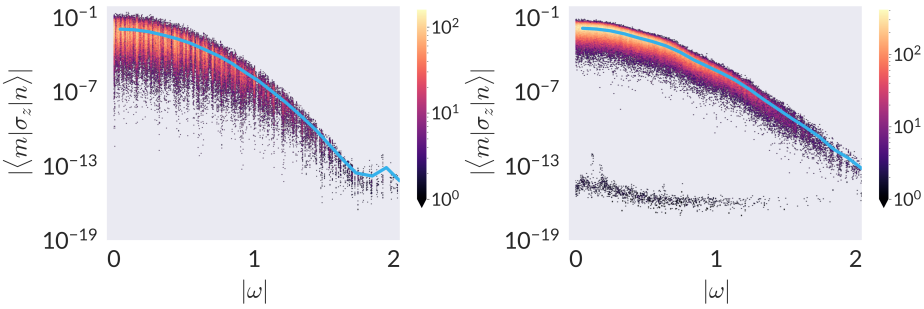
$$\chi(|R_{mn}|/\sigma) = \int_0^{|R_{mn}|/\sigma} P(|x|)dx\,, \tag{2.15}$$

where $P(|x|)$ is the probability distribution of R_{mn}/σ. This is the probability that the absolute value of the matrix element is less than or equal to a particular value $|R_{mn}|$, with everything expressed in units of the standard deviation. In figure 2.10, we show how for σ^z this function increases more sharply near zero in the integrable regime than in the chaotic regime, indicating that the probability distribution is more peaked.

Considering the statistical distribution around the average for the non-thermalizing operator $\langle m|\Gamma|n\rangle$ the fluctuations around the average are again non-Gaussian distributed. In the integrable system, manifestly so. There is in essence no distribution. As the system becomes more chaotic, a distribution develops which is somewhat more peaked than the peaked non-Gaussian distribution for the thermalizing operator. This is seen in figure 2.9.

2.4.4. Dependence on energy difference ω

The remaining ETH-like property to study is the dependence of the mean $f(E,\omega)$ on the energy difference. Fixing the average energy of the states and examining the dependence on the energy difference ω, we again confirm the similarity between the



(a) Integrable

(b) Far from integrable

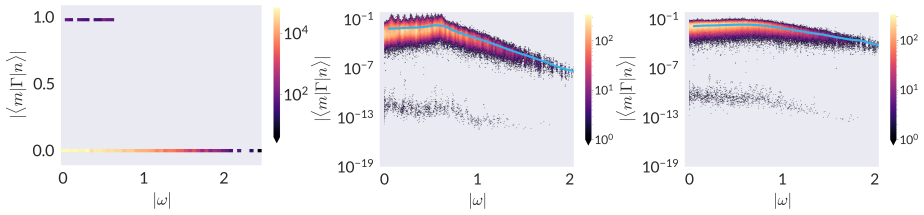
Figure 2.11: Dependence of the absolute value of matrix elements of σ^z on the energy difference between the states at fixed average energy E. The blue line corresponds to a running average over a small energy window, which is equivalent (up to an overall factor of $e^{-S(E)/2}$ to $|f(E,\omega)|$.

matrix elements $\langle m|\sigma^z|n\rangle$ of the thermalizing operator in the integrable and non-integrable theories. This is shown in figure 2.11. We show both the exact answer and a running average over a small energy window. The latter displays the expected smooth dependence of the matrix elements on energy—this time, the dependence on ω of $e^{-S(E)/2}f(E,\omega)$. A curious feature is that the dependence on ω is already noticeable at $\omega = 0$. There is no random matrix theory-like plateau for $\omega < \omega_*$. Studying the energy difference dependence for Γ, on the other hand, does show this cut-off frequency below which the response is RMT-like, once the system has become non-integrable.[1] We do not have an explanation for this distinction between the two operators. Nor does there appear to be a relation between ω_* and the relaxation time Ω^{-1} displayed in figure 2.4. We leave a better understanding of these scales to further study. We note, however, that the trivial ω dependence can be understood by the fact that Γ loses all meaning when far away from integrability. Indeed it was built out of a few single-particle operators, but the physics of the model can no longer be understood in this language.

2.5. Discussion and outlook

In this note, we have demonstrated by explicit examples the differences and similarities between operator thermalization and eigenstate thermalization. We emphasize the point again: while the anstaz for matrix elements in equation (2.1) is such that operators that obey it will relax to their thermal expectation values, satisfying that ansatz for all (or most) operators in the theory is not necessary for there to be some operators that do relax. This is especially true when an average is taken, so that the

[1] In the integrable regime, the dependence on ω is highly erratic: this is because the operator is very dependent on the details of the spectrum and cannot be simply understood in terms of the energies of the states.



(a) Integrable (b) Non-Integrable (c) Far from integrable

Figure 2.12: Dependence of the absolute value of matrix elements of Γ on the energy difference between the states at fixed average energy E. The blue line in the last two figures corresponds to a running average over a small energy window, which is equivalent (up to an overall factor of $e^{-S(E)/2}$) to $|f(E,\omega)|$. $N = 13$

details of the statistical distribution of matrix elements is smoothed over. Indeed, we have shown that σ_i^z in a transverse field Ising chain, is consistent with this ansatz. This is despite the fact that the TFI is an integrable model. The corollary statement that an operator satisfying ETH implies quantum chaos is therefore also manifestly not true.

We have also illustrated how the no-go condition is a feature of integrability: even a small move away from integrability caused our operator Γ to relax, with the relaxation becoming faster as we moved farther away. This move away from integrability was also correlated with the matrix elements approaching a more ETH-like form.

A natural next route of inquiry is to study quenches and non-linear response. So far, we have focused on matrix elements and linear-response two-point functions. However, we can also ask how quenching with operators in different classes might produce states approximating different ensembles. An obvious question is whether there is a connection between operators satisfying the no-go condition and the resulting density matrices approaching a thermal ensemble vs a generalized Gibbs ensemble. This can be studied numerically using the examples we have presented here, but also analytically by closely examining the form that ETH should take in the presence of conserved charges and examining possible (in)compatibility with the no-go condition, equation (2.2).

Finally, one can wonder how far operators can go towards mimicking chaotic properties of the spectrum of theories, often the underlying physics behind ETH. This may be tested by studying the behaviour of out-of-time-order correlators and more generally operator growth. Once again, the example of thermalizing operators in free and integrable theories leads to the obvious question of whether they behave differently under such measures than non-thermalizing opertors. We hope to report on this soon.

Acknowledgements

We are thankful to Marko Kuzmanović, Szilagyi Dániel, Tereza Vakhtel and Jan Zaanen for stimulating discussions. This research was supported in part by Koenraad Schalm's VICI award of the Netherlands Organization for Scientific Research (NWO), by the Netherlands Organization for Scientific Research/Ministry of Science and Education (NWO/OCW), and by the Foundation for Research into Fundamental Matter (FOM). PSG acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [PDF-517202-2018].

3

Replicating Higgs Fields in Ising Gauge Theory: the Registry Order

Attribution

This paper has been previously published as a preprint on arXiv and has been submitted to Physical Review E for publications, and it is currently under the editorial review, under the title ***Replicating Higgs fields in Ising gauge theory: the registry order***, together with Koenraad Schalm and Jan Zaanen.[48]

Abstract

We consider Z_2 gauge field theory coupled to "Higgs" matter fields invoking several copies of such matter, interacting entirely through the gauge fields, the $Z_2 \times Z_2 \times Z_2 \cdots / Z_2$ and the $O(N) \times O(N) \times O(N) \cdots / Z_2$ families of theories. We discover that the Higgs phase of such theories is characterized by a hitherto unidentified "registry" order parameter. This is characterized by a gauge invariant $p = 2^{N_{\rm rep}}/2$ Potts type symmetry where $N_{\rm rep}$ is the number of matter copies. The meaning of this registry is that the different matter copies align their vectors locally in strictly parallel or anti-parallel fashion, even dealing with the continuous $O(2)$ symmetry. Supported by Monte-Carlo simulations, we identify the origin of this registry order in terms of

the gauge interactions mediated by the fluxes ("visons") associated with the Z_2 gauge fields, indirectly imposing the discrete symmetry in the gauge invariant global symmetry controlled effective order parameter theory. In addition, it appears that our simulations reveal a hitherto unidentified "pseudo-universality" associated with the very similar form of the overall phase diagrams of the various theories suggesting a remarkable "governance" by the gauge field part of the dynamics.

3

3.1. Introduction

The physics of gauge fields has its special traits and arguably the simplest incarnation of such an interacting theory, controlled by Ising (Z_2) local symmetry, has played a historical role in the subject, comparable to the (global) Ising model itself in revealing generic principle. Early on, the pure Z_2 gauge theory was introduced by Wegner [10, 11], revealing in a minimal setting the fundamentals in the form of the confining and deconfining phases. Much later it was gradually realized that, different from the confining phase, the deconfining phase is characterized by *topological* order [49–54]. Kitaev rediscovered this in the context of topological quantum computation in the form of his toric code [55] demonstrating the non-Abelian braiding associated with its fluxes. This insight into the Ising deconfining state acquired in a recent era plays a central role in various guises in the modern topological order portfolio of condensed matter physics.

In a next step one can couple in Higgs fields ("matter") and generically a Higgs phase is formed besides the (de)confining phases. Next to the role it played in quantum condensed matter (e.g. [51, 52]), this flourished particularly in the context of liquid crystals. It was early on realized that the Higgs-confinement transition of $O(3)$ matter Z_2 gauge ($O(3)/Z_2$) is equivalent to the (uniaxial) nematic to isotropic transition [56] (see also [53, 54]). Recently this was exploited to study "generalized" nematics characterized by the breaking of the non-abelian point groups in 3D [57].

In fact, this started with the pioneering work by Fradkin and Shenker [12] that focused in on the minimal case where matter is "in the fundamental", e.g. the Higgs field is governed by Z_2 symmetry as the gauge fields. They found that the "maximally ordered" Higgs phase and "maximally disordered" confining phase actually become *indistinguishable*. This confused the community for a long while and perhaps this got resolved for the first time by Huse and Leibler [58] invoking an analogy with amphiphilic phases as of relevance in e.g. biology. We will review this in more detail underneath since this motif will play an important role in the remainder.

We stumbled on a natural extension of this discrete gauge theory portfolio. Although we have not managed to identify any circumstance where this could be of empirical relevance, as a theoretical construct it is entertaining. It adds some new general mo-

tifs, while shedding also an unexpected light on the basic physics of discrete gauge systems. We call this "discrete gauge theory with *replicated* Higgs fields", where various matter fields couple to the same gauge field. This defines a vast landscape of theories, and we will focus on the most elementary examples. We will limit ourselves to three (overall) dimensional systems characterized by a single Z_2 gauge field coupling to "replicated" Z_2 and $O(2)$ Higgs fields.

Let us first define this "replicated" theory. Discrete gauge symmetry can only be handled departing from a UV lattice. Let us consider the Euclidean action that may be interpreted as a thermal problem in $D = d+1$ dimensions, or either as the action of the quantum incarnation in d space dimensions after Wick rotation. Consider a hyper-cubic lattice in D overall dimensions and assign matter (Higgs) fields $\vec{\phi}_i^a$ to the site i with a symmetry that will be specified in a moment. The novelty is in the $a = [1, N_{\rm rep}]$ different "flavor" copies of the matter fields. Next, assign gauge group matrices U_{ij} to the *links* between nearest-neighbor lattice sites. The action is then in full generality of the form,

$$\begin{aligned} S[\{\phi^a, U_{ij}\}] = K\sum_{\square} \mathrm{Tr}\big[U_{12}U_{23}U_{34}U_{41}\big] + \sum_{a=1}^{N_{\rm rep}} J_a \sum_{\langle ij\rangle} U_{ij}\vec{\phi}_i^a\cdot\vec{\phi}_j^a \\ + \sum_{ab} J^L_{ab}\sum_i \vec{\phi}_i^a\cdot\vec{\phi}_i^b \end{aligned} \tag{3.1}$$

in the usual guise of defining the gauge curvature in terms of the Wilson plaquette action (first term), coupling minimally to matter (second term). Once again, the only novelty is in the consideration of the $N_{\rm rep} > 1$ matter field copies. The last term corresponds with the minimal, gauge invariant couplings between the Higgs field "replicas". A crucial ingredient is that all replica's are subjected to coupling to the same gauge field.

We will be focused on the minimal Z_2 gauge field incarnation in a thermal setting, addressing its statistical physics in 3 overall space dimensions. In terms of Ising spins ± 1 with Pauli matrices living on the bonds with operator τ^z_{ij},

$$\begin{aligned} S_{Z_2}[\{\phi^a, U_{ij}\}] = -K\sum_{\square} \tau^z_{12}\tau^z_{23}\tau^z_{34}\tau^z_{41} - \sum_{a=1}^{N_{\rm rep}} J_a \sum_{\langle ij\rangle} \tau^z_{ij}\vec{\phi}_i^a\cdot\vec{\phi}_j^a \\ - \sum_{ab} J^L_{ab}\sum_i \vec{\phi}_i^a\cdot\vec{\phi}_i^b\,. \end{aligned} \tag{3.2}$$

For a single Higgs field this is just the thoroughly studied $O(\cdots)/Z_2$ action, where $O(\cdots)$ refers to the symmetry group of the matter field. We will concentrate on the elementary cases of the matter fields in the fundamental $O(\cdots) \to Z_2$ as well as the case of $O(2)$ matter.

We repeat, different from the single copy versions, we have not managed to identify circumstances where these replicated theories become of relevance to experiment. Perhaps the closest approach are the multi-component superconductors identified and analyzed by Babaev and coworkers [59–61]. These would correspond with two $U(1)$ matter fields sharing the $U(1)$ gauging by electromagnetism. However, the latter is supposedly to be *non-compact* lacking magnetic monopoles: as will become clear, the analogous "gauge fluxes" (or "visons") of the Z_2 version are crucial to the physics we wish to discuss.

Not knowing quite what to expect, we explored this theoretical landscape in first instance through Monte-Carlo simulations with a focus on establishing the nature of the phase diagrams. To establish the location and nature of the phase transition we employ standard methodology (thermodynamics, Binder criterium): details can be found in the appendix.

This revealed surprises that we deem of sufficient interest to report here. These are already revealed by the simplest replicated $Z_2 \times Z_2 \times Z_2 \cdots / Z_2$ version. We first consider the limit where the local couplings between the matter fields (the $J^L_{ab} = 0$ in Eq. (3.2)) are absent for identical Higgs couplings $J_a = J \; \forall \; a$. Although there are no direct couplings between the replicated matter fields, we identify a new, gauge invariant global symmetry. We call this the "registry" order parameter, associated with the *relative* orientation of the local matter fields. As we will explain in Section 3.3, for $N_{\rm rep}$ replicated fields this is governed by a $p = 2^{N_{\rm rep}}/2$ state Potts model. This symmetry is automatically broken in the Higgs phase, while it restores in the confining phase. The consequence is that, different from the single copy version, this renders the Higgs and confining phase to be distinguishable as they are now separated by a second order phase transition. The local couplings J^L_{ab} break this symmetry explicitly turning these transitions into cross-overs, while the onset of the "registry" phase transition in the gauge coupling (K) and matter coupling ($J_a = J$) can be manipulated by choosing matter couplings that are different for the copies $J_a \neq J_{b\neq a}$.

We will then turn to the $O(2) \times O(2) \times \cdots / Z_2$ case (Section 3.5). For a single matter copy the Higgs phase is distinguishable from the confining phase since the former breaks symmetry spontaneously in the form of a ("spin") nematic order parameter, involving for $O(2)$ a second order phase transition. However, invoking more than one copy we find that this is characterized by the same registry order parameter as the Z_2 case for vanishing local couplings. The ramification turns out to be that the confinement-to-Higgs (or isotropic-to-nematic) transition is lifted to a first order one for the reason that *two* global symmetries (registry and nematic) are simultaneously spontaneously broken at this transition.

Last but not least, it sheds further light on a peculiarity associated with the Z_2 gauge systems that we find to be more generic than anticipated. This departs from the

structure of the phase diagram of the Z_2/Z_2 problem, as function of gauge (K) and matter (J) coupling, Fig.3.1. As pointed out already by Fradkin and Shenker [12], by following the large J (top side) and small K (left side) evolution of the couplings no phase transition is encountered between the confining and Higgs "phases", stressing the indistinguishability. However, departing from the tricritical point associated with the deconfining phase one finds a strand of first order transition terminating at a critical endpoint. This was elucidated in full by Huse and Leibler [58], employing a dual language involving both the matter and gauge topological defects, that we will review first to set the stage (Section 3.2). This is associated with a peculiar "amplitude dynamics" and we find that it is remarkably robust, being actually rather insensitive to the presence of gauge invariant global symmetries. This can already be seen in the elementary $O(2)/Z_2$ system where it appears to have been overlooked (see Fig. 3.4) – we are aware of only one publication where the first order "strand" was mentioned without further analysis [62]. We will end with a short discussion of our findings (section 3.6).

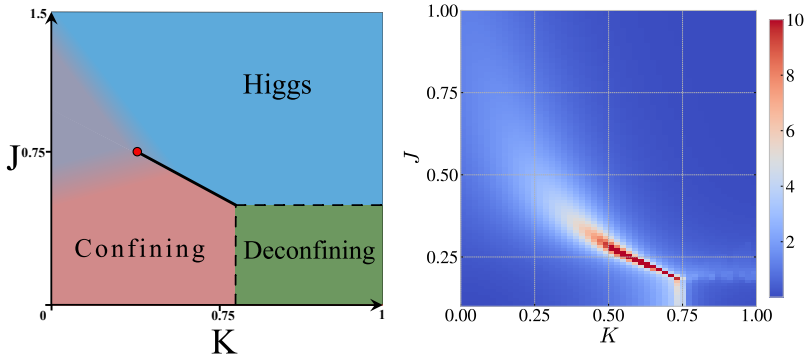


Figure 3.1: The quantitative phase diagram as established by Monte-Carlo of the Z_2/Z_2 theory as function of the gauge coupling K and matter coupling J, see appendix 3.7.1 for details. The left panel gives the overview, further illustrated by the outcomes for the specific heat in the right panel. The dashed- and drawn lines represent second- and first order transitions respectively, that we confirmed by a detailed analysis of the Binder cumulants. This is well understood (see main text): for small J the matter sector is disordered and can be ignored, and one is just dealing with the pure gauge Wegner Ising gauge theory revealing the deconfining- and confining phases separated by a continuous phase transition, where the latter can be viewed as a condensate of gauge fluxes ("visons"). Upon increasing J at large K one enters the "Higgs-like" phase that is famously indistinguishable from the confining phase – a transition is absent as function of J for small K. A peculiarity is the first order line emanating at the tricritical point anchored at the deconfinement transition. This reflects a van der Waals density driven liquid-vapor transition associated with the density of "featureless" domain walls, terminating at a critical end point.

3.2. A short review of Z_2 gauge theory with matter.

Let us first remind the reader of the thorough understanding of Wegner Ising gauge theory, both for matter in the fundamental (Z_2) and the case of matter fields with a larger symmetry than the fundamental. The pure gauge theory played a decisive role in the very early days of Yang-Mills theory by demonstrating the existence of confining/deconfining phase transitions [10] in 3 and higher (overall) dimensions, highlighted by the famous lecture notes by Kogut [11]. Among others, the pure Z_2 gauge is dual to global Ising, while the deconfining phase was much later understood as being characterized by topological order. For a particularly appealing physical interpretation see the "stripe fractionalization" [53, 54].

The essence is the invariance of the theory Eq. (3.2) under the local gauge transformations at each site i,

$$\begin{aligned} |\text{state}\rangle &\rightarrow \prod_j \sigma^x_{ij} |\text{state}\rangle \\ \phi^a_i &\rightarrow -\phi^a_i \,. \end{aligned} \tag{3.3}$$

The bond variables are Ising valued (± 1) and the action is invariant under flipping the signs of all bonds emanating from any site i when simultaneously the matter (vector) fields living on the site revers their signs. For vanishing (and by extension small) matter couplings J_a one is in the realms of the pure gauge theory (see Fig.3.1). This is best understood in terms of the topological excitations, the "gauge fluxes" [11] also called "visons" in the condensed matter literature [52]. The gauge invariant object is the Z_2 valued Wilson plaquette variable $\tau^z_{12}\tau^z_{23}\tau^z_{34}\tau^z_{41}$: for an even or uneven number of positive bond variables τ^z_{ij} this takes a positive or negative value, the latter representing gapped excitations when K is large.

These visons have a similar status as the monopoles of compact QED [63], the difference being that these fluxes have co-dimension $d-2$: in 3 dimensions these correspond with "world lines" (see Fig. 3.2). It is easy to see that a Dirac seam emanates from the line (forming a surface) and for large K these form small closed loops protecting the topological order. As for global $U(1)$ in 3D, upon reducing K these loops grow in size to "blow out" at the transition to the confining phase that can be viewed as a condensate of the "vison particles". It is easy to demonstrate [11] that the (gauge invariant) Wilson loop exhibits a perimeter law when the visons are expelled from the vacuum (deconfinement), turning into an area law in the confining phase (vison condensate). This explains the small J regime of all phase diagrams that we will present.

Let us now turn on the matter couplings J^a to focus in on the case of minimal Z_2 matter "in the fundamental". As we already emphasized, the demonstration by Fradkin and Shenker [12] that the "maximally orderly" and "maximally disorderly" Higgs

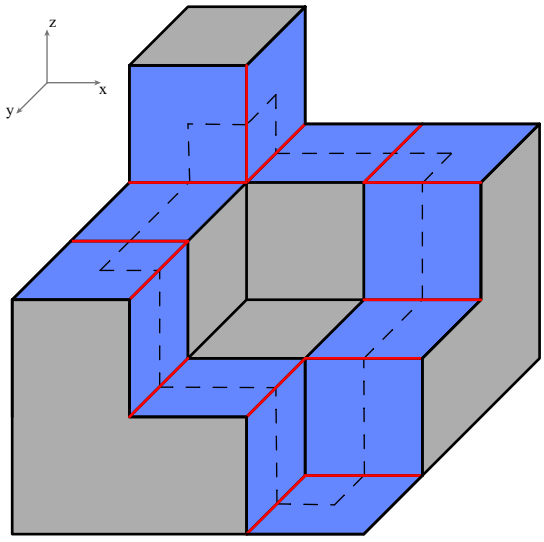


Figure 3.2: An example of a closed Z_2 gauge flux ("vison") loop as these occur in the deconfining (and Higgs "like") phase representing the topological excitation of the gauge theory. The plaquettes indicated by blue represent a gauge invariant flux $\tau^z_{12}\tau^z_{23}\tau^z_{34}\tau^z_{41} = -1$ immersed in a background of positive plaquette values. These have to form a connected line in three dimensions, the dashed line in this illustration.

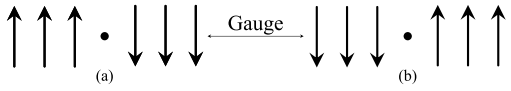


Figure 3.3: The featureless nature of the Higgs "phase" of the Z_2/Z_2 theory explained by inspecting the domain walls in the large J,K regime. Depart from the unitary gauge (upper row) and construct a domain wall costing an energy J/unit cell. Restore the gauge invariance implying that on every site an up spin cannot be distinguished from a down spin. This shows that the domain wall has the gauge invariant meaning of an object carrying energy but nothing else. This is the crucial insight behind the indistinguishability of the Higgs and confining phases.

and confining phases are actually indistinguishable caused initially confusion. But in hindsight it is obvious, for the simple reason that a phase transition requires a *global*, gauge invariant symmetry to be broken. But in the presence of single Z_2 matter, global symmetry is erased. Consider the large K,J limit; the visons are completely expelled and one can choose a unitary gauge fix taking all bond variables to be positive and the matter field living on the sites form an ordered Ising state with the spins, say, pointing up. However, according the gauge transformation Eq. (3.3) on every site this can be swapped to a down spin. This sense of Ising order has no gauge invariant meaning and a global symmetry that is broken cannot be identified, and thereby there is no distinction from the confining phase.

From tracking parameters along the large J limit varying K followed by descending

along J for small K one finds no phase transition proving the indistinguishability of the Higgs and confining phases. The oddity is however in the form of a "strand" of first order transitions emanating from the tricritical point anchored at the pure gauge confining-deconfining transition; the deconfining phase keeps of course its identity characterized by the expelled visons. This will be an important motif in the remainder: it is a peculiarity of the "amplitude sector" of these gauge theories, that was elucidated by Huse and Leibler in a particularly appealing physical setting [58].

Although there is no manifest global symmetry, there is a "material dynamics" at work that becomes obvious considering the topological excitations, see Fig. 3.3. It is instructive to depart from the large K, J limit. Consider the unitary gauge fix where one is dealing with a standard Ising model, characterized by domain walls as topological excitation costing an energy $E \sim J$/unit cell. However, upon restoring gauge invariance it "looses the symmetry": for instance one can swap all spins to the left of the domain wall from up to down and the other way around to the right. Although the matter spins "disappear", the presence of the domain wall as en energetic excitation is a gauge invariant notion. This is the simple clue; upon reducing J such featureless domain walls will start to proliferate.

Let us now decrease K such that (closed) vison loops start to form. It is easy to find out that the Z_2/Z_2 domain walls and visons relate to each other in the same way as the (Abrikosov) flux lines and magnetic monopoles of the compact $U(1)/U(1)$ theory, where the monopole "cuts open" the flux line. The vison loop "cuts open" the domain wall surface. For small J and large K in the "Higgs like" phase one finds accordingly large domain wall surfaces with here and there a small hole ("vesicles" [58]). However, for small K and finite J ("confinement like") there are many visons and accordingly the domain walls are "cut in small pieces" ("platelets" [58]).

This offers the crucial insight regarding the origin of the first order line separating the confining and Higgs like regime. The net density of domain walls takes the role of density in the van der Waals theory dealing with liquid-vapor transitions. This is just the ubiquitous affair where the density changes discontinuous in a first order transition in the pressure-temperature diagram, terminating at a critical point. Huse and Leibler argue that this particular "platelet" versus "vesicle" incarnation is literally related to the behaviour of lipid membranes as of great relevance to e.g. biology[58]. One take-home message of our work is that this peculiar "amplitude dynamics" is surprisingly ubiquitous in the whole landscape of theories defined by Eq. (3.2).

The next general motif that will be important is associated with matter field characterized by a larger symmetry than the gauge sector. In general, this will imply that a gauge invariant *global* symmetry can be identified, that is broken in the Higgs phase, restoring the distinguishability with the confining phase. The simplest example is the single copy $O(N)/Z_2$ system.

This "left over" global symmetry breaking is easy to infer in the large K, J limit. In unitary gauge one finds here an ordered state breaking the $O(N)$ symmetry. However, the gauge transformation transforms the vector into minus itself. Consider $O(3)$ in 3D: this turns the vector in the *director* order parameter of a *uniaxial nematic*. The confinement-Higgs transition becomes in turn equivalent to the liquid crystal nematic-isotropic transition. This was used by Toner et al. [56] to shed light on the origin of the well known first order character of this transition. One way is to consider small K to integrate out perturbatively the gauge fluctuations, the outcome being that one recovers the Landau-de-Gennes theory governed by a rank two traceless symmetric tensor order parameter, allowing for a cubic invariant responsible for the first order transition. However, the dual (topological) language elucidates that *two* global symmetries are now simultaneously broken by the vison condensation as well as the manifest nematic order. We notice that this gauge theory strategy was used recently to completely classify and study "generalized nematics" associated with the breaking of rotational symmetry to any of the (non-Abelian) three dimensional point groups. This is a remarkably rich affair, with the highly symmetric point groups translating in high rank tensor order parameters [57, 64, 65].

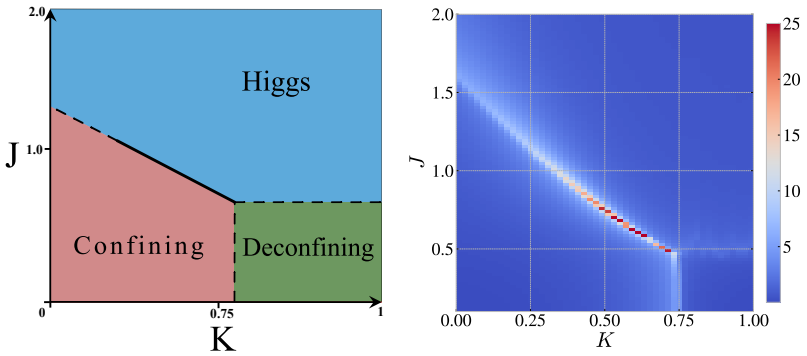


Figure 3.4: In the left panel the phase diagram of the $O(2)/Z_2$ theory, established from the Monte Carlo results for the specific heat c_V in the right panel. This can be directly compared with the Z_2/Z_2 theory, Fig.3.1. One infers that it looks very similar, the only difference being that now the "gap" between the termination of the first order line and the $K \to 0$ limit between the confining- and Higgs phase is now interrupted by a second order phase transition associated with the breaking of the gauge invariant "half-periodic" $O(2)$ symmetry. The fact that in other regards the phase diagrams look so similar is surprising, see the main text.

A final motif that will be useful in the remainder is associated with the "outlier" Abelian matter $O(2)/Z_2$ case that is also well known, especially in the context of "fractionalized" superconductivity (e.g., [52]). Different from the non-Abelian $O(N)$ cases in the small K limit the rank two tensor order parameter simplifies to a simple $O(2)$

with halved periodicity,

$$H_{\text{eff},K\to 0} = -J' \sum_{\langle i,j\rangle} \left(\vec{\Phi}_i \cdot \vec{\Phi}_j\right)^2 = -J'|\Phi|^2 \sum_{\langle i,j\rangle} \cos\left(2(\phi_j - \phi_i)\right) , \tag{3.4}$$

using $\vec{\Phi}_i = |\Phi| e^{i\phi_i}$. This has the obvious ramification that for $K \to 0$ this transition continues to be second order. In Fig.3.4 we show the phase diagram. There is now indeed a second order transition separating the Higgs and confining phases, associated with the disappearance of the "halved periodicity" XY order parameter characterizing the Higgs phase. However, upon increasing K this turns into a first order line again. Strikingly, this first order "strand" is even quantitatively very similar as its analogue in the Z_2/Z_2 case: the main difference is just that the "connection" between confining and Higgs is now closed off by the transition involving the manifest XY order parameter. This is clearly associated with the amplitude of the order parameter that obviously submits to the same "Huse-Leibler" logic, being controlled by the density of the "non-topological" matter defect "fragments" near the tricritical point.

Although it has been previously observed that close to the tricritical point the Higgs-confinement transition has turned first order [62] it took us by surprise that the $O(2)/Z_2$ behavior behaves so similarly as to the Z_2/Z_2 case: the only essential difference is that the "gap" between the critical end point and the $K \to 0$ limit is just "filled" with the nematic-to-isotropic like second order transition, barely affecting even the locus of the end point of the first order line. A priori it is not at all obvious why this is so similar. For instance, the matter topological excitations, as identified in unitary gauge, are now XY-vortices with a quantized rotation associated with the halved periodicity. These are lines (and not surfaces) in 3D, in stark contrast with the Ising domain walls of the Z_2/Z_2 theory. We will encounter underneath other variations on this theme, invariably revealing this somewhat mysterious quantitative universality of the Huse-Leibler motif.

3.3. Replicating the Z_2 matter fields: two copies.

After these preliminaries let us now turn to the main subject: the family of "replicated" Ising gauge theories. The simplest case is the Z_2 matter theory with two matter copies: $Z_2 \times Z_2/Z_2$. In fact, the most interesting case is the truly minimal one where we take the same matter field coupling $J_1 = J_2 = J$ and set the local couplings to zero $J_L^{12} = 0$. Under these circumstances the matter fields couple only through the Ising gauge fields.

Let us start with the simplest example of a Z_2 gauge theory with replicated matter: the $Z_2 \times Z_2/Z_2$ case. In Fig.3.5 we show the phase diagram as established by our Monte-Carlo simulations. We infer that this is a very close sibling of the $O(2)/Z_2$ phase diagram that we just discussed Fig.3.4. Next to the ubiquitous Higgs-

deconfining transition, the transitions between the Higgs- and confining phase looks very similar, including the first order line emanating from the tricritical point turning second order roughly at the locus of the Z_2/Z_2 critical endpoint.

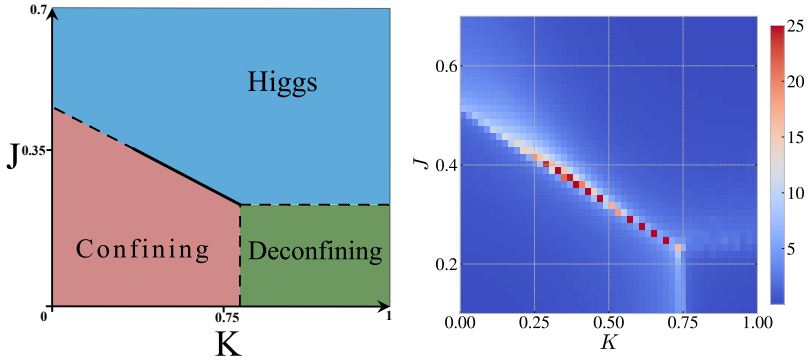


Figure 3.5: The phase diagram of the "two copy" $Z_2 \times Z_2/Z_2$ theory for $J_1 = J_2 = J$ and $J_L^{12} = 0$ illustrated by the Monte Carlo results for the specific heat c_V in the right panel. This looks very similar as to the $O(2)/Z_2$ case (Fig.3.4) in turn similar to the Z_2/Z_2 case except for the connection between Higgs- and confinement is closed off by an honest second order transition. As explained in the text, one can now identify an Ising valued gauge invariant "registry" order parameter that breaks the symmetry spontaneously in the Higgs phase having a similar role as the nematic order parameter of the $O(2)/Z_2$ case.

What is going on here? As explained earlier, for the transition between confining and Higgs at low K to be a true second order phase transition, there has to be a broken global symmetry. In fact this is indeed controlled by a gauge invariant order parameter with a global Ising symmetry that may not be directly obvious to the reader, but it is quite simple. Consider again the unitary gauge with all bond spins +1. Because we have two matter fields, we are now dealing with two *independent* Ising spin systems living on the sites that will both be ordered for large J, K. As we emphasized in Section 3.2, after restoring the gauge invariance both Ising spin systems "loose their symmetry" according to the Z_2/Z_2 rule book. But now we observe that the relative orientation of these two spin systems actually corresponds with a gauge invariant, global Z_2 symmetry!

Take the matter spins "1" to be pointing in the positive direction, and the "2" spins can be either parallel or anti-parallel to the "1" spins. Let us now see what happens with this "registry" upon restoring the gauge invariance, see Fig.3.6. Consider parallel registry on a particular site like $\uparrow_1\uparrow_2$ and under a gauge transformation both spins flip – a single sector $\uparrow, \downarrow$ is not gauge invariant. But it follows that the *relative* orientation of the matter spins in both sectors can be either *parallel* or *anti-parallel* and after restoring the gauge invariance it continues to be distinguishable whether one is dealing with locally parallel or anti-parallel configurations: $\downarrow_1\downarrow_2 \leftrightarrow \uparrow_1\uparrow_2$ versus $\uparrow_1\downarrow_2 \leftrightarrow \downarrow_1\uparrow_2$! This is what we call the "registry order parameter" which carries clearly

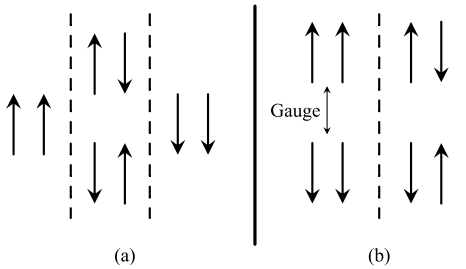


Figure 3.6: The "registry" order parameter of the two copy $Z_2 \times Z_2 / Z_2$ theory. Depart from the large J, K limit and take a unitary gauge fix as in Fig.3.3. One can now construct two types of domain walls associated with the two copies (upper line). However, upon restoring gauge invariance only two of the four on-site configurations can be distinguished: the spins are locally either parallel- or anti-parallel. This is the global Z_2 symmetry that is broken in the Higgs phase.

a global Z_2 charge!

This "registry" symmetry breaks spontaneously in the Higgs phase, causing a two fold degenerate ground state: either the "parallel" or "anti-parallel" registry takes over. This registry order parameter $\langle \phi^1 \phi^2 \rangle$ is easy to measure and it is precisely what we find in the Monte-Carlo simulation of the Higgs phase, see Fig.3.7.

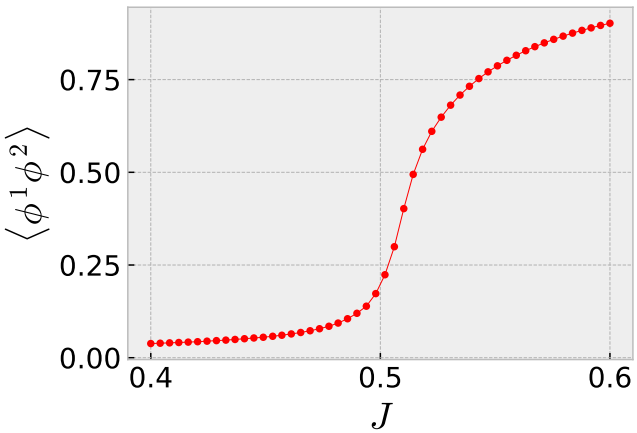


Figure 3.7: Expectation value of the registry order parameter $\langle \phi^1 \phi^2 \rangle$ of the $Z_2 \times Z_2 / Z_2$ theory, along the $K = 0$ slice.

At first sight this may be a bit confusing given that the matter fields interact only via the gauge couplings. However, the simple logic in the previous paragraph just reveals that this replicated system has in the Higgs phase a doubly degenerate ground state associated with "parallel" and "anti-parallel" registry. It is easy to construct gauge

invariant domain walls between domains with opposite registry, as the "featureless" domain walls of the Z_2/Z_2 theory now acquire the gauge invariant meaning that they represent a jump in the registry order. Notice that the phase diagram is in all regards other than the second order Higgs-confinement transition a near quantitative copy of the Z_2/Z_2 phase diagram. Given the lessons of the $O(2)/Z_2$ theory this is perhaps not surprising since the "microscopy" of the $Z_2 \times Z_2/Z_2$ theory is a close cousin of the $O(2)/Z_2$ case.

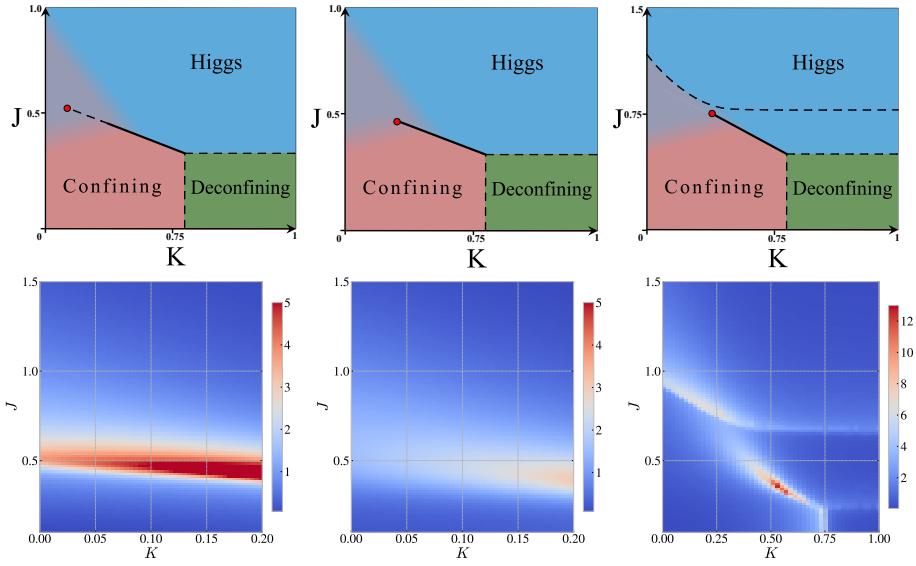


Figure 3.8: The phase diagrams of the $Z_2 \times Z_2/Z_2$ theory, with the Monte Carlo data for c_V in the second row, for a finite $J_L = 0.1$ (**a**) and $J_L = 1$ (**b**). Notice that the Monte Carlo results are shown over a small interval in K to highlight the most relevant changes. This local coupling acts as a field breaking the registry symmetry explicitly, turning the $J_L = 0$ Ising phase transition into a cross over. The consequence is that for any finite J_L the indistinguishability of Higgs and confinement is restored. The small strand of second-order like (dashed line) transition for small J_L (**a**) may well be related to a rapid crossover that can not be distinguished within our numerical accuracy (Binder criterium) from a real transition. For large J_L (**b**) the phase diagram becomes nearly identical to the one of the Z_2/Z_2 theory itself. Another interesting game is to vary the relative strength of the matter coupling of the two copies for $J_L = 0$. In (**c**) we show a typical example dealing with different matter couplings $J = J_1 = 3J_2$.

To illustrate these matters further, let us consider what happens when the local J_L^{12} interaction in Eq. (3.2) is switched on. It is immediately obvious that this relates directly to the registry: one sees immediately departing from unitary gauge that this lifts the degeneracy of the parallel and anti-parallel registry configurations. This acts as a field breaking the registry Ising symmetry explicitly! This should have the effect to turn the registry phase transition into a cross over, with the effect that yet again the confining and Higgs phases become indistinguishable again, and this is what we find is going on according to our MC simulations: see Fig.3.8 (**a**,**b**).

A next freedom one can exploit is to take $J_L = 0$ but change instead the relative magnitude of the matter couplings, $J = J_1 \neq J_2$. This is an entertaining affair highlighting the unusual nature of the registry order. Upon reducing J_2 the locus on the vertical axis where this second field changes from order to disorder shifts upwards along the vertical axis in terms of the "dominating" copy governed by $J_1 = J$. The effect is that at the J_1 transition in the small K regime the J_2 coupled matter field is still *disordered* while the order in both fields is required for the existence of the registry order! Hence, holding $J_1 \neq J_2$ fixed while varying K, there is a window where the registry order disappears in the small K regime, switching on again when K has become sufficiently large to reach the J where also the second field becomes also prone to order again, see Fig.3.8(**c**). The outcome is a critical end point where now a line of second order transitions starts that is subsequently turned into the first order Huse-Leibler affair! We notice that this peculiar behaviour is only seen in a small J_2/J_1 interval; for $J_2 \leq 0.25J$ the phase diagram becomes again the one of the single copy Z_2/Z_2 theory.

Up to this point we have demonstrated that in terms of the degrees of freedom of the gauge theory a quantity can be identified characterized by a global symmetry that is broken in the Higgs phase – the registry order parameter. However, what is the nature of the *dynamics* responsible for the stability of this order? Inspecting this deep in the Higgs phase (large K, J) employing the unitary gauge is not informative. The reason is that this is rooted in "gauge field interactions" that are unusual in the sense that the discrete nature of the Z_2 gauge fields implies that these "gauge forces" are characterized by a mass scale that becomes large deep in the Higgs phase. In a statistical physics language any gauge field mediated force may be viewed as an "order-out-of disorder" phenomenon – the fluctuations of the gauge fields are responsible for the interactions between the (gauge invariant) matter fields.

Hence, the limit to consider is $K = 0$: the visons are the dynamical degrees of freedom of the discrete gauge theory and these come for free in this limit. Given that the disclinations are bound states of matter defects (domain walls) and the visons, the stability of the Higgs phase itself is entirely due to the cost associated with the former, while the Z_2 gauge field is maximally fluctuating. Its only energy cost comes from the coupling to the matter fields. The gauge fields can therefore be straightforwardly integrated out with the outcome that one obtains the (gauge invariant) Landau-de-Gennes order parameter theory [56]. In the general case one is dealing with the point groups associated with the rotational symmetry of the nematic-type state; for Abelian point groups in two dimensions that are encoded by Z_N gauge fields one recovers in this way the simple "p-adic" nematics [66], while this procedure has been shown to be instrumental to derive the high rank tensor de-Gennes order parameters associated with the non-Abelian point groups in three dimensions [57, 64, 65].

The outcome for the elementary Z_2 gauge theory is simple: integrating out the gauge fields when $K = 0$ leads to the simple "square of the Hamiltonian" gauge invariant effective theory as in Eq. (3.4) [56]. In full generality, departing from the replicated theory with arbitrary matter field symmetry Eq. (3.2) when $J_L^{ij} = 0$ and all $J_i = J$ one obtains,

$$H_{\text{eff},K=0,} = -J' \sum_{\langle i,j \rangle} \left(\sum_{a=1}^{N_{\text{rep}}} \vec{\phi}_i^a \cdot \vec{\phi}_j^a \right)^2 . \tag{3.5}$$

Let us first consider the single replica Z_2 matter field. Here $\vec{\phi}_i \to \sigma_i^z$ and we infer that the effective Hamiltonian becomes $\sum_{\langle i,j \rangle} (\sigma_i^z \sigma_j^z)^2 \to$ constant: this is the essence of the Fradkin-Shenker observation [12], no gauge invariant degree of freedom can be identified distinguishing the Higgs and confining phases implying that these are indistinghuishable. But let us now consider two identical Z_2 copies,

$$\begin{aligned} H_{\text{eff},K=0,N_{\text{rep}}=2} &= -J' \sum_{\langle i,j \rangle} \left(\sum_{a=1}^{2} (\sigma^z)_i^a \cdot (\sigma^z)_j^a \right)^2 \\ &= \text{constant} - J' \sum_{\langle i,j \rangle} \left((\sigma^z)_i^{(1)} (\sigma^z)_i^{(2)} \right) \left((\sigma^z)_j^{(1)} (\sigma^z)_j^{(2)} \right) . \end{aligned} \tag{3.6}$$

This simple affair reveals the origin of the "registry dynamics": the combination $\left((\sigma^z)_i^{(1)} (\sigma^z)_i^{(2)} \right)$ takes the (global) Z_2 values ± 1 for parallel and anti-parallel registry and Eq. (3.6) is just an Ising Hamiltonian associated with the registry degrees of freedom.

In hindsight this is elementary. As for the uniaxial nematics of Toner *et al.* [56], the gauge theory in the strong coupling regime ($K \to 0$) is in fact a redundant parametrization of the "director" gauge invariant de Gennes type theories. The additional richness of the gauge theory is associated with K becoming large, i.e. the weak coupling regime of the gauge theory. For instance, the topologically ordered *deconfining* phase has no physical identification dealing with the "molecular" nematic liquid crystals. Additional microscopic structure is required, with perhaps the "stripe fractionalization" [53, 54] being the most elementary example of how this can happen.

This is also underlying the difficulty to recognize this simple motif "deep" in the Higgs phase, for large J and $K \to \infty$. Departing from the unitary gauge one easily identifies the registry as gauge invariant degree of freedom (as in the above) but at first sight the dynamics stabilizing it is obscure. The reason is of course that the visons are now highly energetic excitations requiring an energy $E \sim K$ per unit length associated with the effect that the gauge symmetry is discrete and the gauge fields are massive . However, there is no phase transition in the Higgs phase at large J, varying K from zero to infinity. This implies that regardless the magnitude of the virtual visons their fluctuations *always* suffice to hard wire the registry order. This may be viewed as an order-out-of-disorder phenomenon pushed to its extreme.

3.4. The case of many Z_2 matter fields.

Having established the rules for two Z_2 matter fields copies, how does this generalize to many copies? Let us depart again from Eq.(3.2) for Z_2 matter and consider an arbitrary number of matter field copies $N_{\rm rep}$. As before the circumstances optimal for the registry type order are associated with setting all local couplings to vanish, $J_L^{ab} = 0$, and taking the matter fields couplings to be equal: $J_a = J$. It is in fact easy to find out by induction how the registry spontaneous symmetry breaking of the two copy case generalizes to many copies.

Let us consider the three copy case: $Z_2 \times Z_2 \times Z_2/Z_2$. We just proceed as before, departing from deep in the Higgs phase (large K, J) and using the unitary gauge. At every site the matter fields can occur in 2^3 different configurations, see Fig.3.9. Identifying these with domains one would find accordingly 8 different domain walls associated with flipping one spin keeping the other spins fixed. However, upon restoring the gauge invariance amounting to flipping *all* spins on the site, it follows that configurations are pairwise associated, e.g. $\uparrow\uparrow\uparrow \leftrightarrow \downarrow\downarrow\downarrow$. Accordingly, one finds 4 distinct gauge invariant vacuum states, separated by "single spin" domain walls. This is governed by a $p = 4$ state *Potts model*! It is easy to check that the Ising registry theory Eq.(3.6) generalizes to the 4-state Potts theory by considering the "squared" de-Gennes type effective theory in the $K \to 0$ limit.

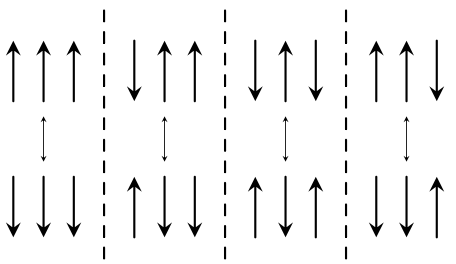


Figure 3.9: The registry order parameter (see main text) in the case of the three copy $Z_2 \times Z_2 \times Z_2/Z_2$ theory. Depart from unitary gauge fix and the three matter fields can locally form eight configurations. However, under gauge transformations half of them are redundant, e.g. $\uparrow\uparrow\uparrow \leftrightarrow \downarrow\downarrow\downarrow$. The gauge invariant registry order parameter takes therefore four different physical realizations and it is governed by a $p = 4$ state Potts model.

This is confirmed by our MC simulations. In Fig.3.10 we show the phase diagram of the three copy model. As anticipated, this looks very similar as to the two copy case, Fig.3.5. Although the dynamics is quite different — the 4 distinct "gauged" domain walls associated with $p = 4$ Potts — the Huse-Leibler first order strand is barely affected with yet again the registry order being responsible for the continuous (Potts model) phase transition distinguishing the confining and Higgs phases. In Fig.3.11 we show a typical realization after a partially annealed Monte-Carlo "quench" deep in

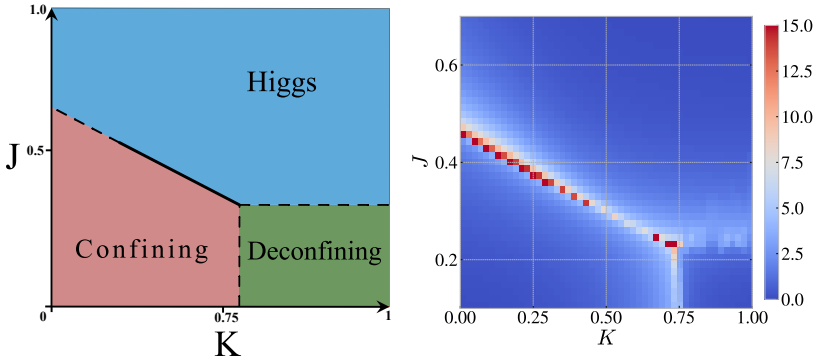


Figure 3.10: The phase diagram of the three copy $Z_2 \times Z_2 \times Z_2/Z_2$ theory as in Fig.3.1 and Fig.3.4 for the "minimal" $J_a = J$ and $J^L_{ab} = 0$ case. This looks very similar to Fig.3.5 and the strand of the first order transitions is barely changed. The specific heat c_V data suggests a first order transition all the way down to $K = 0$, but a Binder cumulant study reveals this change to second order as sketched. A major distinction now is that dynamics of (continuous) confining to Higgs transition is governed by a $p = 4$ state *Potts model*.

the Higgs phase: one discerns the 4 distinct Potts domains separated by the "registry domain walls" confirming this simple analysis.

As for the two copy case one can now proceed by switching on various J^{ab}_L local couplings, "gluing together" the local matter copies with the expected results that we checked. Involving one coupling between two of the three sectors, the 4-state Pott symmetry is lifted by the explicit symmetry breaking to the effective two copy registry Ising symmetry. Similarly, coupling all copies with each other diminishes the registry spontaneous symmetry breaking such that confinement and Higgs become indistinguishable. In the same guise one can detune the matter couplings J_a, turning into a variation of the matters we discussed in the previous section.

The two and three copy cases reveal the counting rules and by induction we can generalize this now to an arbitrary number copies. We just proceed as for the three copy case. Given $N_{\rm rep}$ copies there are a total of $2^{N_{\rm rep}}$ local configurations in unitary gauge. The next observation is that these configurations are pair wise identified with each other by the gauge transformation. The result is that the registry symmetry is now captured by a $p = 2^{N_{\rm rep}}/2$ state Potts model.

3.5. Replicating $O(2)$ matter.

The "matter in the fundamental" Z_2 matter case is special, and what to expect when the symmetry of the matter field is raised relative to the Z_2 gauge symmetry? For this purpose we focused in on the minimal extension: the replicated $O(2)$ matter fields

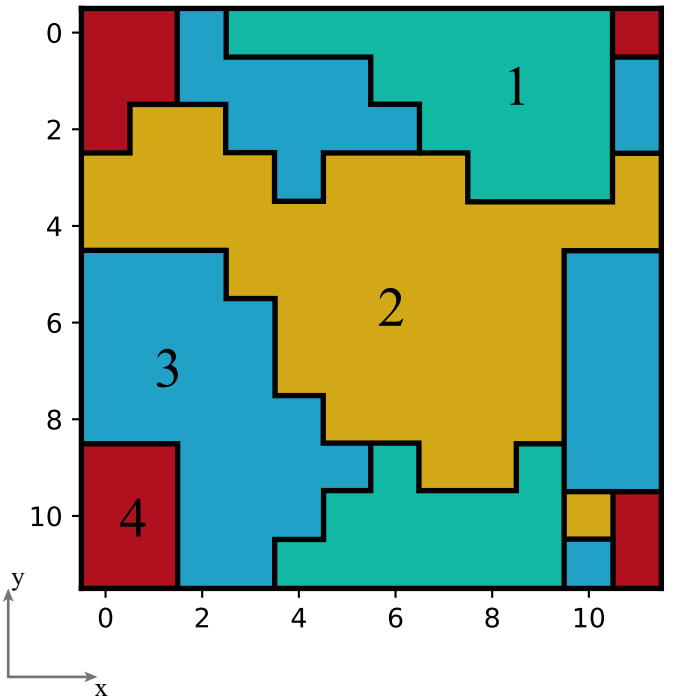


Figure 3.11: Planar snapshot along the z axis of a Monte Carlo quench for the three copy model deep in the Higgs phase $J_1 = J_2 = J_3 = 0.8$ and $K = 0$. This reveals the presence of the 4-state Potts model domains of Fig.3.9 separated by the "registry" domain walls. Domain 1 - are $\uparrow\uparrow\uparrow\leftrightarrow\downarrow\downarrow\downarrow$, domain 2 - are $\downarrow\uparrow\uparrow\leftrightarrow\uparrow\downarrow\downarrow$, domain 3 - are $\downarrow\uparrow\downarrow\leftrightarrow\uparrow\downarrow\uparrow$, domain 4 - are $\uparrow\uparrow\downarrow\leftrightarrow\downarrow\downarrow\uparrow$. Every domain wall differs from the neighboring domain by a single flip.

gauged by Z_2. As we discussed in Section 3.2, for a single copy the Higgs phase is now characterized by the nematic-like order, distinguishing it from the confining phase through the presence of a second order phase transition.

In Fig.3.12 we show the phase diagram of the two copy $O(2) \times O(2)/Z_2$ case for the (usual) choice $J = J_1 = J_2, J_L^{12} = 0$. This looks yet again very similar as to the other cases, the main difference being that now the transition between the Higgs and confining phase has turned into a *first order* transition for *all* K. Inspecting the "strength" of the first order transition exploiting the Binder criterium (see Appendix 3.7.2) we find that close to the tricritical point the transition looks quite like the other cases: this is clearly driven by the "Huse-Leibler" amplitude fluctuations. Upon reducing K the transition becomes an increasingly weak first order transition, but it continues to be first order all the way down to $K \to 0$.

As we will argue, this first order behavior is due to the fact that *two order parameters*

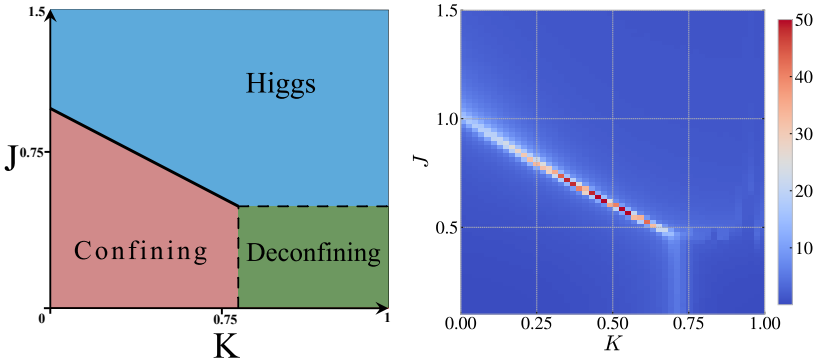


Figure 3.12: The phase diagram of the $O(2) \times O(2)/Z_2$ (left) and raw MC data for c_V (right) for $J_1 = J_2 = J$ and $J_L^{12} = 0$. This is yet again remarkably similar as to the other phase diagrams. The difference is that the Higgs-confinement phase transition is now first order all the way to $K = 0$ contrasting with its second order behavior for the single copy version, Fig.3.4. This due to the occurrence of a Ising type registry order parameter in the Higgs phase which is of the same kind as for the Z_2 matter cases, that disappears simultaneously with the nematic type order parameter upon entering the confining phase. The first order nature is confirmed by a Binder cumulant study (Fig.3.15**(c)** in Appendix 3.7.2).

governed by two independent global symmetries vanish simultaneously at the Higgs to confinement transition. In fact, the replicating has the effect that the "accidental" second order nature of this transition for $O(2)/Z_2$ becomes similar to the generic first order transition of the $O(N)/Z_2$ system with $N \geq 3$ as argued by Ref. [56].

What are the two symmetries that are broken in the Higgs phase? This is actually a bit more of a subtle affair than for the simple Z_2-replica's. As we will argue and confirm with the Monte Carlo, one type of symmetry is associated with the nematic type "halved periodicity" XY as for a single $O(2)/Z_2$, actually applying to both copies individually. But these are non-locally coupled together by the gauge fluctuations in a way that they submit to a perfect Z_2 registry symmetry, that is macroscopically the same symmetry as the registry Z_2 of the $Z_2 \times Z_2/Z_2$ theory.

This is yet again easy to deduce by zooming in on the maximal gauge fluctuations, the $K = 0$ case. As discussed in Section 3.3, this is of the universal form Eq.(3.5). For the two copy $O(2) \times O(2)/Z_2$ case it follows immediately,

$$\begin{aligned} H_{O(2)\times O(2),K\to 0} &= -J' \sum_{\langle i,j\rangle} \left(\vec{\Phi}_{i1}\cdot\vec{\Phi}_{j1} + \vec{\Phi}_{i2}\cdot\vec{\Phi}_{j2}\right)^2 \\ &\sim -J' \sum_{\langle i,j\rangle} \left((\vec{\Phi}_{i1}\cdot\vec{\Phi}_{j1})^2 + (\vec{\Phi}_{i2}\cdot\vec{\Phi}_{j2})^2 + 2(\vec{\Phi}_{i1}\cdot\vec{\Phi}_{j1}) \times (\vec{\Phi}_{i2}\cdot\vec{\Phi}_{j2})\right) . \end{aligned} \tag{3.7}$$

The first and second term just represent the director order parameter – for $O(2)$ just the halved periodicity – and upon ignoring the last term one is just dealing with two

completely decoupled identical "$O(2)$ nematics". This last term encapsulates the interactions between these two copies as induced by the fluctuating disclinations. It is obvious that also this term is governed by an invariance under $O(2)$ rotations of every copy separately – there is surely no effective single site "anisotropy" at work reducing it to the on site Z_2 registry revealed by Eq. (3.6) of the $Z_2 \times Z_2 / Z_2$ case.

However, the registry is now hidden in the "synchronization" imposed by the gauge fluctuations associated with the relative orientation of the two spin configurations on neighbouring sites. Parameterize $\vec{\Phi} = |\Phi| e^{i\phi}$, such that $\phi_{ja} = \phi_{ia} + \nabla_{ij}\phi_a$ and the interaction term becomes,

$$-J' \sum_{\langle i,j \rangle} (\vec{\Phi}_{i1} \cdot \vec{\Phi}_{j1}) \times (\vec{\Phi}_{i2} \cdot \vec{\Phi}_{j2}) = -J' |\Phi|^2 \sum_{\langle i,j \rangle} \cos\Big(\nabla_{ij}\phi_1\Big) \cos\Big(\nabla_{ij}\phi_2\Big) \,. \tag{3.8}$$

This interaction term is governed by a global Z_2 symmetry! It originates in the synchrony of gauge invariant "remnants" of the two copies: the interaction term is minimal either for *both* copies being parallel on neighbouring sites or *both* antiparallel where $\nabla_{ij}\phi_a = 0$ or $\nabla_{ij}\phi_a = \pi$. This signals the two fold degeneracy of the Ising order. In Fig.3.13 we illustrate how to construct the Ising domain walls associated with this registry order.

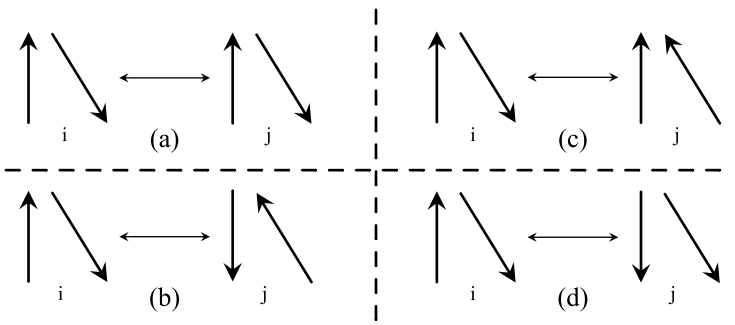


Figure 3.13: Illustration of the registry order of the $O(2) \times O(2)/Z_2$ theory: the construction of the registry Ising domain wall. We see that in contrast with the multiple Z_2 matter fields — the *local onsite* orientation of different matter fields is irrelevant for the registry. The two copies acquire independent orientations on the $O(2)$ (half) circle. What matters is how two matter fields change *together* from site to site. The strong gauge field fluctuations "glue" together matter fields in a sense that they would have to simultaneously change together in the same direction, see Eqs.(3.7),(3.8). In all cases **(a)**-**(d)** we depart from the same unitary gauge two spin reference configuration on site i. In **(a)** we just consider the ground state configuration on site j in unitary gauge. By performing a gauge transforming on site j we obtain **(b)**, gauge equivalent to **(a)**. But now consider **(c)**, with an anti-parallel orientation of the second spin on site j, gauge equivalent with **(d)**. The case **(a)**, **(b)** corresponds with $\nabla_{ij}\phi_1 = \nabla_{ij}\phi_2 = 0$. We see that these configurations minimize the interaction energy Eq.(3.8). While **(c)**, **(d)** instead has $\nabla_{ij}\phi_1 = 0, \nabla_{ij}\phi_2 = \pi$. Compared to **(a)**-**(b)** this would cost us $2J'$ energy. The configuration **(a)**-**(b)** corresponds with a registry domain wall between *parallel* and *anti-parallel* orientations of $O(2)$ matter directors at the neighboring sites.

This can be illustrated by a Monte-Carlo quench from "high temperature", departing

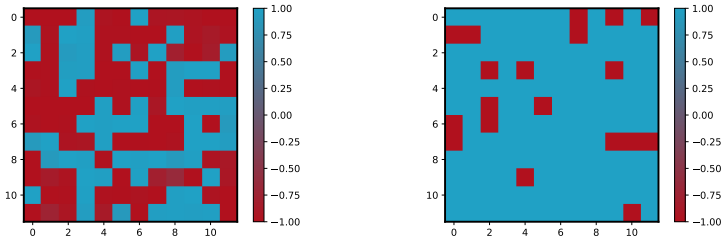


Figure 3.14: 2D snapshot along z axis of a system on $12 \times 12 \times 12$ lattice with periodic b.c for $O(2) \times O(2)/Z_2$ model. We a single frozen configuration of fields computed using a Monte Carlo quench deep in the Higgs phase: $J_1 = J_2 = 5$ and $K = 0$. **(Left)** Snapshot of $\cos\left(\nabla_{ij}\phi_1\right)$. This elucidates a simple registry order (parallel/anti-parallel), in the form of registry domains that are frozen in. The color coding denotes the cosine of the relative orientation of the spins on the neighboring sites: Domain 1 (parallel) - corresponds with $\uparrow\uparrow \leftrightarrow \downarrow\downarrow$, and domain 2 (anti-parallel) - accordingly $\downarrow\uparrow \leftrightarrow \uparrow\downarrow$. **(Right)** Snapshot of $\exp(2i\phi_1)$ director order parameter showing homogeneous distributions, being annealed, with some local fluctuations due to the finite size and temperature effects.

from random configurations and partially annealing the system in the Higgs regime, similar to Fig.3.11. This is shown in Fig.3.14. One infers that one form of order is of the nematic-type (XY like) associated with the orientation of the director of one of the the copies that we find to be homogeneous in this snapshot – this is annealed. However we can also track the *relative orientation of a single field between neighboring sites* $\nabla_{ij}\phi_\alpha$, for $\alpha = 1$. Fig.3.14 (Left) shows that this is either parallel $\nabla_{ij}\phi_1 = 0$ or anti-parallel $\nabla_{ij}\phi_1 = \pi$ as illustrated in Fig.3.13 and explained after the text. As in Fig.3.11 we observe domains of the two different parallel and anti-parallel registry separated by "registry domain walls" of the kind similar to the Z_2 matter.

Similar to the $Z_2 \times Z_2/Z_2$ case, this registry order is critically dependent on the J_L^{12} being zero – upon switching it on this acts as an explicit symmetry breaking of the registry turning the first order transition near $K = 0$ into second order, exhibiting a phase diagram that is like the single copy $O(2)/Z_2$ case (Fig.3.4). The effect of unbalancing the matter couplings ($J_1 \neq J_2$) has very similar effects as illustrated in Fig.3.8, although now a second order transition is left behind when the registry order switches off.

In summary, the "registry-sector" of the $O(2)$ case behaves more or less identically as in the Z_2 case. The difference is that in the $O(2)$ case one also has to account simultaneously for the nematic type order characterizing this Higgs phase. This renders the phase transition first order all the way to $K = 0$. The topological defects of this nematic-like state will start to populate the vacuum upon lowering K, J. These are the disclinations, in turn being a "confined" combination of the vortex-type matter defect, and the Z_2 gauge flux/vison, e.g. [56, 62]. The matter-vortices of the two copies share a single gauge flux. Upon integrating out these "topological fluctua-

tions", an interaction mediated by the Z_2 gauge fields develops which is responsible for the registry order. Similar to the Z_2 case, the limit where one can easily deduce the effects of the visons "gluing" the copies into registry, this is most easily deduced in the $K = 0$ limit which is entirely controlled by the matter interaction J.

3.6. Discussion and conclusions

Gauge field theory is of course well known to have its own rules. In this paper we have focused in on the simplest of all gauge symmetries, the Z_2 variety, as the simplest theory revealing the characteristic phase structure characterized by confinement, deconfinement and the Higgs phase. By introducing the matter replicas we discovered a new set of phenomena. In this pursuit we have heavily leaned on the unbiased Monte-Carlo simulations. Puzzled by the outcomes we discovered the new phenomenon of "registry order parameter". As we discussed in Sections 3.3 and 3.4, it is very easy to identify the origin in the Z_2 matter versions, although it is a bit less obvious and arguably more entertaining for the $O(2)$ version in Section 3.5. It took us by surprise, given that the origin of the induced "gauge interaction" that is responsible for the registry symmetry breaking is of a kind that is rather unfamiliar.

The mechanism is revealed by considering the extreme strongly coupled limit of the gauge fields, $K = 0$. In terms of the degrees of freedom of the gauge theory, the mechanism is unusual, highlighted by the $O(2)$ case. The physical degrees of freedom associated with the disordering of the Higgs phase are the nematic-type disclinations that are in turn confined combinations of matter vortices and the fluxes of the gauge field (the visons). Although both matter fields carry their own vortices, these "share" a single vison. In the $K = 0$ limit the latter come for free and upon integrating these out one finds that the de-Gennes type effective order parameter theory is endowed with the registry "Ising" Hamiltonian, Eq . (3.6), and Eq.(3.8) respectively. Interestingly, the discrete nature of the gauge theory has eventually the effect to generate the Ising type registry global symmetry breaking.

We have only inspected the most elementary forms of such gauge theoretical systems. These are just a point in the vast landscape of all gauge theories, up to the non-Abelian Yang-Mills theory behind e.g. the Standard Model. It would be quite interesting to find out what happens with this "registry order" upon systematically raising the symmetries involved. What happens in the "replicated" Z_2 gauge theory involving the non-Abelian $O(N)$ matter fields with $N \geq 3$? What happens raising the Z_2 gauge symmetry to the non-Abelian 3D point group symmetries as in Ref. [57]? Even more fundamental, what is the fate of registry dealing with *continuous* gauge symmetry, starting with the Abelian $U(1)$ of compact electrodynamics [63]?

Finally, another aspect also caught us by surprise in the Monte Carlo outcomes for

the various phase diagrams. The Huse-Leibler mechanism for the first order transition between the "Higgs" and "confinement-like phase" emanating from the tricritical point appears to be surprisingly universal. Eventually this is a quantitative affair. The mechanism as understood for the Z_2/Z_2 case does in this regard rely on the specifics of this particular theory: the "vesicles" versus "platelet" affair. This surely works differently involving continuous symmetry – the $O(2)$ cases. But surprisingly the "first order strand" is even quantitatively barely affected by these fundamental differences. The reason for this is presently unclear to us and it may be of interest to have a closer look at the origin of this "quasi-universality".

Acknowledgments

We thank T. Senthil and K. Liu for discussions. This research was supported in part by the Dutch Research Council (NWO) project 680-91-116 (*Planckian Dissipation and Quantum Thermalisation: From Black Hole Answers to Strange Metal Questions.*), and by the Dutch Research Council/Ministry of Education. The numerical computations were carried out on the Dutch national Cartesius and Snellius national supercomputing facilities with the support of the SURF Cooperative as well as on the ALICE-cluster of Leiden University. We are grateful for their help.

3.7. Appendix

3.7.1. Monte Carlo Simulations

The Monte Carlo simulations were performed on a $d \times d \times d$ grid with periodic boundary conditions. Grid size in all simulations was $d = 12$ in order to avoid finite size effects and still get reasonable computational times. We used most of the time a number of measurement sweeps $N = 6000000$; thermalization sets in typically after 1/3 of the sweeps. Near the critical points we checked this by tracking the evolution of the various quantities as function of the number of steps, taking as many steps as needed for the quantity to saturate. For the updating rules we used the Metropolis-Hastings algorithm with the acceptance ratio $A(n,n') = \min(1, e^{-\Delta E_{n,n'}})$ where $\Delta E_{n,n'}$ is the energy difference between states n and n' that differ in a single matter field or gauge field flip. Phase diagrams were obtained by vertically scanning along different values of J's, using annealing order to improve convergence accuracy. The whole phase diagram was run on a remote cluster where each process was associated with a single value of K scanning along the J axis. We noted that the longest sweeps were required to equilibrate when deep in the deconfining regime. Flipping single "bond spins" using the Metropolis-Hastings algorithm leads to highly energetic configurations and accordingly to long thermalization times. But this did not pose any difficulty since

the physics in this regime is simple.

3.7.2. Determining order of phase transition

We used several quantities in order to determine the order of a phase transition. For the $O(2)$ matter fields we tracked the local nematic magnetization $m = \langle |e^{2i\theta_i}| \rangle$ and local registry order parameter $R = \langle \theta_i^1 \cdot \theta_i^2 \rangle$. For Z_2 matter fields we only encountered the local registry order parameter. Note that formally the $O(2) \times O(2)/Z_2$ registry order parameter is non-local in that it is the nearest neighbor difference. In practice this implies also a local order parameter, which is easy to understand after performing a "block-spin" averaging RG-step. We also measured the specific heat computed as $C_V = \frac{1}{d^3}(\langle E^2 \rangle - \langle E \rangle)$. In order to see the exact point of phase transition we employed the Binder ratio defined as:

$$U = \frac{1}{2}\left[3 - \frac{\langle m^4 \rangle}{\langle m^2 \rangle^2}\right] . \tag{3.9}$$

In the $\lim_{T\to 0} U = 1$ while for $\lim_{T\to\infty} U = 0$. Since this ratio is dimensionless, plotting curves of different sizes clusters their intersection point which represents the exact value of the phase transition. In our case shape of the Binder curve is more important than the exact intersection point. A smooth transition from 0 to 1 in a sigmoid fashion indicates a second order transition while a sharp dip that diverges with system size indicates a first order behavior. Especially the transition associated with the small K regime of the $O(2) \times O(2)/Z_2$ is quite weakly first order, and this is manifested by a Binder ratio dip that does not diverge with the system size, see e.g. and Fig.3.15((c)). When going to larger system sizes, dip in the Binder curve become very narrow. In that case, refining values of J in that regions helps capture it, otherwise the peak is easily missed.

Besides the Binder ratio we also inspected histograms of values occurred during the run of a simulation for an average spatial order parameter and the local energy in order to observe the detailed behavior around the critical point where more than one peak signals the phase separation associated with first order transitions. Histogram are created by counting the occurrences of observed value in specific predetermined bins. Even with fluctuations happening during the run of a simulation, these graphs reveal where are the points of most concentrations around which quantity varies. In most simulations size of the single bin was 10^{-5} in single unit of observed quantity, which helps in a resolution of very closely placed peaks. As an example of usefulness of distribution histograms in determining the order of a phase transition we present "registry" and energy distributions for $Z_2 \times Z_2/Z_2$ model for single value of $K = 0.55$ in the regime of "Huse-Leibler" first order line emanating from the tricritical point, Fig.3.16.

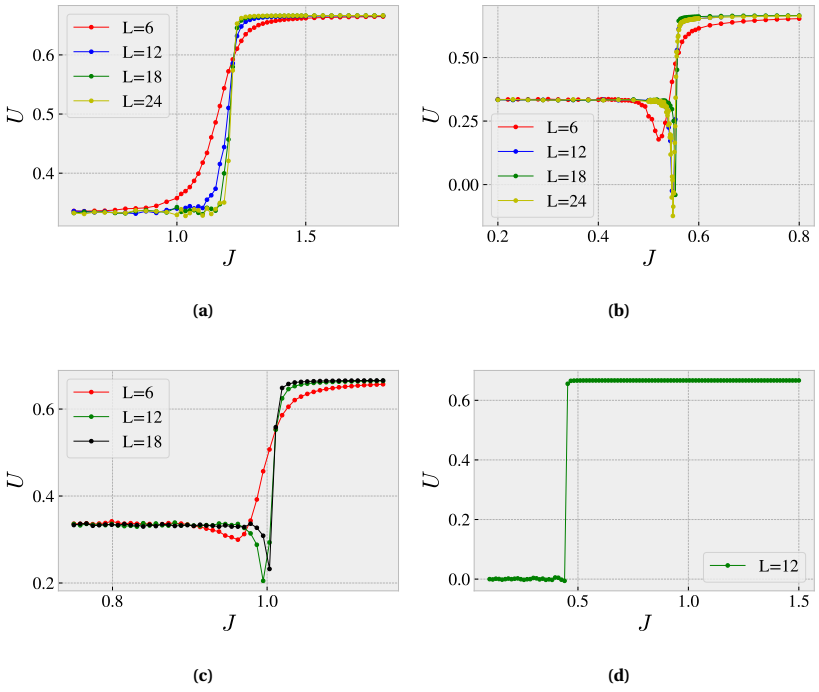


Figure 3.15: Examples of the Binder ratio as function of system size: (**a**): A typical example of a second order transition from confining to Higgs for $O(2)/Z_2$ at $K = 0$, (**b**): An example of a first order transition for $O(2)/Z_2$ at value of K in the range of first order - "Huse-Leibler" line emanating from the tricritical point (**c**): An example of the Binder ratio for a weak first order transition, showing its behavior for $O(2) \times O(2)/Z_2$ for $K = 0$ as function of J. (**d**) An example of the Binder ratio for $Z_2 \times Z_2 \times Z_2/Z_2$ along $K = 0$ with a clear indication of a second order phase transition.

All the phase diagrams presented in this paper are graphs of specific heat. But for determining the precise nature of the phase order that is not enough. As it can be seen from raw data graphs absolute values of specific heat can be an indicator of the order, but can't be completely trusted, because these values depend on the model and are not universal. We only used specific heat as an indicator of where the transition might be Fig.3.17, but analysis using Binder and histograms were done to determine the order of the transition.

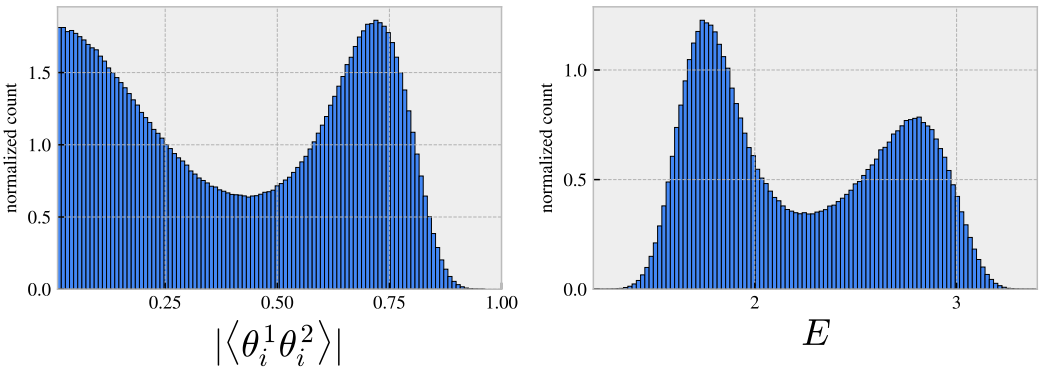


Figure 3.16: Example of average order parameter ($\langle\theta_i^1 \cdot \theta_i^2\rangle$) and energy ($E$) distributions values during a single run of a simulation, on the "Huse-Leibler" first order line $J = 0.29, K = 0.55$, for $Z_2 \times Z_2 / Z_2$ model with the usual $J = J_1 = J_2$ and $J_L^{12} = 0$. We see the appearance of two peaks in these distributions indicating the coexistence of two phases.

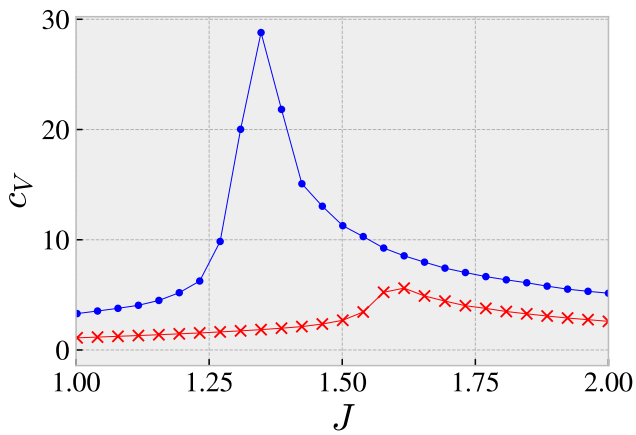


Figure 3.17: Example of the specific heat c_V for $O(2) \times O(2) / Z_2$ (blue-dots) and $O(2) / Z_2$ (red-cross) along the $K = 0$ slice (Higgs to confining transition). These two models have a different order of a phase transition which might be seen from the amplitude of a c_V divergence. Usually sharper and bigger divergences indicate the first order has occurred. However, one has to be careful due to finite size effects, and for this reason we also study the Binder cumulants. Better use for this graphs is in roughly locating where the transition happens. This information can be used to run the simulation on a more finely spaced grid around the transition in order to get better convergence and more precise point of transition using finite size methods.

4

Dense Entanglement in Critical States

Abstract

At critical points the entanglement between microscopic degrees of freedom is thought to be maximal, and proportional to the number of dynamical fields. In 1D systems this is analytically known and numerically verified through the knowledge that the bond dimension in tensor network states represents the upper bound on the amount of entanglement a system can represent. Here we test this in 2D systems using Carleo & Troyer neural-network quantum states in solving many-body quantum systems at their criticality. We postulate that for a neural-network quantum state (NQS) at criticality the entanglement is bounded by the ratio of the number of visible (N) and hidden nodes (M), $\alpha = M/N$. Computing the entanglement as a function of α at criticality in Z_2^n/Z_2 lattice gauge theories, allows us to study entanglement at criticality for differing number of dynamical fields. Surprisingly we do not find a linear relation.

4.1. Introduction

Though the idea precedes it by a number of years, the notion to use entanglement to classify the ground states of quantum many body systems really took off with the discovery of the topological insulator. Ordinary states of matter with quasi-particle expectations around a trivial IR fixed point are short range entangled states. Topo-

logical insulators with a gap that separates the bulk spectrum from the topological (edge) modes are long range sparsely entangled states. Sparse, because the number of topological modes is very small compared to the exponentially large number of excitable modes in the full Hilbert space. It also raises the immediate question whether there are long range *densely* entangled states of matter. Such states are arguably the most quantum many-body state possible and have been coined quantum supreme matter.

Frustrated systems and quantum spin liquids are possible candidates. But another possible candidate is a theory right at the non-trivial IR fixed point of a second order phase transition where the correlation length is infinite. As no length scale remains, there cannot be either short or long range entanglement. Entanglement must exist at all scales.[1] This is exemplified by the classic calculation of the entanglement entropy in 1+1D conformal field theories [69]

$$S_{\text{1D CFT}} = \frac{c}{3}\ln\left(\ell/a\right). \tag{4.1}$$

Here ℓ is the size of the region (the area) for which the entanglement with the complement is computed; a is a UV cut-off, and c is the central charge of the theory. Since the central charge is a measure of the number of degrees of freedom in the system, the scaling of the entanglement entropy with c shows that all degrees are involved and entanglement is both dense and long range.

An illustrative example of the dense entanglement in critical states is from a study of such 1D systems using the Multi-Scale Entanglement Renormalization Ansatz (MERA) [70]. Similar to Matrix Product States (MPS), these are variational descriptions of many-body-ground states designed to track (up to area-scaling) entanglement in terms of the bond-dimension of the variational state. MERA improves on MPS by engineering in critical behavior from the start. Tuning such a 1D MERA system to criticality one indeed sees that to ensure the same accuracy in the ground state energy, the minimal bond dimension must grow exponentially with the central charge. Since the bond dimension is designed to scale as $D \sim \exp\left(S_{\text{1D CFT}}\right)$, this includes the scaling with the central charge, consistent with Eq.(4.1).

The aim of this paper is to verify that this similar exponential increase in entanglement at critical states also happens in 2D systems. An extension of MERA to 2D systems is notoriously difficult. However, the advent of neural network machine learning techniques, has given us an inroad to this question. From Neural Quantum States [71, 72], where the many-body groundstate is contructed as a variational wavefunction based on a Restricted Boltzmann Machine neural network, one can also get an estimate of the entanglement or rather the entanglement entropy between two spatially separated parts of the groundstate wavefunction [73]. The power of this ap-

[1] The remarkable connections between quantum entanglement and emergent geometry in holographic theories suggest that the non-trivial IR fixed point dual to extremal black holes are of this type [67, 68].

proach is that it is not limited to 1D systems [72]. It works in principle in any dimension and is only limited by computational time.

The question that remains then is which 2D systems to use. Though one can simply study the approach to a critical point, one would need a notion of a central charge to make a fully equivalent statement compared to 1D systems. Of course there is no notion of a central charge in 2D systems. However, using that the central charge encodes for the number of degrees of freedom, we can rely on a recent finding that there is a nice sequence of critical points in classical 3D $Z_2 \times Z_2 \times \ldots \times Z_2/Z_2$ gauge theory [48].[2] With each additional Z_2 matter factor the number of degrees of freedom increases, yet in all other aspects the critical points are similar. The thermodynamics of these classical 3D systems corresponds to the quantum groundstate of 2D gauged transverse field Ising models [74–76]. It is then natural to suppose that one finds an increase in the entanglement in the groundstates of 2D Z_2 gauged transverse field $Z_2 \times Z_2 \times \ldots \times Z_2$ Ising at criticality proportional to the number of matter fields involved. That all matter fields are involved is suggested by the nature of the phase transition: it belongs to the $p = 2^{N_{\text{rep}}-1}$-Potts universality class [48].

In section 4.2 we set up the RBM based Neural Quantum State variational approximation to the groundstate wavefunction of $(Z_2)^n/Z_2$ gauged transverse field Ising. Computationally we will limit our attention to $n = 2,3,4$. We then compute the Entanglement Entropy between two parts of the system in section 4.3. The results are partially surprising. We see the rise in entanglement entropy as we approach the critical point corresponding with the absence of a scale and hence densification of entanglement. However, at criticality we do not see the expected increase in the entanglement entropy corresponding to the increase in the number of matter fields. In fact the entropy exhibits a puzzling behavior with increasing matter fields. We discuss possible explanations for this unexpected finding in the conclusion section 4.5.

4.2. Neural Quantum State approximation to groundstates of $Z_2 \times Z_2 \times \ldots \times Z_2/Z_2$ transverse field "Ising" lattice gauge theory

4.2.1. NQS from RBM Set-Up: application to Z_2 gauge theory

A Neural Quantum State (NQS) is a variational wave function ansatz based on a Restricted Boltzmann Machine (RBM). It was introduced by Carleo & Troyer [71] inspired by the use of RBMs in machine learning problem but now applied to minimize the ground state of a given Hamiltonian. Consider a system with N spin-1/2

[2] This is inspired by orbifold models of 1D systems.

spins labeled as $S = (s_1, s_2, \ldots, s_N)$ with $s_i \in \{-1, 1\}$. Then the RBM represents a wave function in the following way: The physical spins are complemented by hidden spins $H = (h_1, \ldots, h_M)$ with $h_i = \{-1, 1\}$. Then one posits the variational function for the quantum state:

$$\Psi_M(S, \mathscr{W}) = \sum_{h_i} e^{\sum_j a_j s_j + \sum_i b_i h_i + \sum_{ij} W_{ij} h_i \sigma_j^z}, \tag{4.2}$$

which depends on the values of the weights as $\mathscr{W} = (a_i, b_j, W_{ij})$. These weights form a network (Fig.4.1) and the expression resembles a Boltzmann sum over the h_i configurations.

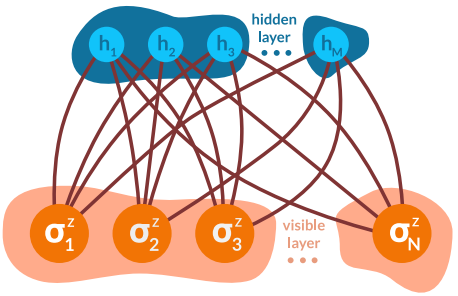


Figure 4.1: Illustration of a restricted Boltzman machine (RBM) neural network.

The qualification Restricted refers to the fact that the weights (interactions) only connect visible and hidden spins, and there are no weights amongst hidden spins. Because of this lack of inter-hidden layer connectivity of an RBM all hidden variables can be traced out which leaves us with:

$$\Psi(S, \mathscr{W}) = e^{\sum_i a_i s_i} \prod_{i=1}^{M} 2\cosh\left(b_i + \sum_j W_{ij} s_j\right), \tag{4.3}$$

Following the algorithm from [71] we can train this network to efficiently represent the ground state of our Hamiltonian.

The reason we choose a NQS variational method is that one can readily compute the entanglement entropy in the (approximate) groundstate. To exhibit dense entanglement at criticality in 2D systems, we wish to compute the entanglement entropy on a set of related theories that has tuneable set of degrees of freedom. A family of such theories are $Z_2 \times Z_2 \ldots Z_2 / Z_2$ multiple Ising matter fields gauged with $\mathbb{Z}_2$ symmetry on a 2+1D lattice [48]. These are extensions of the well known Z_2/Z_2 lattice gauge theory with Higgs fields from Fradkin and Shenker [75]. On the sites of a lattice, labeled by $\vec{r}$, we have n-Ising matter fields ($\sigma^i(\vec{r})$, $i = 1 \ldots n$) and on the links in the direction of the lattice vector $\hat{e}_\mu$ we have an Ising gauge field ($U(\vec{r}, \hat{e}_\mu)$. The action of

a $d+1$ dimensional model on a hypercubic lattice is defined as:

$$\begin{aligned} S(\sigma(\vec{r}), U(\vec{r}, \hat{e}_\mu)) &= J \sum_{i=1}^{n} \sum_{\vec{r},\mu} \sigma^i(\vec{r}) U(\vec{r}, \hat{e}_\mu) \sigma^i(\vec{r} + \hat{e}_\mu) \\ &+ K \sum_{\vec{r},\mu\nu} U(\vec{r}, \hat{e}_\mu) U(\vec{r} + \hat{e}_\mu, \hat{e}_\nu) U(\vec{r} + \hat{e}_\nu + \hat{e}_\mu, -\hat{e}_\mu) U(\vec{r} + \hat{e}_\nu, -\hat{e}_\nu) \, . \end{aligned} \tag{4.4}$$

This action is invariant under the local $\mathbb{Z}_2$ gauge transformations:

$$\begin{aligned} \sigma^i(\vec{r}) &\rightarrow = \sigma^i(\vec{r}) s(\vec{r}) \, , \\ U(\vec{r}, \hat{e}_\mu) &\rightarrow s(\vec{r}) U_\mu(\vec{r}, \hat{e}_\mu) s(\vec{r} + \hat{e}_\mu) \, , \end{aligned} \tag{4.5}$$

where $s(\vec{r}) = \pm 1$. Though for $n = 1$ there is famously no phase transition at $K = 0$ as a function of J illustrating the formal equivalence between the confining ($J < 0$) and Higgs ($J > 0$) phase of the Z_2/Z_2 theory, for any $n \geq 2$ there is a second order phase transition at finite J between an disordered $J < 0$ and ordered phase $J > 0$ [48]. The phase transition belongs to the $p = 2^{n-1}$-Potts universality class and is characterized by an expectation value of the (gauge invariant) registry order parameters $\langle \sigma^1 \sigma^2 \rangle$, $\langle \sigma^1 \sigma^3 \rangle$, …, $\langle \sigma^{n-1} \sigma^n \rangle$. We will use the product of all order parameters to measure all of them simulataneously $\langle \sigma_1^{n-1} \sigma_2^{n-1} \ldots \sigma_n^{n-1} \rangle = \prod_{i<j} \langle \sigma_i \sigma_j \rangle + \ldots$. This implies that the 2D quantum system corresponding to this system has a quantum phase transition with critical behavior at the corresponding point. The corresponding quantum Hamiltonian can be derived using transfer matrix formalism. Following [74, 76] we find a Z_2 gauged transverse field Ising Hamiltonian

$$\begin{aligned} H = &- \sum_{i,\vec{r}} \sigma_1^i(\vec{r}) - \sum_{\vec{r},\mu} \tau_1(\vec{r}, \hat{e}_\mu) - \lambda \sum_i \sum_{\vec{r},\mu} \sigma_3^i(\vec{r}) \tau_3^\mu(\vec{r}) \sigma_3^i(\vec{r} + \hat{e}_\mu) \\ &- \omega \sum_{\vec{r},\mu\nu} \tau_3(\vec{r}, \hat{e}_\mu) \tau_3(\vec{r} + \hat{e}_\mu, \hat{e}_\nu) \tau_3(\vec{r} + \hat{e}_\nu, \hat{e}_\mu) \tau_3(\vec{r}, \hat{e}_\nu) \, , \end{aligned} \tag{4.6}$$

where now the matter fields σ^i and the gauge field τ are Pauli matrices acting on sites and links of a 2D lattice, and we used that for a Z_2 symmetry the link is its own Hermitian conjugate $\tau(\vec{r} + \hat{e}_\mu, -\hat{e}_\mu) = \tau(\hat{r}, \hat{e}_\mu)$. The coupling coefficients λ and ω can be related to K and J in the 2+1D classical action formulation, but the precise relation is unimportant. We shall set $\omega = 0$ and keep λ undetermined. What is important for our NQS variational approach is that physical states of this theory must satisfy the gauge invariance constraint. So therefore must the NQS itself. The local gauge transformations are generated by $G(\vec{r}) = \prod_i \sigma_1^i(\vec{r}) \prod_\mu \tau_1(\vec{r}, \hat{e}_\mu)$ at each site $\vec{r}$. Every physical state $|\psi\rangle$ must therefore obey

$$G(\vec{r}) |\psi\rangle = |\psi\rangle \tag{4.7}$$

for every site $\vec{r}$ of the lattice. This constraint can easily be visualized in the τ_1 and σ_1 basis Fig.4.2, basically every site needs to have $n + m = 2k, k \in \mathbb{Z}$ with n the number of down spins on and m the number of down-links emanating from the site.

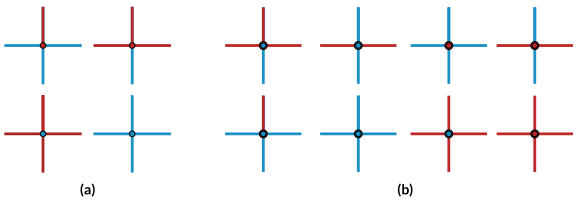


Figure 4.2: Visualization of the gauge constraint in τ_1 and σ_1 basis. Red color represents down spins (-1) and blue up spins (+1). **(a)** Different gauge invariant combinations in the case of Z_2/Z_2 theory. **(b)** Different gauge invariant configurations for the $Z_2 \times Z_2/Z_2$ theory. For the visual purposes spins at the nodes of a lattice for two fields are different sizes, so they can be distinguished in the diagram.

The NQS wavefunction itself is approximated by MCMC sampling. Due to the gauge constraint, this has to be done with care, because the naive Hilbert space is now larger than the space of physical states. We employ a simple updating procedure that always flips two matter spins, a matter spin and a link, or two links on a site to remain within the physical Hilbert space. For alternate approaches to NQS states for gauged lattice models, see [77].

4.3. Entanglement entropy

The reason we resort to NQS to approximate the ground state is that entanglement or rather the entanglement entropy can be readily computed for such wavefunctions. Entanglement entropy is of course only a partial measure of entanglement, but to first approximation it should be able to quantify its denseness. Given a system described by a density matrix ρ and divided into two parts, A and B, then the entanglement entropy between these parts is defined as:

$$S_A = -\operatorname{tr}\left[\rho_A \log \rho_A\right] , \quad \rho_A = \operatorname{Tr}_B \rho \, . \tag{4.8}$$

If the original state is pure, as is the case here, then $S_B = S_A$.[3]

Quite generally, for a system decomposable into two parts A and B, the (ground)state can be written as

$$|\Psi\rangle = \sum_{i,j} c_{i,j} \left|i\right\rangle_A \otimes \left|j\right\rangle_B \, . \tag{4.9}$$

The coefficients of the expansion $c_{i,j}$ can be combined in one probability-amplitude matrix where each element at the coordinates i, j would be a probability-amplitude that we find the system with part A being in the state i and part B being in the state j.

[3] An important comment is that one has to be careful in computing the entanglement entropy in gauged theories, see e.g. [78]. As our algorithm specifically only limits to physical states, this is not an issue, and we can use the standard expression Eq.((4.8)).

The entanglement entropy can be computed using Schmidt/Singular Value Decomposition (SVD); see [73] in the context of NQS or e.g. [79] in the contect of MCMC. We write the probability-amplitude matrix as:

$$c_{i,j} = U_{i,k}\Sigma_{k,l}V^{\dagger}_{l,j} \,. \tag{4.10}$$

If we label with N_A and N_B the sizes of parts A and B respectively, then U, Σ and $V^{\dagger}$ are matrices of dimensions $N_A \times N_A$, $N_A \times N_B$ and $N_B \times N_B$, where $\Sigma_{kl} = \sigma_k \delta_{kl}$ is a "diagonal" matrix with singular (i.e. non-negative real) values on the main diagonal and zeros otherwise. Using this decomposition the entanglement entropy is easily seen to equal

$$S_A = -\sum_{i=0}^{\min(N_A,N_B)} \sigma_i^2 \log \sigma_i^2 \,. \tag{4.11}$$

In the variational NQS approach the wavefunction is represented probabilistically by an ensemble of $N_{\text{ens.}}$ states with $N_{\text{ens.}} = 10^4$ as default choice. To compute the entanglement entropy from this subrepresentation, we follow [73]. These $N_{\text{ens.}}$ correspond to $N_{\text{conf}} \ll N_{\text{ens.}}$ different spin configurations, where the multiplicity $n_{\vec{s}}$ of each spin configuration is directly related to its weight $p_{\vec{s}} = n_{\vec{s}}/N_{\text{ens.}}$ in the ensemble. Algorithmically we can easily read off the non-vanishing spin configurations in subsystem A and subsystem B, and construct the non-vanishing components of the matrix $c_{i,j} = \langle i, j|\psi_{\text{NQS}}\rangle$. We hierarchically order the absolute value of the $|c_{i,j}|$. We make a reduced ansatz by only keeping the $N_{\text{red.}}$ largest values. This gives an $N^{(A)}_{\text{red.}} \times N^{(B)}_{\text{red.}} \geq N_{\text{red.}}$ matrix of $c^{\text{red.}}_{i,j}$. In case one is interested in the entanglement entropy between exactly one half of the system and the other, then by symmetry $N^{(A)}_{\text{red.}} \times N^{(B)}_{\text{red.}} = N_{\text{red.}}$, and $N_{\text{red.}}$ should be chosen to be an exact square. We then compute the entanglement entropy of $c^{\text{red.}}_{i,j}$ by Schmidt decomposition. The accuracy of the entanglement entropy is controlled by the truncation $q = N_{\text{red.}}/N_{\text{ens}}$, and we can compare this to the accuracy in the groundstate energy when only sampled over the $N_{\text{red.}}$ most important contributors to the ensemble.

There is one point one needs to pay special attention to. When computing the entanglement entropy for gauge theory in the above described way, it can happen that in constructing the matrix $c_{i,j}$ that the final state resulted from combining states i and j is not gauge invariant. In that case the probability of system reaching that state is zero. When creating matrix $c_{i,j}$ we need to check all elements and if the state they came from is not gauge invariant set those to zero. Doing this we ensure that only gauge invariant states contribute to the entanglement entropy.

4.4. Dense entanglement at criticality or not?

The 2D lattice we choose will be square of size $N_{\text{lattice}} = 3 \times 3$ with periodic boundary conditions (PBC). We will consider the sequence of $(Z_2)^n/Z_2$ gauge theories that have

$n = \{2,3,4\}$. More than 4 matter fields or more lattice sites becomes computationally expensive. The NQS variational ansatz will have with N_{lattice} visible and M hidden nodes, with the ratio of two labeled as $\alpha = M/N$. We try four different configurations $\alpha = \{1,2,3,4\}$. The accuracy of the untruncated groundstate energy is expected to scale polynomially in α [70]; $\Delta E_{\text{g.s}} = a\alpha^{-b}$. Unlike [70], the exact groundstate energy is not known for our models. However, studying the convergence of the NQS for $\alpha = 1,2,3,4$ we can roughly see that increasing the number of hidden knows gives an improvement that decreases relatively to the number of hidden nodes, if we ignore the lowest result $\alpha = 1$. This is represented in Fig.4.3, where we have sampled over 50 different initial conditions, and estimated the accuracy of the groundstate by using bootstrap over those 50 initial configurations.

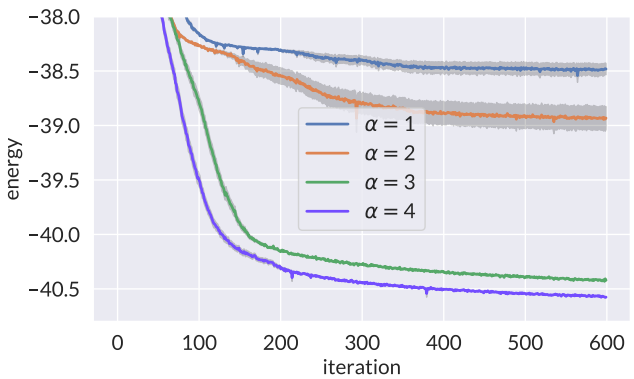


Figure 4.3: The approach towards the optimal NQS ground-state wavefunction for the Z^2/Z model measured through its energy as a function of learning epoch for $\alpha = 1,2,3,4$ for the value $\lambda = 0.948, \omega = 0$. Considering the $\alpha = 1$ result an outlier, one sees an improved convergence as the number of hidden nodes α is increased. The average over 50 initial conditions is given as well as the standard error computed using bootstrap over 500 resamples.

In this finite size system, there is no true instaneous phase transition, but its incipience is clearly visible in both the specific heat and the development of a finite order parameter for the registry symmetry. Fig.4.4 illustrates this. We clearly see the incipient second order phase transition as predicted for this model in [48].

We can now test how entanglement also in 2D systems becomes denser *both* as we approach the critical point, *and* as we increase the number of degrees of freedom analogous to the central charge in 1D systems. Fig.4.6 gives our results of the entanglement between 2/3 and 1/3 of the system. In the case of 3×3 system, dividing system exactly in half is not possible, the method we opted out for is the divisions along the secondary diagonal as in the Fig.4.5.

Fig.4.6 and specifically the middle row of Fig.4.6 shows the results of the entangle-

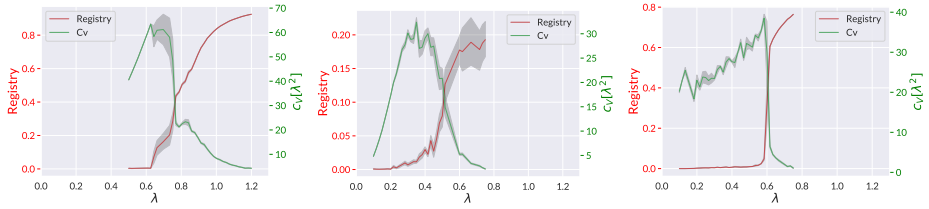


Figure 4.4: The specific heat $c_V = \frac{\partial \langle E \rangle_{NQS}}{\partial \lambda}$ (green) and registry order parameter $\langle \mathscr{O}_n \rangle = \langle \sigma_1^{n-1} \ldots \sigma_n^{n-1} \rangle_{NQS}$ (red) for the Z_2^n/Z_2 gauge theory for $n = 2,3,4$ for $\omega = 0$ as a function of λ determined from the NQS wavefunction for $\alpha = 3$. The mean and standard deviation after are computed using bootstrap.

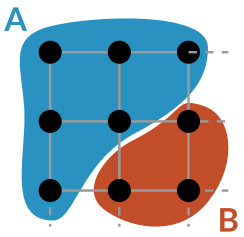


Figure 4.5: Illustration of the system division for a 3 × 3 lattice used in our computations.

ment entropy for the sequence of Z_2^n/Z_2 theories as a function of the relative coupling $\lambda - \lambda_c/\lambda_c$ w.r.t. the criticial point. Initially the entanglement entropy does increase when going from 2 to 3 fields, but then stays the same when we added another 4th field. Though the numerical results are not super smooth, and there is a slight increase for $\alpha = 4$, we expect a linear increase and this is clearly not there. There are several possible explanations for this. One is that finite size effects due to computational limitations do have a direct impact. This cannot be ruled out, but the fact that the entanglement entropy does not change much with the increase in α suggests otherwise. Already $\alpha = 1$ appears sufficient to represent all the entanglement in the system. Another possible explanation for the relation between entanglement entropy and the number of matter fields is that increasing number of fields would require an increase in the $N_{\text{red.}}$. From the right figure in the bottom row of Fig.4.6 we see that increase of $N_{\text{red.}}$ lead to the increase in the entanglement entropy. We are again computationally limited here as a further increase in the number of required states would cause memory issues.

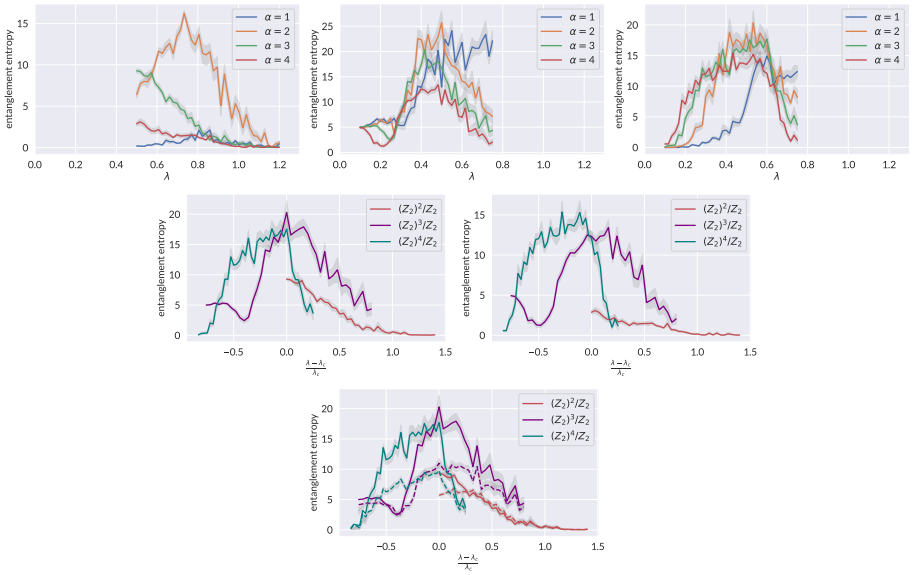


Figure 4.6: The entanglement entropy $S_{2/3}$ of 2/3 of the 3×3 system entangled with the other 1/3 for the $N_{\text{red.}} = 10^4$ truncated (see text) NQS groundstate of Z_2^n/Z_2 gauge theories averaged over 50 initial conditions. The top row shows the dependence on the number of hidden variables as a function of λ for $N_{\text{red.}} = 10^4$: Left is the result for $n = 2$ for $\alpha = 1,2,3,4$; middle for $n = 3$; right for $n = 4$. This estimates the error in the entanglement entropy. The middle row shows the dependence on n rescaled to $\lambda_{\text{red.}} = \frac{\lambda-\lambda_c}{\lambda_c}$; left for $\alpha = 3$, right for $\alpha = 4$ for $N_{\text{red.}} = 10^4$. No discernible increase in the entanglement entropy as a function of n is seen. The bottom row shows the dependence on $N_{\text{red.}} = 10^4$ (solid line) and $N_{\text{red.}} = 10^3$ (dashed line).

4.5. Discussion

Having presented our results, we must leave a verification whether entanglement entropy scales with the number of degrees of freedom also at a 2D critical point an open question. Surprisingly, within the limitations of our numerics in the Z_2^n/Z_2 sequence of models we studied this appears not to be the case. We cannot fully rule out that a technical/computational limitation is the cause, but we have performed extensive tests and the code faithfully reproduces the known 1D results (see Appendix). One possibility to overcome computational limitations is to move away from the traditional Monte Carlo sampling and using more direct approach like equivariant flow-based sampling [80] or generative models [81]. We leave this for future research.

Acknowledgments

We are especially grateful to Giuseppe Carleo and Matija Medvidović for their help in setting up NQS for gauged models. We thank Jan Zaanen for discussions and his

insistence that this should be tested in models with $d > 1$. This research was supported in part by the Dutch Research Council (NWO) project 680-91-116 (*Planckian Dissipation and Quantum Thermalisation: From Black Hole Answers to Strange Metal Questions.*) and by the Dutch Research Council (NWO)/Ministry of Education.

4.6. Appendix

4.6.1. NQS States and Entanglement for 1D transverse field Ising Model

Given that we have such an unexpected and curious result, it behooves an in depth exhibition of the validity of our approach. Here we compute the known entanglement entropy in a 1D transverse Ising model using exactly the same algorithm. These results agree with the theoretical expectation as well as the numerical NQS results of [73]. We have parametrized the 1D transverse field Ising model Hamiltonian analogous to Eq.((4.6))

$$H_{\text{TI-1D}} = -\sum_{\vec{r}} \sigma_1(\vec{r}) - \lambda \sum_{\vec{r},\mu} \sigma_3(\vec{r})\sigma_3(\vec{r}+\hat{e}_\mu) \tag{4.12}$$

and use a $N = 16$ site system with periodic boundary conditions. The results are on Fig.4.7. One sees the increase in entanglement as one approaches the critical value $\lambda_c = XX$ consistent with the notion of dense entanglement.

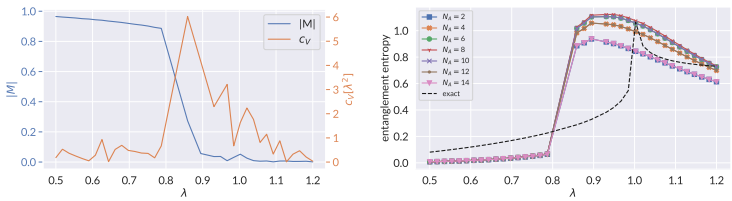


Figure 4.7: Left: The specific heat and Z_2 order parameter of the 1D Ising model as a function of λ. Right: The entanglement entropy both as a function of λ and as a function of boundary site $s = 2, 4, 6, 8, 10, 12, 14$ between subsystem A and subsystem B (the boundary site is the last site that is included in A.). The dashed line is the CFT result in the continuum limit.

5

Criticality versus Uniformity in Deep Neural Networks

Attribution

This paper has been previously published as a preprint on arXiv and has been submitted to Journal of Machine Learning Research for publications, and it is currently under the editorial review, under the title ***Criticality versus uniformity in deep neural networks***, together with Jurriaan de Gier, Kevin T. Grosvenor, Ro Jefferson, Koenraad Schalm and Eliot Schwander.[82]

Abstract

Deep feedforward networks initialized along the edge of chaos exhibit exponentially superior training ability as quantified by maximum trainable depth. In this work, we explore the effect of saturation of the tanh activation function along the edge of chaos. In particular, we determine the line of uniformity in phase space along which the post-activation distribution has maximum entropy. This line intersects the edge of chaos, and indicates the regime beyond which saturation of the activation function begins to impede training efficiency. Our results suggest that initialization along the edge of chaos is a necessary but not sufficient condition for optimal trainability.

5.1. Introduction

Over the past decade or so, deep learning has emerged as one of the most powerful tools for processing and analyzing data, and has proven successful on an increasingly wide range of computational challenges. These remarkable feats include highly accurate image classification [83], advanced generative modelling of images [84], natural language processing [85], accurate protein structure predictions [86], and besting humans in a wide range of games [87]. Key to these neural networks' success is the extremely large number of parameters—generally speaking, the *expressivity* of a neural network increases with depth [88]. Expressivity refers to the range of functions that a network can approximate, with the network being understood as simply a function from the space of inputs to the space of outputs. However, the price we must pay for larger and more powerful networks is that they are more difficult to train; for example, the risk of vanishing or exploding gradients is exacerbated with depth [89]. Hence, an improved understanding of how the network parameters impact trainability is highly valuable, as even small improvements in the initialization of deep neural networks can make intractable problems tractable.

In this work, we study trainability in deep random feedforward neural networks. Such networks are frequently used in the literature due to their analytical tractability: the phase space is two-dimensional and parameterized by the variances of the initial weight and bias distributions: σ_w^2 and σ_b^2.[1] This makes them useful models for investigating general features of deep networks. In particular, we will be concerned with the behavior of the pre- and post-activations, in terms of both their distributions as well as the accuracy of the network on a classic image classification task, namely MNIST (numerical digit recognition) and CIFAR-10 (colored images, which we convert to grayscale).

More specifically, we build on previous work [14, 15] which demonstrated the presence of an order-to-chaos phase transition in this class of deep networks. Intuitively, correlations in the input that we wish to learn are exponentially suppressed with depth in the ordered (analogously, low-temperature) phase, and washed-out by noise in the chaotic (high-temperature) phase; these two phases are characterized by vanishing or exploding gradients, respectively. The boundary between these two phases is a critical line called the *edge of chaos*,[2] which is a continuous phase transition characterized by a diverging correlation length ξ for the layer-to-layer two-point function of the neurons. Since the correlation length sets the depth scale at which information can propagate, this theoretically enables networks of arbitrary depth to be trained at criticality (more generally, networks are trainable provided their depth

[1] As is standard in the literature, we restrict to zero-mean networks, as initializing with a small non-zero mean does not qualitatively change our results.

[2] Technically, this should be called the edge of stability, but we will use edge of chaos synonymously with criticality for consistency with the literature.

does not exceed the scale set by ξ). In other words, the deeper the network, the closer one must lie to the edge of chaos; this was demonstrated in [15] along a slice of parameter space at bias variance 0.05 and weight variance ranging from 1 to 4, and subsequently generalized/corroborated in, e.g., [90–92]

Several questions naturally arise from the above work. First, given that the network parameters will evolve under training in order to minimize the specified cost function and, in particular, develop interdependencies, why does the choice of initialization have such a decisive effect on network performance?[3] Indeed, it was observed in [92] that the hidden-layer pre-activation distributions (as quantified by their variance) rapidly approach some asymptotic value within 10 or fewer layers, and then remain relatively unchanged for arbitrarily many additional layers. We corroborate this fact at the level of the post-activation in fig. 5.6 of appendix 5.6.1.

Second, what role does the particular distribution of post-activations in a given layer play in determining network performance? For example, the activation function considered in [15] is hyperbolic tangent, which we adopt henceforth. When $\sigma_b^2 \ll 1$ and $\sigma_w^2 \lesssim 1$, the pre-activations z of the hidden layers are approximately Gaussian-distributed with small variance (cf. (5.8)). In this case, $\tanh(z) \approx z$, so the network behaves like a linear network. These are quite restrictive, being incapable of representing functions whose output data are non-linearly separable and cannot be generated by a combination of linearly separable data. In the opposite extreme, for large values of σ_w^2 and σ_b^2, the pre-activation variance becomes so large that the post-activation distribution becomes peaked at ± 1. In other words, large pre-activation variance saturates the tanh, causing it to behave like a discrete step-function. One expects this regime to also impair trainability, since the gradients on which the back-propagation algorithm depends become vanishingly small everywhere except near the origin.[4] Thus, it seems that one should seek to remain somewhere between these two extremes. Quantifying this is one of the main motivations for the present work.

In particular, note that in both the linear and the saturation regimes, one expects the expressibility of the network to be poor. In contrast, between these extremes lies a region in which the post-activation distribution is approximately uniform, and hence we might expect the expressibility of the network to be maximized at this point. To see this, recall that the uniform distribution has maximum entropy, which measures the number of possible states any particular system can have; a step function, in contrast, can only store a single bit of information, and hence has a low entropy of $\ln 2$. This leads to the conjecture that networks whose internal distributions are approximately uniform, i.e., maximally entropic, have higher expressibility, and hence

[3] In other words, why does the network remain near the initialization regime (e.g., the edge of chaos) as it evolves?

[4] Recall that the updates to the weights and biases under gradient descent contain products of the derivatives of the activation functions in all higher layers.

might enjoy a performance advantage. Of course, given approximately Gaussian preactivations, the post-activation distribution of tanh cannot be exactly uniform, but we can quantify the degree of uniformity via the relative entropy (defined below). In fact, we will show that there is a *line of uniformity* on the (σ_w^2, σ_b^2) phase space along which the post-activation distribution is as uniform as possible. This line intersects the aforementioned edge of chaos (see fig. 5.1), and the relative importance of lying near this line is the primary question we shall explore below.

We shall begin by deriving an expression for the line of uniformity, defined by the condition that the distribution of the final hidden layer minimizes the relative entropy with respect to the uniform distribution. The computation uses many of the same ingredients as [15], and the interested reader is encouraged to turn there for more background. We then examine proximity to this line in relation to the edge of chaos considered in previous works.

5

We find that for deep networks away from the edge of chaos, the exponential suppression dominates, and no benefit from uniformity is observed. However, along the edge of chaos – where the suppression is only polynomial – we find a relatively sharp fall-off in the post-training accuracy to the right of the line of uniformity. The location of this fall-off depends on the learning rate, since decreasing the learning rate can increase the final accuracy, but at the cost of additional computing time (see fig. 5.2). This suggests that criticality is a necessary but not sufficient condition for optimal trainability.

This dependence on other hyperparameters illustrates that optimal trainability is not just a matter of final accuracy but also of efficiency, i.e., how quickly the final accuracy is reached. Since computational limits exist, we shall rely on an intuitive notion of efficiency per epochs in addition to accuracy; that is, we consider the accuracy achieved after a fixed number of training epochs. It is conceivable that in the limit of infinite training epochs accuracy differences disappear, so that formally, the configurations are equally good. In a practical sense however, they clearly are not.

Note that there can obviously be very many notions of efficiency depending on which resource(s) one considers most valuable. Here, we are implicitly prioritizing training time, i.e., number of epochs. If one were to put the premium on floating point operations used in training, then one would instead measure efficiency as in [93]. Yet another concept called learning efficiency has to do with how much time it takes to run a learning algorithm and, in particular, how this scales with the size of the input space [94].

Returning to our main question, to isolate the effects of uniformity *away* from the edge of chaos, we also examine networks which are both shallow (i.e., not yet exponentially suppressed) and narrow (i.e., low expressibility per layer), and confirm

that training efficiency, in the sense described above, degrades to the right of the line of uniformity (i.e., away from the origin), though final accuracy need not. In contrast to the edge of chaos, the line of uniformity is not a sharp phase boundary, but it does indicate coarsely the parameter boundary where activation saturation starts to affect training efficiency. This not only establishes the more obvious point that, even in deep random feedforward toy models on the edge of chaos, backpropagation training depends sensitively on activation function choice, as earlier emphasized in [95, 96], but also that for a given activation function choice there are optimal points or regions on the edge of chaos itself.

5.2. The line of uniformity

We can estimate the location of the line of uniformity by capitalizing on the fact that wide networks, with a large number N of neurons in each hidden layer, are approximate Gaussian processes. At finite N, the neurons in a given layer are not independent due to their shared dependence on the neurons in the previous layer. Physically however, the non-Gaussianities that can be seen by marginalizing over the previous layer(s) can be thought of as interactions that are $1/N$ suppressed [97, 98]. Hence, in the limit $N \to \infty$, the distribution of pre-activations becomes Gaussian, essentially by the central limit theorem. This greatly simplifies the analysis, and is the reason for the widespread use of such models in previous studies, including [15].[5]

Thus, at large-N, the distribution of pre-activations z for any hidden layer takes the form

$$p(z;\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{z^2}{2\sigma^2}}\,, \tag{5.1}$$

where σ^2 is the variance, and we assume the mean $\mu = 0$ since adding a small finite mean does not qualitatively change our results. If the activation function $\phi(z)$ is one-to-one and once-differentiable, then the distribution of post-activations x will be given by

$$p_\phi(x;\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma\,\phi'\big(\phi^{-1}(x)\big)}\, e^{-\frac{\phi^{-1}(x)^2}{2\sigma^2}}\,. \tag{5.2}$$

Concretely, for $\phi(z) = \tanh(z)$, this yields

$$p_\phi(x;\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma(1-x^2)}\, e^{-\frac{\operatorname{arctanh}(x)^2}{2\sigma^2}}\,, \tag{5.3}$$

with $x \in [-1,1]$. The corresponding variance is given by

$$\sigma_\phi^2 = \int_{-1}^{1} \mathrm{d}x\, x^2\, p_\phi\big(x;\sigma^2\big)\,. \tag{5.4}$$

[5] One will often see the phrase "mean-field theory" used in place of the central limit theorem in this context; however, as pointed out in [98], this is not technically correct, and mean-field theory does not necessarily correspond to the $N \to \infty$ limit.

As mentioned above, we quantify the uniformity of the post-activation distribution p_ϕ by the relative entropy or Kullback-Leibler divergence with respect to the uniform distribution p_{uni},

$$S(p_{\text{uni}}||p_\phi) = \int_{-1}^{1} \mathrm{d}x\, p_{\text{uni}}(x) \ln \frac{p_{\text{uni}}(x)}{p_\phi(x)} \,. \tag{5.5}$$

Substituting in (5.3) and $p_{\text{uni}} = \frac{1}{2}$, this yields

$$S(p_{\text{uni}}||p_\phi) = \frac{1}{2} \ln\left(8\pi\sigma^2\right) + \frac{\pi^2}{24\sigma^2} - 2 \,. \tag{5.6}$$

This has a minimum at

$$\sigma^2_{\text{min}} = \frac{\pi^2}{12} \approx 0.822 \,. \tag{5.7}$$

Therefore, we wish to find the set of points (σ_w^2, σ_b^2) at which the variance of the final hidden layer is σ^2_{min}; this will define the line of uniformity. To proceed, we use the recursion relation

$$\sigma_\ell^2 = \sigma_w^2\, \sigma^2_{\phi,\ell-1} + \sigma_b^2 \,, \tag{5.8}$$

which follows from the large-N condition discussed above (i.e., the neurons on any given layer can be treated as i.i.d. random variables). Note that this is exactly the same as eq. (3) of [15], where our σ_ℓ^2 is their q^ℓ_{aa} and our $\sigma^2_{\phi,\ell-1}$ is the corresponding integral expression.[6] This recursion relation ostensibly requires the variance of the first hidden layer, σ_1^2, as an input. However, it turns out that (5.8) quickly converges to a fixed value σ_*^2, which (by definition) is a function of σ_w^2 and σ_b^2, but not of σ_1^2:

$$\sigma_*^2 = \sigma_w^2\, \sigma^2_{\phi,*} + \sigma_b^2, \tag{5.9}$$

where $\sigma^2_{\phi,*}$ is σ^2_ϕ evaluated at σ_*^2; see [14] for further discussion of this convergence. In appendix 5.6.1, we have demonstrated numerically that the corresponding post-activation distribution indeed converges rapidly to one which depends only on the initialization point (σ_w^2, σ_b^2).

Now, consider a fixed value of σ_*^2 (and hence also of $\sigma^2_{\phi,*}$). Then we can consider (5.9) as an expression for σ_b^2 as a function of σ_w^2, which defines a line in phase space of the form

$$\sigma_b^2 = \sigma_*^2 - \sigma^2_{\phi,*}\, \sigma_w^2. \tag{5.10}$$

where σ_*^2 is the y-intercept, and $-\sigma^2_{\phi,*}$ is the slope. Since the relative entropy (5.6) of the final hidden layer is only a function of its variance, the lines of constant σ_* given by (5.10) are also lines of constant relative entropy. In particular, the line of

[6] Explicitly, the variance can be written as $\sigma^2_\phi = \int \mathscr{D}z \left[\phi(\sigma z)\right]^2$, where $\mathscr{D}z = \frac{\mathrm{d}z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ is the standard Gaussian measure.

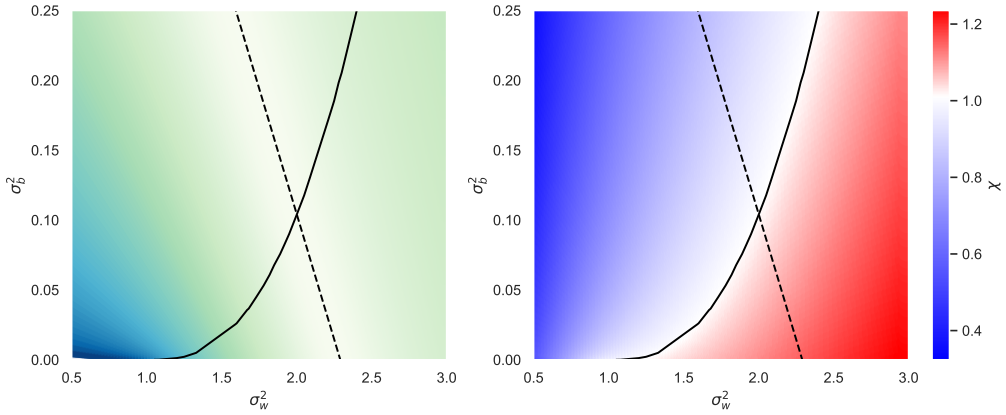


Figure 5.1: (Left) Contour plot of the logarithm of the relative entropy in the (σ_w^2, σ_b^2) plane. The dashed line is the line of uniformity—saturation increases to the right of it and linearity increases to the left of it. (Right) Contour plot of $\chi = e^{-1/\xi}$. The ordered/low-temperature phase is shaded blue, while the chaotic/high-temperature phase is shaded red. In both, the solid black line is the edge of chaos, while the dashed black line is the line of uniformity.

uniformity (minimum relative entropy) is given by (5.10) with $\sigma_*^2 = \sigma_{\text{min}}^2 = \frac{\pi^2}{12}$, cf. eq. (5.7). There is no closed-form expression for $\sigma_{\phi,\text{min}}^2$, but we can evaluate (5.4) numerically to obtain $\sigma_{\phi,\text{min}}^2 \approx 0.359$. In summary, the line of uniformity (LOU) is given by

$$\text{LOU}: \quad \sigma_b^2 = \sigma_{\text{min}}^2 - \sigma_{\phi,\text{min}}^2 \sigma_w^2, \tag{5.11}$$

with $\sigma_{\text{min}}^2 = \frac{\pi^2}{12} \approx 0.822$ and $\sigma_{\phi,\text{min}}^2 \approx 0.359$. In the left panel of fig. 5.1, we present a contour plot of the logarithm of the relative entropy. The line of uniformity is the dashed black line—to the left of it, as one approaches the origin, is the linear regime; and to the right, the activation becomes more and more saturated. For comparison, the edge of chaos is the solid black line.

5.3. The edge of chaos

The method for computing the edge of chaos as a function of σ_w^2 and σ_b^2 is described in [14, 15]. Once we have σ_*^2, as described previously, then we can define the quantities

$$\chi = \sigma_w^2 \int \mathscr{D}z \left[\phi'(\sigma_* z)\right]^2, \qquad \xi = -\frac{1}{\ln \chi}, \tag{5.12}$$

where $\mathscr{D}z$ is the standard Gaussian measure, cf. footnote 6, and ξ is the correlation length mentioned in the introduction (note that this is denoted ξ_c in [15]).

The meaning of χ will be discussed in the next paragraph, while the meaning of ξ is as follows: we consider two identical copies of the network and feed them slightly different inputs. Then, we can study the correlation (i.e., covariance) between a neuron in one copy and the same neuron in the second copy as a function of the layer. This correlation will decay exponentially for deeper layers with a characteristic length scale, ξ. (Strictly speaking, this is only true in the ordered phase: in the chaotic phase, the quantity ξ is complex-valued and cannot be interpreted as a correlation length). The edge of chaos is defined as the critical point, where the correlation length ξ diverges.

As discussed in more detail in [14, 15], χ is obtained as the derivative of the aforementioned covariance with respect to that in the previous layer, and probes the stability of the fixed point when the covariance is unity: $\chi > 1$ implies that we approach this point from below (unstable), while $\chi < 1$ implies that we approach this point from above (stable).[7] The edge of chaos corresponds to $\chi = 1$, where ξ diverges.

To find the edge of chaos, we can scan over the space of tuples (σ_w, σ_*) to find those which satisfy the condition $\chi = 1$. We then feed these into (5.8) to find the corresponding value of σ_b. In this manner, we can find arbitrarily many points on the edge of chaos (EOC). Within some finite range of σ_w^2 values, we can find a good fit to the EOC. In the range $1 \leq \sigma_w^2 \leq 10$, a good polynomial fit is

$$\text{EOC}: \quad \sigma_b^2 = \sum_{n=2}^{9} \frac{c_n}{n!} (\sigma_w^2 - 1)^n , \tag{5.13}$$

with fit coefficients

n	c_n	n	c_n
2	0.0190	6	−1.15
3	0.778	7	0.769
4	−1.07	8	−0.328
5	1.25	9	0.0672

(5.14)

Of course, we can reduce the number of fit coefficients needed by reducing the range of σ_w^2 values over which we require the fit to be good.

The form of this fit is designed such that it contains the point $(\sigma_w^2, \sigma_b^2) = (1, 0)$, and that the edge of chaos has zero slope at this point. We justify these conditions analytically in appendix 5.6.2. In the right plot in fig. 5.1, we present a contour plot of χ. Again, the edge of chaos is drawn as a solid black line and the line of uniformity as a dashed line. The point of intersection of the edge of chaos and line of uniformity is found to be

$$(\sigma_w^2, \sigma_b^2)_{\text{intersect}} = (2.00, 0.104) . \tag{5.15}$$

[7] See [99] for a pedagogical explanation.

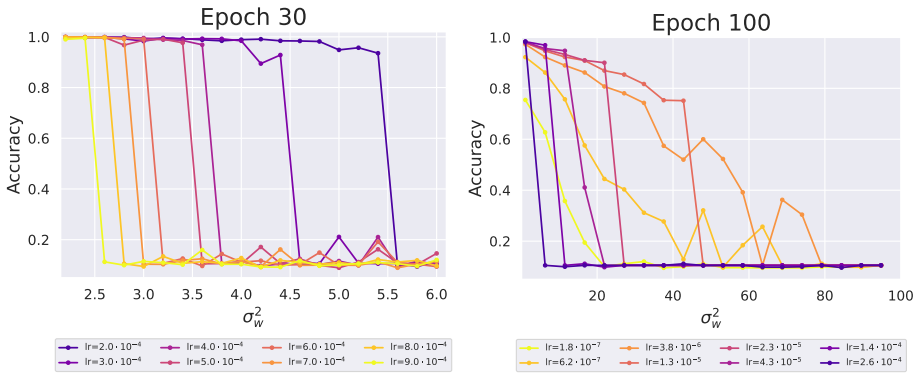


Figure 5.2: Accuracies on MNIST for distributions of initial weights along the edge of chaos in a deep ($L = 100$), wide ($N = 784$) neural network with tanh activation function, for a range of learning rates, after 30 epochs (left) and 100 epochs (right). We observe a drop-off in accuracy beyond a value σ_w^2 which is up to an order of magnitude larger than the point at which the line of uniformity is crossed. For learning rates of the order typically used in the literature, this point is near the intersection of the LOU and the EOC, but moves to higher values of σ_w^2 for smaller learning rates. When learning rates become extremely small ($r < 10^{-5}$), learning becomes highly inefficient, and the drop-off less sharp for the training duration considered. Networks were trained via stochastic gradient descent with batch size 64 and momentum 0.8.

5.4. The impact of uniformity along the edge of chaos

To the right of the line of uniformity, neurons begin to saturate the tanh activation function, i.e., approach ± 1. This implies that backpropagation based on gradient descent should be less efficient, and hence networks should reach a lower accuracy in a fixed amount of training time. The Google Brain collaboration has already established that at the edge of chaos, learning accuracy is enhanced due to polynomial rather than exponential decay of correlations as a function of network depth [15]. Combining the two insights, optimal learning should therefore take place on the edge of chaos near the line of uniformity.

To test this hypothesis, we have performed the MNIST image classification task in networks ranging up to a depth of $L = 100$ hidden layers at various points along the edge of chaos. The resulting learning accuracy is shown in fig. 5.2. We see that this expectation is partially validated. On the left side of the line of uniformity – but to the right of the linear regime – all points on the edge of chaos are equally good at learning. But beyond a certain point, which lies to the right of the intersection point (5.15) of the edge of chaos and line of uniformity, the final accuracy decreases. However, this drop-off point is substantially (up to an order of magnitude) displaced to the right of the intersection point, indicating that the line of uniformity is perhaps better thought of as a region rather than a narrow band, and depends on hyperparameters (such as the learning rate) as mentioned above. Nevertheless, for typical learning rates used in the literature of order 10^{-3}, such as used in [15], the drop-off

point at approximately $\sigma_w^2 \sim 2.5$ is indeed fairly close to the intersection between the line of uniformity and the edge of chaos at $\sigma_w^2 = 2$.

We repeated this exercise for the CIFAR-10 image classification task, and present the corresponding results in fig. 5.3. We converted the colored images to grayscale to reduce the input size by a factor of 3. The drop-off in accuracy along the edge of chaos towards larger values of σ_w^2 is still present, though the effect is not as dramatic as it is for MNIST. This is not surprising as CIFAR is a much more difficult task than MNIST and so we expect that the saturation of slightly more or fewer neurons will have a much less decisive effect. We note however that in the regime of extremely small learning rates, where training MNIST becomes highly inefficient, the MNIST and CIFAR results appear similar insofar as neither exhibits the obvious sharp drop-off observed for MNIST at the higher learning rates generally used in practice.

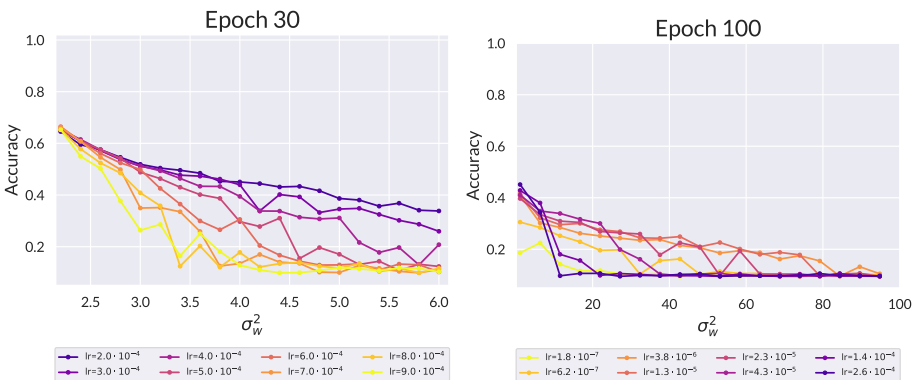


Figure 5.3: Accuracies on CIFAR-10 for distributions of initial weights along the edge of chaos in a deep ($L = 100$), wide ($N = 1024$) neural network with tanh activation function, for a range of learning rates, after 30 epochs (left) and 100 epochs (right). The drop-off in accuracy towards higher values of σ_w^2 here is much more gradual than the sharp drop-offs observed for MNIST in fig. 5.2 (at all but the lowest learning rates, $r < 10^{-5}$). Networks were trained via stochastic gradient descent with batch size 64 and momentum 0.8.

Thus, the line of uniformity is not a sharp boundary, unlike the edge of chaos. This is somewhat inherent in its definition, which selects proximity to the uniform distribution of final hidden layer weights as a condition for efficient learning based on the entropic argument given above, but does not specify any particular fall-off behavior. The line of uniformity does, however, give an estimate of where the saturation of the activation function should start to affect learning, and by extension, the point at which saturation of the activation function begins to hinder learning efficiency. To summarize: on the left side of the line of uniformity, the distributions are sufficiently narrow that saturation of the tanh activation function does not occur, and all initial weight distributions along the EOC learn equally well. Conversely, on the right side of uniformity, neurons saturate the activation function and hence hamper learning, even along the EOC. This is our main observation. Importantly, we note that the

studies by [15] were performed to the left of the point where the line of uniformity crosses the edge of chaos and hence at optimal efficiency.

Before moving on to our final set of experiments, we note that the above conclusion is of course specific to saturating activation functions, specifically tanh. This is one motivation for the use of non-saturating activation functions such as ReLU or SWISH, though the unbounded nature of such functions presents its own set of training difficulties. While a similar analysis of uniformity, as quantified by the maximally entropic distribution, for non-saturating activation functions is beyond the scope of this work, a brief inspection of learning efficiency along the EOC for SWISH shows no loss of accuracy in agreement with the absence of saturation effects; see appendix 5.6.4.[8]

5.5. Uniformity away from the EOC

Thus far, we have examined the impact of uniformity on training efficiency along the edge of chaos. Now, we would like to explore whether the line of uniformity still affords training advantages even for networks initialized far from criticality. In attempting to exhibit this however, one quickly finds that the edge of chaos represents a far more dominant effect than the line of uniformity. A close inspection of the learning accuracy of deep (L=300) and wide (N=784) MNIST learning networks shows that there is no discernible difference in learning accuracy away from the edge of chaos: it is simply poor everywhere (see fig. 5.7 in appendix 5.6.3, also [15].) This can be understood from the form of the correlation functions: away from the edge of chaos, correlations damp exponentially $\sim e^{-L/\xi}$. For a deep network, this exponential damping will erase any finer difference in accuracy results. Along the edge of chaos, the damping is only polynomial and, therefore, the finer difference remains, as seen in fig. 5.2. In shallow networks however, the exponential damping does not have sufficient time to compound, and if the network is also narrow and hence has low expressibility per layer, we can explore the effect of uniformity even away from criticality in such models.

Furthermore, it is common lore that efficient backpropagation needs sufficient gradients, and that such gradients are absent if most of the post-activation functions saturate to a fixed asymptotic value. However, if a sufficient number of weight and bias values are such that there remain trainable paths through a saturated landscape, the model will still learn, even though, distribution-wise, most of the neurons have saturated. Therefore, the inefficiency due to saturation discussed above can be displayed more clearly by choosing narrower networks with smaller N, where we might expect that uniformity – that is, maximally entropic distributions – may afford the

[8] For both SWISH and tanh, the edge of chaos is a line of critical initalizations through phase space, while for ReLU it is only a single point [97].

most advantage.

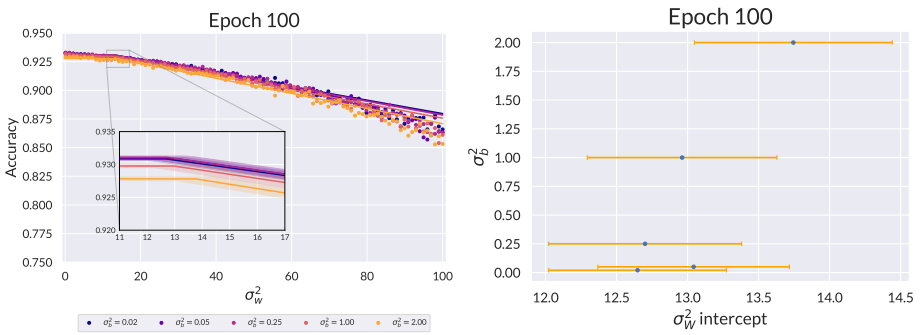


Figure 5.4: For small networks, the learning efficiency exhibits threshold behavior as a function of σ_w^2. Shown are results for MNIST trained on a $N = 8$, $L = 1$ network sampled over 50 network initializations. The inset shows the fits in the threshold region. The bottom figure shows that the location of this threshold in σ_w^2 decreases with increasing σ_b^2 consistent with the trend implied by the line of uniformity threshold. As explained in the text, there is a multiplicative factor involved and the large-N analysis cannot be straightforwardly transplanted to this small-N case. The uncertainty bars are propagated from the uncertainties in the accuracy versus σ_w^2 data points.

The effect of lying near uniformity is therefore strongest in shallow, narrow networks rather than deep, wide networks where the edge of chaos effect dominates. For these small networks, some of the asymptotic analysis above locating the LOU and EOC does not immediately apply, since the network is unable to reach the asymptotic value σ_*^2 of the pre-activation variance.[9] At the same time, the input variance and mean, σ_0^2 and μ_0, actually *do* matter in this case and, with this information, we can roughly estimate the location of the line of uniformity. For example, for $L = 1$, we have $\sigma_1^2 = \sigma_w^2(\sigma_0^2+\mu_0^2)+\sigma_b^2$ and the line of uniformity would be where $\sigma_1^2 = \sigma_{\text{min}}^2 = \frac{\pi^2}{12}$. For example, for MNIST, $\sigma_0^2 \approx 0.095$ and $\mu_0^2 \approx 0.017$, so the line of uniformity can be estimated as $\sigma_b^2 \approx \frac{\pi^2}{12} - 0.112\sigma_w^2$. Equivalently, for fixed σ_b^2, this gives a σ_w^2-threshold of $\sigma_w^2 \sim 7.35 + 8.93\sigma_b^2$ beyond which we expect saturation effects to decrease training efficiency. For $L = 2$, we would iterate the above process once more, passing through the activation function; this gives an estimated threshold of $\sigma_w^2 \approx 3.5 + 8.93\sigma_b^2$.

Results for $L = 1$ are shown in fig. 5.4, and results for $L = 2$ are shown in fig. 5.5. As predicted, we observe that the accuracy retains a high, approximately constant value up to a σ_b^2-dependent threshold for σ_w^2, and then decays approximately linearly thereafter. To determine the threshold empirically, we fit the data to a function

[9] In this sense, we may take "shallow" to mean $L \leq 5$, since as shown in fig. 5.6, by $L \approx 6$, the network has reached σ_*^2. Strictly speaking however, the predictions for the EOC as well as the LOU are ill-defined in narrow networks, since these are no longer approximately Gaussian, and also appear to be beyond the reach of current perturbative approaches [98].

of the form

$$A_{\text{fit}}(\sigma_w^2) = A_{\max} - r(\sigma_w^2 - \sigma_{w,\text{thr}}^2)\,\Theta(\sigma_w^2 - \sigma_{w,\text{thr}}^2), \tag{5.16}$$

where $A_{\max}$ is the maximum accuracy, $\sigma_{w,\text{thr}}^2$ is the threshold value, r is the rate of linear decay, and Θ is the Heaviside step function. Each accuracy vs. σ_w^2 data point is an average over 20 instantiations of the network and thus comes with its own variance. These propagate into uncertainty bars for the three fit parameters. We plot the threshold for different values of σ_b^2 in fig. 5.4 for $L = 1$. This qualitatively confirms our expectations, though the empirical value of the threshold is about a factor of 2 greater than the analytical prediction, and the slope about a factor of 8 smaller. However, given that we are applying a large-N analysis to a relatively narrow network ($N = 8$), an $\mathcal{O}(1)$ quantitative discrepancy is reasonable. For $L = 2$ the corresponding results are presented in fig. 5.5, again showing qualitative agreement. The empirical threshold in this case is about a factor of 4 greater than the theoretical value, and the slope is a factor of 8 smaller.

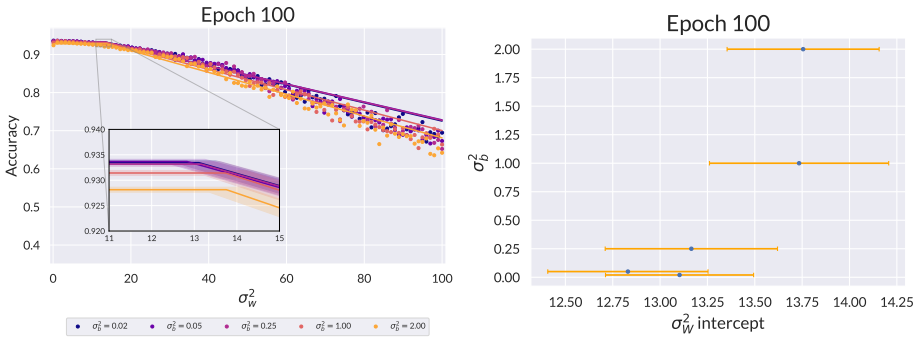


Figure 5.5: Small-network threshold behavior as in fig. 5.4, for MNIST trained on a network with $N = 8$ and $L = 2$, sampled over 50 initial conditions drawn from (σ_w^2, σ_b^2). The bottom figure shows that the location of this threshold decreases with σ_b^2 consistent with the trend implied by the line of uniformity.

Conclusion In this work, we establish that for deep random feedforward networks along the edge of chaos, the efficiency of training via stochastic gradient descent still depends on non-saturation of the activation function. Similar points have been made previously in [95, 96], which compared the performance of difference activation functions initialized at one point on their respective edges of chaos. However, what we demonstrate for the tanh activation function is that not all points on the edge of chaos are equally efficient at learning. Within a fixed number of training epochs (~ 100), activation function saturation eventually impedes learning if we push the weight and bias variances too far to the right of the line of uniformity, defined to be where the final layer post-activation is most uniformly distributed, i.e., maximally entropic. Unlike the edge of chaos, which separates chaotic and ordered outputs, the line of uniformity does not mark an abrupt change in the overall be-

havior of the network. Rather, it simply indicates roughly the point where the saturation of the activation function begins to impede learning. We demonstrate this for shallow and narrow networks as well, where the exponential damping of neuron correlations away from the edge of chaos becomes much less of a decisive factor in determining training efficiency.

Acknowledgments

This research was supported in part by the Dutch Research Council (NWO) project 680-91-116 (*Planckian Dissipation and Quantum Thermalisation: From Black Hole Answers to Strange Metal Questions.*) and by the Dutch Research Council (NWO)/Ministry of Education. K.T.G. has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101024967.

5.6. Appendix

5.6.1. Independence of σ_*^2 on σ_1^2

The exact pre- or post-activation distribution at a given layer obviously does depend on σ_1^2, the pre-activation variance at the first hidden layer. This dependence is generated via the recursion relation (5.8). However, at the fixed point, the asymptotic distributions do not depend on σ_1^2. Indeed, the relation that the asymptotic pre-activation variance satisfies is eq. (5.9), which does not depend on σ_1^2 at all. We can demonstrate this fact by plotting the evolution of the post-activation distribution for fixed σ_w^2 and σ_b^2, but for many values of σ_1^2. In fig. 5.6, we show this for $(\sigma_w^2, \sigma_b^2) = (1.76, 0.05)$ for several values of σ_1^2, both less than and greater than σ_*^2 which turns out to be $\sigma_*^2 \approx 0.57$ in this case. When $\sigma_1^2 < \sigma_*^2$, the post-activation distribution starts out narrower and spreads out, whereas when $\sigma_1^2 > \sigma_*^2$ it starts out more peaked at ± 1 and then flattens out.

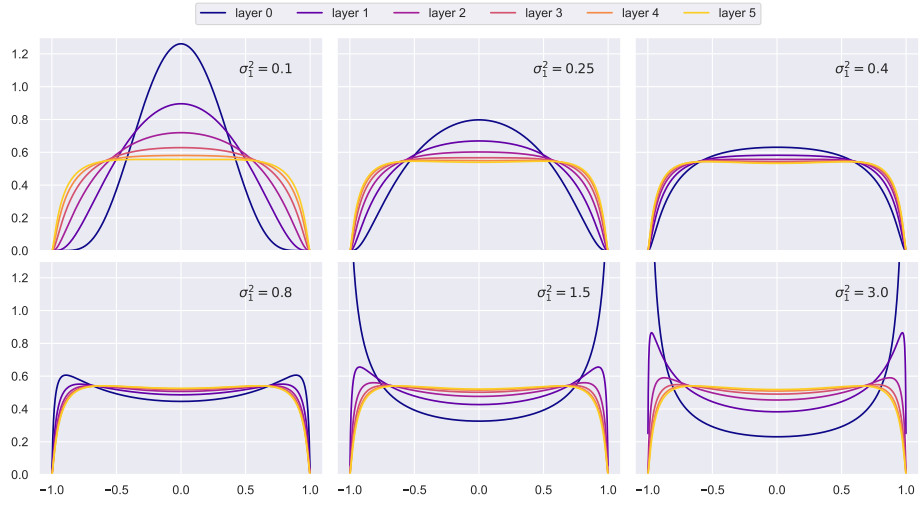


Figure 5.6: Layer-to-layer evolution of the post-activation distribution at $(\sigma_w^2, \sigma_b^2) = (1.76, 0.05)$ for six different values of the first hidden layer pre-activation variance σ_1^2. The post-activations converge to the asymptotic distribution within about five layers.

5.6.2. Analytic Details of the Fixed Point Computation

In this appendix, we will show that the edge of chaos contains the point $(\sigma_w^2, \sigma_b^2) = (1, 0)$ and has zero slope there. At this point, the fixed-point equation (5.9) reads

$$\sigma_*^2 = \sigma_{\phi,*}^2 \,. \tag{5.17}$$

The left-hand side is the fixed-point pre-activation variance, whereas the right-hand side is the corresponding post-activation variance. As long as $|\phi(z)| < |z|$, which is the case for $\phi(z) = \tanh(z)$ except at $z = 0$, the variance of the post-activation will always be smaller than that of the pre-activation. Therefore, the only solution at this point is $\sigma_*^2 = \sigma_{\phi,*}^2 = 0$ and thus at this point $\phi'(\sigma_* z) = \operatorname{sech}^2(0) = 1$ and $\chi = 1$ or $\xi = \infty$. Hence, this point is on the edge of chaos.

Now, consider eq. (5.10), but now along the edge of chaos rather than the lines of constant σ_*^2. Let σ_w^2 be our independent parameter along the edge of chaos and take a derivative with respect to it:

$$\frac{\partial \sigma_b^2}{\partial \sigma_w^2} = \left(1 - \sigma_w^2 \frac{\partial \sigma_{\phi,*}^2}{\partial \sigma_*^2}\right) \frac{\partial \sigma_*^2}{\partial \sigma_w^2} - \sigma_{\phi,*}^2 , \tag{5.18}$$

where we have used the fact that $\sigma_{\phi,*}^2$ depends on σ_w^2 only through its dependence on σ_*^2.

5

To compute the derivative $\frac{\partial \sigma_{\phi,*}^2}{\partial \sigma_*^2}$, it is convenient to first rewrite the integral expression for σ_ϕ^2 in (5.4) by changing back to the original pre-activation variable z:

$$\sigma_\phi^2 = \int_{-1}^{1} \mathrm{d}x\, p_\phi(x;\sigma^2)\, x^2 = \int_{-\infty}^{\infty} \mathrm{d}z\, p(z;\sigma^2)\, \phi(z)^2 . \tag{5.19}$$

We can easily compute the various derivatives of the pre-activation distribution:

$$\frac{\partial p(z;\sigma^2)}{\partial \sigma^2} = \left(\frac{z^2}{\sigma^2} - 1\right) \frac{p(z;\sigma^2)}{2\sigma^2}, \quad \frac{\partial^2 p(z;\sigma^2)}{\partial z^2} = \left(\frac{z^2}{\sigma^2} - 1\right) \frac{p(z;\sigma^2)}{\sigma^2} = 2\, \frac{\partial p(z;\sigma^2)}{\partial \sigma^2} . \tag{5.20}$$

Therefore, using integration by parts, and the fact that we can ignore boundary terms due to the fast fall-off of the Gaussian, we find

$$\frac{\partial \sigma_\phi^2}{\partial \sigma^2} = \int \mathrm{d}z\, \frac{\partial p(z;\sigma^2)}{\partial \sigma^2}\, \phi(z)^2 = \frac{1}{2} \int \mathrm{d}z\, \frac{\partial^2 p(z;\sigma^2)}{\partial z^2}\, \phi(z)^2 = \int \mathrm{d}z\, p(z;\sigma^2) \big(\phi'(z)^2 + \phi(z)\, \phi''(z)\big) . \tag{5.21}$$

By rescaling the variable to σz, the first integral term above can be written as

$$\int \mathrm{d}z\, p(z;\sigma^2) \phi'(z)^2 = \int \mathscr{D}z \left[\phi'(\sigma z)\right]^2 . \tag{5.22}$$

Note that when this is evaluated at σ_*^2 and multiplied by σ_w^2, we get precisely χ, as defined in (5.12). Let us give a name to the remaining integral in (5.21) evaluated at σ_*^2. For future convenience, we will put a relative minus sign in the definition below, the reason being that, for $\phi = \tanh$, the object $\phi\,\phi''$ is *negative* semi-definite:

$$\tilde{\chi} = -\sigma_w^2 \int \mathrm{d}z\, p(z;\sigma_*^2)\, \phi(z)\, \phi''(z) = -\sigma_w^2 \int \mathscr{D}z \phi(\sigma_* z)\, \phi''(\sigma_* z) . \tag{5.23}$$

Then, (5.21) evaluated at σ_*^2 and multiplied by σ_w^2 reads

$$\sigma_2^2 \frac{\partial \sigma_{\phi,*}^2}{\partial \sigma_*^2} = \chi - \tilde{\chi} \,. \tag{5.24}$$

Now, let us define

$$\tilde{\xi} = -\frac{1}{\ln(\chi - \tilde{\chi})} \,. \tag{5.25}$$

This is precisely the object called ξ_q in [15], which is the length scale that controls the exponential decay of information propagation through the neural network from a single input.

Plugging eq. (5.24) back into eq. (5.18) gives

$$\frac{\partial \sigma_b^2}{\partial \sigma_w^2} = (1 - \chi + \tilde{\chi}) \frac{\partial \sigma_*^2}{\partial \sigma_w^2} - \sigma_{\phi,*}^2 \,. \tag{5.26}$$

Along the edge of chaos, $\chi = 1$, and so

$$\frac{\partial \sigma_b^2}{\partial \sigma_w^2} = \tilde{\chi} \frac{\partial \sigma_*^2}{\partial \sigma_w^2} - \sigma_{\phi,*}^2 \,, \tag{5.27}$$

Now, we can establish a simple bound on $\tilde{\chi}$ by virtue of the fact that $|\phi(z)| \leq |z|$, for $\phi = \tanh$. To do this, let us first rewrite $\tilde{\chi}$ using the identity

$$\phi''(z) = -2 \tanh(z) \operatorname{sech}^2(z) = -2 \phi(z) \phi'(z) \,. \tag{5.28}$$

Therefore,

$$\phi(z) \phi''(z) = -2 \phi(z)^2 \phi'(z) = -\frac{2}{3} \left[\phi(z)^3\right]' \,, \tag{5.29}$$

and

$$\tilde{\chi} = \frac{2\sigma_2^2}{3} \int \mathrm{d}z \, p(z; \sigma_*^2) \left[\phi(z)^3\right]' = -\frac{2\sigma_w^2}{3} \int \mathrm{d}z \, \frac{\partial p(z; \sigma_*^2)}{\partial z} \phi(z)^3 = \frac{2\sigma_w^2}{3\sigma_*^2} \int \mathrm{d}z \, p(z; \sigma_*^2) \, z \phi(z)^3 \,. \tag{5.30}$$

Therefore, since $|\phi(z)| \leq |z|$ for $\phi = \tanh$,

$$0 \leq \tilde{\chi} \leq \frac{2\sigma_w^2}{3\sigma_*^2} \int \mathrm{d}z \, p(z; \sigma_*^2) \, z^4 = 2 \sigma_w^2 \sigma_*^2 \,. \tag{5.31}$$

Therefore, since we have already shown that $\sigma_*^2 = \sigma_{\phi,*}^2 = 0$ at the point $(\sigma_w^2, \sigma_b^2) = (1, 0)$, it follows that $\tilde{\chi} = 0$ at this point as well and, from eq. (5.27),

$$\left. \frac{\partial \sigma_b^2}{\partial \sigma_w^2} \right|_{(\sigma_w^2, \sigma_b^2) = (1,0)} = 0 \,. \tag{5.32}$$

In other words, the edge of chaos has zero slope at the point $(\sigma_w^2, \sigma_b^2) = (1, 0)$.

5.6.3. Implementation Details

Throughout this work, we have used a vanilla feedforward neural network of L hidden layers, each having the same depth N. As described, initial weights and biases are drawn from zero-mean Gaussian distributions with $\frac{\sigma_w^2}{N}$ and σ_b^2 respectively. Both MNIST and CIFAR-10 were trained using the standard cross-entropy loss function and no optimizer. This reproduces the results of [15] (see fig. 5.7), confirming critical behavior.

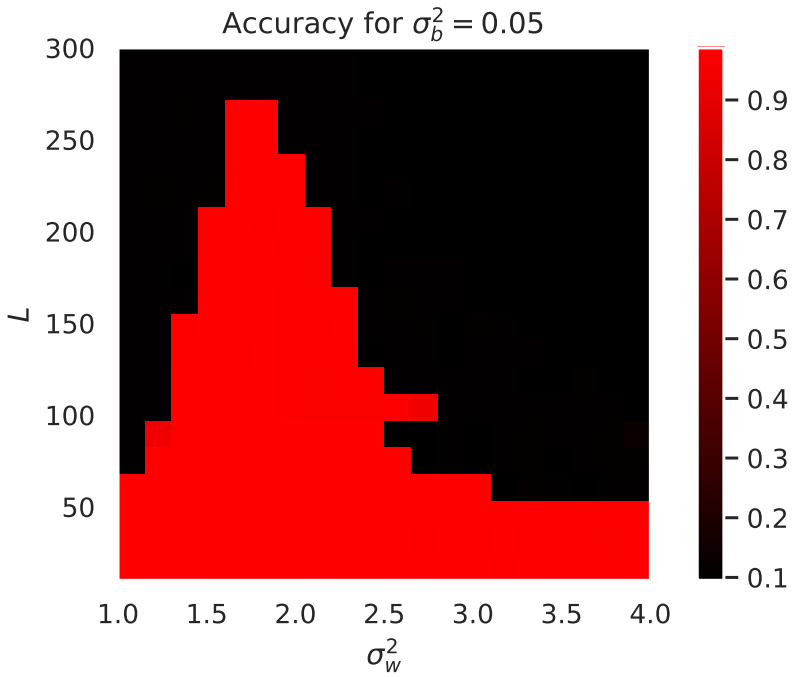


Figure 5.7: Optimal learning for deep neural networks at the edge of chaos as first shown by [15]. Shown is learning efficiency for MNIST training as a function of network depth L with $N = 784$ and choice of initial weight distribution σ_w^2 holding the initial bias distribution $\sigma_b^2 = 0.05$ fixed. At the edge of chaos $(\sigma_w^2, \sigma_b^2) = (1.76, 0.05)$, learning remains efficient even for very deep networks, but eventually ($L \sim 270$) goes down. This same behavior has been observed for deep feedforward networks in [15, 92]. The learning rate used is $\ell = 10^{-3}$ for $L < 100$ and $\ell = 10^{-4}$ for $L \geq 100$.

5.6.4. SWISH activation function

Throughout the text, we examined the impact of saturation via the line of uniformity for the tanh activation function. For non-saturating activation functions, it is an open question whether a similar notion of uniformity exists. While a full analysis of this is beyond the scope of this work, in this appendix we offer some preliminary

results for the SWISH activation function,

$$\text{swish}(z) = \frac{z}{1+e^{-z}}\,, \tag{5.33}$$

which also features a line of critical points separating an ordered and chaotic phase. Note that unlike the EOC for tanh, which increases with increasing σ_w^2, the EOC for SWISH decreases with increasing σ_w^2, which prevents us from examining the impact of large weight variances. Conversely, for small values of σ_w^2, the corresponding value of σ_b^2 becomes so large that we are unable to satisfy the critical detection criteria $\chi = 1$ discussed in the main text.[10] The EOC for SWISH is plotted in fig. 5.8, which shows a computable range of approximately $\sigma_w^2 \in [1.97, 3.4]$. The same figure also shows the accuracy for an $L = 40$ network with SWISH activation function trained along the EOC, demonstrating no deterioration of performance within this range, which confirms the absence of saturation effects. See also [95, 96].

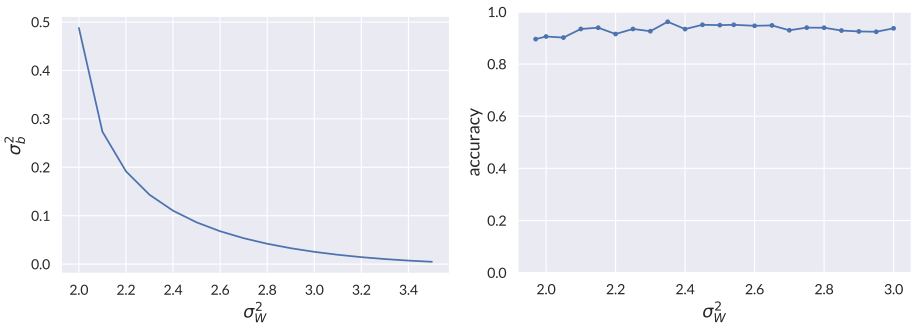


Figure 5.8: (Left) Edge of chaos for SWISH activation function. (Right) Accuracy for a feedforward network with $L = 40$ layers trained on MNIST for 21 equally-spaced points along the SWISH EOC. Over the limited range for which the criticality condition $\chi = 1$ is satisfied, we observe no significant differences in accuracy, though a slightly lower learning rate was used for the left-most two points; we believe this to be due to the large values of σ_b^2 in this regime.

[10] We do not claim that the EOC stops beyond this point, rather that it cannot be computed from the central limit method used in [14, 15]. It is conceivable that this could be computed via the NN/QFT correspondence developed in [98], but this has not been attempted for SWISH.

Bibliography

[1] S. Sachdev, *Quantum Phase Transitions*. Cambridge University Press, 2 ed., 2011.

[2] M. Newman and G. Barkema, *Monte Carlo Methods in Statistical Physics*. Clarendon Press, 1999.

[3] K. Binder, *Critical properties from monte carlo coarse graining and renormalization*, *Phys. Rev. Lett.* **47** (Aug, 1981) 693–696.

[4] G. Carleo and M. Troyer, *Solving the quantum many-body problem with artificial neural networks*, *Science* **355** (2017), no. 6325 602–606, [https://www.science.org/doi/pdf/10.1126/science.aag2302].

[5] A. Dawid, J. Arnold, B. Requena, A. Gresch, M. Płodzień, K. Donatella, K. A. Nicoli, P. Stornati, R. Koch, M. Büttner, R. Okuła, G. Muñoz-Gil, R. A. Vargas-Hernández, A. Cervera-Lierta, J. Carrasquilla, V. Dunjko, M. Gabrié, P. Huembeli, E. van Nieuwenburg, F. Vicentini, L. Wang, S. J. Wetzel, G. Carleo, E. Greplová, R. Krems, F. Marquardt, M. Tomza, M. Lewenstein, and A. Dauphin, *Modern applications of machine learning in quantum sciences*, 2022.

[6] K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators*, *Neural Networks* **2** (1989), no. 5 359–366.

[7] M. Srednicki, *Chaos and quantum thermalization*, *Phys. Rev. E* **50** (Aug, 1994) 888–901.

[8] P. Sabella-Garnier, K. Schalm, T. Vakhtel, and J. Zaanen, *Thermalization/relaxation in integrable and free field theories: an operator thermalization hypothesis*, 2019.

[9] T. LeBlond and M. Rigol, *Eigenstate thermalization for observables that break hamiltonian symmetries and its counterpart in interacting integrable systems*, *Phys. Rev. E* **102** (Dec, 2020) 062113.

[10] F. J. Wegner, *Duality in generalized ising models and phase transitions without local order parameters*, *Journal of Mathematical Physics* **12** (1971), no. 10 2259–2272, [https://doi.org/10.1063/1.1665530].

[11] J. B. Kogut, *An introduction to lattice gauge theory and spin systems*, *Rev. Mod. Phys.* **51** (Oct, 1979) 659–713.

[12] E. Fradkin and S. H. Shenker, *Phase diagrams of lattice gauge theories with higgs fields*, *Phys. Rev. D* **19** (Jun, 1979) 3682–3697.

[13] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, *A high-bias, low-variance introduction to machine learning for physicists*, *Physics Reports* **810** (may, 2019) 1–124.

[14] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, *Exponential expressivity in deep neural networks through transient chaos*, .

[15] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, *Deep Information Propagation*, *arXiv e-prints* (Nov., 2016) arXiv:1611.01232, [arXiv:1611.01232].

[16] A. Bukva, P. Sabella-Garnier, and K. Schalm, *Operator thermalization vs eigenstate thermalization*, 2019.

[17] T. Langen, R. Geiger, and J. Schmiedmayer, *Ultracold atoms out of equilibrium*, *Annual Review of Condensed Matter Physics* **6** (2015), no. 1 201–217, [https://doi.org/10.1146/annurev-conmatphys-031214-014548].

[18] A. M. Kaufman, M. E. Tai, A. Lukin, M. Rispoli, R. Schittko, P. M. Preiss, and M. Greiner, *Quantum thermalization through entanglement in an isolated many-body system*, *Science* **353** (2016), no. 6301 794–800, [https://science.sciencemag.org/content/353/6301/794.full.pdf].

[19] P. Bordia, H. Lüschen, S. Scherg, S. Gopalakrishnan, M. Knap, U. Schneider, and I. Bloch, *Probing slow relaxation and many-body localization in two-dimensional quasiperiodic systems*, *Phys. Rev. X* **7** (Nov, 2017) 041047.

[20] G. Bentsen, T. Hashizume, A. S. Buyskikh, E. J. Davis, A. J. Daley, S. S. Gubser, and M. Schleier-Smith, *Treelike interactions and fast scrambling with cold atoms*, *Phys. Rev. Lett.* **123** (Sep, 2019) 130601.

[21] H. Labuhn, D. Barredo, S. Ravets, S. De Léséleuc, T. Macrì, T. Lahaye, and A. Browaeys, *Tunable two-dimensional arrays of single rydberg atoms for realizing quantum ising models*, *Nature* **534** (2016), no. 7609 667.

[22] C. J. Turner, A. A. Michailidis, D. A. Abanin, M. Serbyn, and Z. Papić, *Weak ergodicity breaking from quantum many-body scars*, *Nature Physics* **14** (2018), no. 7 745–749.

[23] A. Dymarsky, N. Lashkari, and H. Liu, *Subsystem ETH*, *Phys. Rev.* **E97** (2018) 012140, [arXiv:1611.08764].

[24] E. M. Brehm, D. Das, and S. Datta, *Probing thermality beyond the diagonal*, *Phys. Rev.* **D98** (2018), no. 12 126015, [arXiv:1804.07924].

[25] A. Romero-Bermúdez, P. Sabella-Garnier, and K. Schalm, *A Cardy formula for off-diagonal three-point coefficients; or, how the geometry behind the horizon gets disentangled*, *JHEP* **09** (2018) 005, [arXiv:1804.08899].

[26] Y. Hikida, Y. Kusuki, and T. Takayanagi, *Eigenstate thermalization hypothesis and modular invariance of two-dimensional conformal field theories*, *Phys. Rev.* **D98** (2018), no. 2 026003, [arXiv:1804.09658].

[27] T. Anous and J. Sonner, *Phases of scrambling in eigenstates*, [arXiv:1903.03143].

[28] P. Nayak, J. Sonner, and M. Vielma, *Eigenstate Thermalisation in the conformal Sachdev-Ye-Kitaev model: an analytic approach*, [arXiv:1903.00478].

[29] M. Besken, S. Datta, and P. Kraus, *Quantum thermalization and virasoro symmetry*, 2019.

[30] J. M. Deutsch, *Quantum statistical mechanics in a closed system*, *Phys. Rev. A* **43** (Feb, 1991) 2046–2049.

[31] M. Srednicki, *Chaos and quantum thermalization*, *Phys. Rev. E* **50** (Aug, 1994) 888–901.

[32] V. I. Yukalov, *Equilibration and thermalization in finite quantum systems*, *Laser Phys. Lett.* **8** (2011) 485–507, [arXiv:1201.2781].

[33] L. D'Alessio, Y. Kafri, A. Polkovnikov, and M. Rigol, *From quantum chaos and eigenstate thermalization to statistical mechanics and thermodynamics*, *Adv. Phys.* **65** (2016), no. 3 239–362, [arXiv:1509.06411].

[34] S. Grozdanov and J. Polonyi, *Dynamics of the electric current in an ideal electron gas: A sound mode inside the quasiparticles*, *Phys. Rev.* **D92** (2015), no. 6 065009, [arXiv:1501.06620].

[35] I. Amado, B. Sundborg, L. Thorlacius, and N. Wintergerst, *Black holes from large N singlet models*, *JHEP* **03** (2018) 075, [arXiv:1712.06963].

[36] D. E. Parker, X. Cao, A. Avdoshkin, T. Scaffidi, and E. Altman, *A Universal Operator Growth Hypothesis*, [arXiv:1812.08657].

[37] M. Medenjak, B. Buca, and D. Jaksch, *The isolated heisenberg magnet as a quantum time crystal*, [1905.08266].

[38] S. Banerjee, J. Engelsöy, J. Larana-Aragon, B. Sundborg, L. Thorlacius, and N. Wintergerst, *Quenched coupling, entangled equilibria, and correlated composite operators: a tale of two $O(N)$ models*, [arXiv:1903.12242].

[39] B. Craps, M. De Clerck, O. Evnin, and S. Khetrapal, *Energy level splitting for weakly interacting bosons in a harmonic trap*, [arXiv:1903.04974].

[40] P. Sabella-Garnier, K. Schalm, T. Vakhtel, and J. Zaanen, *Thermalization/Relaxation in integrable and free field theories: an Operator Thermalization Hypothesis*, [arXiv:1906.02597].

[41] D. Fioretto and G. Mussardo, *Quantum quenches in integrable field theories*, *New Journal of Physics* **12** (may, 2010) 055015.

[42] A. C. Cassidy, C. W. Clark, and M. Rigol, *Generalized thermalization in an integrable lattice system*, *Phys. Rev. Lett.* **106** (Apr, 2011) 140405.

[43] V. Alba, *Eigenstate thermalization hypothesis and integrability in quantum spin chains*, *Phys. Rev. B* **91** (Apr, 2015) 155123.

[44] J. Riddell and M. P. Müller, *Generalized eigenstate typicality in translation-invariant quasifree fermionic models*, *Phys. Rev. B* **97** (Jan, 2018) 035129.

[45] S. Byju, K. Lochan, and S. Shankaranarayanan, *Generalized thermalization in quenched free Fermionic models*, [arXiv:1808.07742].

[46] P. Banerjee, A. Gaikwad, A. Kaushal, and G. Mandal, *Quantum quench and thermalization to GGE in arbitrary dimensions and the odd-even effect*, [arXiv:1910.02404].

[47] M. C. Bañuls, J. I. Cirac, and M. B. Hastings, *Strong and weak thermalization of infinite nonintegrable quantum systems*, *Phys. Rev. Lett.* **106** (Feb, 2011) 050405.

[48] A. Bukva, K. Schalm, and J. Zaanen, *Replicating Higgs fields in Ising gauge theory: The registry order*, May, 2023.

[49] X. G. Wen, *Mean-field theory of spin-liquid states with finite energy gap and topological orders*, *Phys. Rev. B* **44** (Aug, 1991) 2664–2672.

[50] F. Alexander Bais, P. van Driel, and M. de Wild Propitius, *Quantum symmetries in discrete gauge theories*, *Physics Letters B* **280** (1992), no. 1 63–70.

[51] T. Hansson, V. Oganesyan, and S. Sondhi, *Superconductors are topologically ordered*, *Annals of Physics* **313** (2004), no. 2 497–538.

[52] T. Senthil and M. P. A. Fisher, *Z_2 gauge theory of electron fractionalization in strongly correlated systems*, *Phys. Rev. B* **62** (Sep, 2000) 7850–7881.

[53] H. V. Kruis, I. P. McCulloch, Z. Nussinov, and J. Zaanen, *Geometry and topological order in the luttinger liquid state*, *Europhysics Letters* **65** (feb, 2004) 512.

[54] Y. Zhang, E. Demler, and S. Sachdev, *Competing orders in a magnetic field: Spin and charge order in the cuprate superconductors*, *Phys. Rev. B* **66** (Sep, 2002) 094501.

[55] A. Kitaev, *Fault-tolerant quantum computation by anyons*, *Annals of Physics* **303** (2003), no. 1 2–30.

[56] P. E. Lammert, D. S. Rokhsar, and J. Toner, *Topology and nematic ordering, Phys. Rev. Lett.* **70** (Mar, 1993) 1650–1653.

[57] K. Liu, J. Nissinen, R.-J. Slager, K. Wu, and J. Zaanen, *Generalized liquid crystals: Giant fluctuations and the vestigial chiral order of i, o, and t matter, Phys. Rev. X* **6** (Oct, 2016) 041025.

[58] D. A. Huse and S. Leibler, *Are sponge phases of membranes experimental gauge-higgs systems?, Phys. Rev. Lett.* **66** (Jan, 1991) 437–440.

[59] E. Babaev, A. Sudbø, and N. W. Ashcroft, *A superconductor to superfluid phase transition in liquid metallic hydrogen, Nature* **431** (Oct., 2004) 666–668.

[60] E. Babaev and N. W. Ashcroft, *Violation of the london law and onsager–feynman quantization in multicomponent superconductors, Nature Physics* **3** (June, 2007) 530–533.

[61] E. Babaev, J. Carlström, and M. Speight, *Type-1.5 superconducting state from an intrinsic proximity effect in two-band superconductors, Phys. Rev. Lett.* **105** (Aug, 2010) 067003.

[62] T. Senthil and O. Motrunich, *Microscopic models for fractionalized phases in strongly correlated systems, Phys. Rev. B* **66** (Nov, 2002) 205104.

[63] A. Polyakov, *Quark confinement and topology of gauge theories, Nuclear Physics B* **120** (Mar., 1977) 429–458.

[64] J. Nissinen, K. Liu, R.-J. Slager, K. Wu, and J. Zaanen, *Classification of point-group-symmetric orientational ordering tensors, Phys. Rev. E* **94** (Aug, 2016) 022701.

[65] K. Liu, J. Nissinen, J. de Boer, R.-J. Slager, and J. Zaanen, *Hierarchy of orientational phases and axial anisotropies in the gauge theoretical description of generalized nematic liquid crystals, Phys. Rev. E* **95** (Feb, 2017) 022704.

[66] K. Liu, J. Nissinen, Z. Nussinov, R.-J. Slager, K. Wu, and J. Zaanen, *Classification of nematic order in 2 + 1 dimensions: Dislocation melting and $o(2)/Z_N$ lattice gauge theory, Phys. Rev. B* **91** (Feb, 2015) 075103.

[67] J. Zaanen, Y. Liu, Y.-W. Sun, and K. Schalm, *Holographic Duality in Condensed Matter Physics.* Cambridge University Press, Cambridge, 2015.

[68] S. A. Hartnoll, A. Lucas, and S. Sachdev, *Holographic quantum matter*, Mar., 2018.

[69] C. Holzhey, F. Larsen, and F. Wilczek, *Geometric and renormalized entropy in conformal field theory, Nucl. Phys. B* **424** (1994) 443–467, [hep-th/9403108].

[70] G. Evenbly and G. Vidal, *Quantum Criticality with the Multi-scale Entanglement Renormalization Ansatz*, pp. 99–130. No. arXiv:1109.5334. arXiv, Berlin, Heidelberg, 2013. arXiv:1109.5334.

[71] G. Carleo and M. Troyer, *Solving the quantum many-body problem with artificial neural networks*, *Science* **355** (2017), no. 6325 602–606, [https://www.science.org/doi/pdf/10.1126/science.aag2302].

[72] F. Vicentini et al., *NetKet 3: Machine Learning Toolbox for Many-Body Quantum Systems*, [arXiv:2112.10526].

[73] H.-Q. Shi, X.-Y. Sun, and D.-F. Zeng, *Neural-network quantum state of transverse-field ising model*, *Communications in Theoretical Physics* **71** (nov, 2019) 1379, [arXiv:1905.11066].

[74] E. Fradkin and L. Susskind, *Order and disorder in gauge systems and magnets*, *Phys. Rev. D* **17** (May, 1978) 2637–2658.

[75] E. Fradkin and S. H. Shenker, *Phase diagrams of lattice gauge theories with Higgs fields*, *Physical Review D* **19** (June, 1979) 3682–3697.

[76] J. B. Kogut, *An introduction to lattice gauge theory and spin systems*, *Rev. Mod. Phys.* **51** (Oct, 1979) 659–713.

[77] D. Luo, G. Carleo, B. K. Clark, and J. Stokes, *Gauge equivariant neural networks for quantum lattice gauge theories*, *Physical Review Letters* **127** (Dec., 2021) 276402, [arXiv:2012.05232].

[78] S. Ghosh, R. M. Soni, and S. P. Trivedi, *On The Entanglement Entropy For Gauge Theories*, *JHEP* **09** (2015) 069, [arXiv:1501.02593].

[79] H. Matsueda, *Entanglement Entropy and Entanglement Spectrum for Two-Dimensional Classical Spin Configuration*, *arXiv e-prints* (Sept., 2011) arXiv:1109.0104, [arXiv:1109.0104].

[80] G. Kanwar, M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, S. Racanière, D. J. Rezende, and P. E. Shanahan, *Equivariant flow-based sampling for lattice gauge theory*, *Physical Review Letters* **125** (sep, 2020).

[81] M. Medvidovic, J. Carrasquilla, L. E. Hayward, and B. Kulchytskyy, *Generative models for sampling of lattice field theories*, 2021.

[82] A. Bukva, J. de Gier, K. T. Grosvenor, R. Jefferson, K. Schalm, and E. Schwander, *Criticality versus uniformity in deep neural networks*, 2023.

[83] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[84] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, *Zero-Shot Text-to-Image Generation*, *arXiv e-prints* (Feb., 2021) arXiv:2102.12092, [arXiv:2102.12092].

[85] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, *Language Models are Few-Shot Learners*, *arXiv e-prints* (May, 2020) arXiv:2005.14165, [arXiv:2005.14165].

[86] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, *Highly accurate protein structure prediction with alphafold*, *Nature* **596** (2021), no. 7873 583–589.

[87] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver, *Mastering atari, go, chess and shogi by planning with a learned model*, *Nature* **588** (2020), no. 7839 604–609.

[88] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, *On the Expressive Power of Deep Neural Networks*, *arXiv e-prints* (June, 2016) arXiv:1606.05336, [arXiv:1606.05336].

[89] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Unsupervised learning techniques.* O'Reilly Media, Incorporated, 2019.

[90] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and J. Pennington, *Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks*, .

[91] M. Chen, J. Pennington, and S. S. Schoenholz, *Dynamical isometry and a mean field theory of rnns: Gating enables signal propagation in recurrent neural networks*, .

[92] J. Erdmenger, K. T. Grosvenor, and R. Jefferson, *Towards quantifying information flows: relative entropy in deep neural networks and the renormalization group*, *SciPost Phys.* **12** (2022), no. 1 041, [arXiv:2107.06898].

[93] D. Hernandez and T. B. Brown, *Measuring the Algorithmic Efficiency of Neural Networks*, *arXiv e-prints* (May, 2020) arXiv:2005.04305, [arXiv:2005.04305].

[94] R. Livni, S. Shalev-Shwartz, and O. Shamir, *On the Computational Efficiency of Training Neural Networks*, *arXiv e-prints* (Oct., 2014) arXiv:1410.1141, [arXiv:1410.1141].

[95] S. Hayou, A. Doucet, and J. Rousseau, *On the Selection of Initialization and Activation Function for Deep Neural Networks*, *arXiv e-prints* (May, 2018) arXiv:1805.08266, [arXiv:1805.08266].

[96] S. Hayou, A. Doucet, and J. Rousseau, *On the impact of the activation function on deep neural networks training*, in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, pp. 2672–2680, PMLR, 09–15 Jun, 2019. arXiv:1902.06853.

[97] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory*, [arXiv:2106.10165].

[98] K. T. Grosvenor and R. Jefferson, *The edge of chaos: quantum field theory and deep neural networks*, *SciPost Phys.* **12** (2022), no. 3 081, [arXiv:2109.13247].

[99] R. Jefferson, "Criticality in deep neural nets." https://rojefferson.blog/2020/06/19/criticality-in-deep-neural-nets/, 2020. Accessed: 2022-11-17.

Samenvatting

Vooruitgang op het gebied van computerhardware wordt doorgaans opgevolgd door toepassingen die proberen de nieuwe hardware maximaal te benutten. Deze vooruitgangen vinden ook hun weg naar de wetenschap, waar ze ons helpen om de bestaande grenzen van wat mogelijk is te verleggen.

Aan het begin van dit proefschrift, in hoofdstuk 2, contrasteren we de huidige consensus over thermalisatie in gesloten kwantumsystemen, de "eigentoestand thermalisatie hypothese" (ETH), tegenover een recent ontdekte "operator thermalisatie hypothese" (OTH), doormiddel van het bestuderen van thermalisatiedynamiek in gesloten unitaire kwantumsystemen. We hebben aangetoond hoe de twee hypotheses verschillend zijn, en toch in sommige opzichten vergelijkbaar. No-go-voorwaarden die worden opgelegd door de OTH zijn een kenmerk van de integreerbaarheid van de theorie, en slechts een kleine beweging van deze voorwaarden vandaan zorgt ervoor dat de matrixelementen hun ETH-vorm benaderen. Voor het oplossen van deze grote "eigen-problems" was rekenkracht nodig van het lokale computerraster van de Universiteit Leiden.

Een van de meest alomtegenwoordige numerieke methoden in de wetenschap is de Monte Carlo-simulatie, oorspronkelijk ontwikkeld door Stanisław Ulam toen hij werkte aan kernwapens in het Los Alamos National Laboratory. De hoofdgedachte van Monte Carlo simulatie is het willekeurig nemen van steekproeven van de waarden van een integrand om bij benadering de waarde van een integraal te berekenen. Hiermee heeft Monte Carlo-simulatie heeft een revolutie teweeggebracht in de wetenschap en het hedendaagse computing. In hoofdstuk 3 bestuderen we de zogenaamde "rooster-ijktheorie". Voortbouwend op de fundamenten die Wegner heeft gelegd bij zijn realisatie van de pure Z_2-ijktheorie, en een uitbreiding van het werk van Fradkin en Shenker, construeren we Z_2-ijktheorie gekoppeld aan de verschillende materievelden. Ondanks het feit dat ze niet interageren, leiden deze materievelden tot interessante verschijnselen wanneer ze worden doorgemeten met hetzelfde ijkveld. Er is namelijk een nieuwe "registry" -orde in de Higgs-fase ontstaan, wat betekent dat lokaal verschillende kopieën van materievelden hun vectoren parallel en anti-parallel uitlijnen, zelfs in de aanwezigheid van continue $O(2)$-symmetrie van materie velden.

Door Monte Carlo-simulaties uit te voeren voor grotere 3D-roosters hebben we een aantal interessante kenmerken ontdekt van faseovergangen die voorheen overschaduwd werden door de effecten van eindige grootte.

De afgelopen jaren zijn we getuige geweest van een overweldigende ontwikkeling van machine learning-technieken, geïnspireerd door de nieuwe generaties grafische kaarten. De industrie leidde voornamelijk het onderzoek naar nieuwe toepassingen; toch hebben deze nieuwe en enerverende toepassingen hun weg gevonden naar de wetenschap, vooral de natuurkunde. Een methode, "genaamd" neurale kwantumtoestanden (NQS), betreft het gebruik van een neuraal netwerk om toestanden van zeer verstrengelde kwantum toestanden te representeren. Een bijzondere architectuur van neurale netwerken, genaamd "Restricted Boltzmann-machines" (RBM), is zeer geschikt voor de taak, gedeeltelijk omdat het niet-lokale correlatie dankzij het ontwerp omvat. In hoofdstuk 4 onderzoeken we verstrengelingsentropie en de schaling ervan voor dezelfde ijktheorieën als in hoofdstuk 3, nu uitgedrukt als een 2D-kwantumtheorie in één dimensie lager. We stellen vast dat de verwachte lineaire relatie tussen het totale aantal materievelden en verstrengelingsentropie niet aanwezig is.

De manier waarop de natuurkunde de toepassingen van machine learning heeft omarmd, kan ook andersom worden gebruikt, om enkele van de successen te rechtvaardigen en deze verder te verbeteren. In navolging van deze mantra passen we in hoofdstuk 5 de inzichten uit de statistische fysica toe om de computationele mechanica van diepe neurale netwerken te bestuderen. We onderzoeken de wijze waarop de initiële parameterverdeling voor de gewichten en biases kan leiden naar twee verschillende faseregimes van het netwerk, en hoe het kiezen van het optimale punt binnen dit fasediagram de uiteindelijke nauwkeurigheid van het netwerk bepaalt, onder de voorwaarde van een gelijke trainingstijd. We stellen vast dat het initialiseren van gewichten en biases volgens de lijn van faseovergang een noodzakelijke maar geen voldoende voorwaarde is voor optimale trainbaarheid.

Summary

Advancements in computing hardware are usually followed by emerging applications trying to utilize the new hardware to the maximum. These advances also find their way into science, where they help us push the boundaries of what has been possible so far.

At the beginning of the thesis, in Chapter 2, we contrast the current consensus answer about thermalization in closed quantum systems, the eigenstate thermalization hypothesis (ETH), with a recently discovered operator thermalization hypothesis (OTH) by studying thermalization dynamics in closed unitary quantum systems. We showed how the two are different and yet similar in some regards. No-go conditions imposed by the operator thermalization hypothesis are a feature of the integrability of the theory, and just a slight move away from it would make matrix elements approach their ETH form. Solving these big eigen problems required the firepower of the local computing grid at Leiden University.

One of the most ubiquitous numerical methods in science is Monte Carlo simulation, initially developed by Stanisław Ulam while working on nuclear weapons at Los Alamos National Laboratory. The main idea of Monte Carlo is to randomly sample the values of an integrand to compute the approximate value of an integral, revolutionized science and modern-day computing. In Chapter 3, we turn to the lattice gauge theory. Building on the foundations set by Wegner in his realization of the pure Z_2 gauge theory and expanding the work of Fradkin and Shenker, we construct Z_2 gauge theory coupled to the several matter fields. Even though they are non-interacting, these matter fields lead to exciting phenomena when gauged through the same gauge field. Namely, a new "registry" order in the Higgs phase has emerged, meaning that locally different copies of matter fields align their vectors in a parallel and anti-parallel fashion, even in the case of continuous $O(2)$ symmetry of matter fields. Running Monte Carlo simulations for bigger 3D lattice sizes, we have discovered some exciting characteristics of phase transitions previously obscured by the finite-size effects.

In recent years we have witnessed an overwhelming development of machine-learning techniques inspired by the new generations of graphic cards. The industry predominantly led the research of new applications; nevertheless, these new and exciting applications have found their way into science, especially physics. One

method, named neural quantum states (NQS), considers using a neural network to represent a quantum state of highly entangled systems. A particular architecture of neural networks called Restricted Boltzmann machines (RBM) is very well situated for the task, partially because it includes non-local correlation by design. In Chapter 4, we explore the entanglement entropy and its scaling of the same gauge theories from Chapter 3, now expressed as a 2D quantum theory in one lower dimension. We find that the expected linear relation between the total number of matter fields and entanglement entropy is not present.

In the same way, that physics has embraced applications of machine learning, it also can be used the other way around, to justify some of its successes and further improve upon them. Following this mantra in Chapter 5 we apply the insights from statistical physics to study the computational mechanics of deep neural networks. Precisely how the initial parameter distribution for the weights and biases can lead to two different phase regimes of the network and how choosing the optimal point in this phase diagram can make the final accuracy of the network change given equal training time. We find that initializing weights and biases along the line of phase transition is necessary but not sufficient condition for optimal trainability.

List of Publications

1. " New approaches for boosting to uniformity",

   A. Rogozhnikov, **A. Bukva**, V. Gligorov, A. Ustyuzhanin and M. Williams
   JINST, 11 (2015) T03002.

2. "Operator thermalization vs eigenstate thermalization",

   A. Bukva, Philippe Sabella-Garnier, Koenraad Schalm,
   e-Print: 1911.06292.

3. "Criticality versus uniformity in deep neural networks",

   Aleksandar Bukva, Jurriaan de Gier, Kevin T. Grosvenor, Ro Jefferson, Koenraad Schalm, Eliot Schwander
   e-Print: 2304.04784, submitted to *JMLR*.

4. "Replicating Higgs fields in Ising gauge theory: the registry order",

   Aleksandar Bukva, Koenraad Schalm, Jan Zaanen
   e-Print: 2305.02400, submitted to *Phys.Rev.E*.

Curriculum Vitae

I took my first breath on May 20th, 1994, in Kikinda, a small city in the north of Serbia. After finishing elementary school in Kikinda, I moved to Novi Sad to attend one of Serbia's best and oldest high schools, gymnasium Jovan Jovanovic Zmaj.

High school was the place where I discovered that I wanted to study physics later in life. Besides the regular high school curriculum, I attended the physics seminar at Petnica science center. I was quickly introduced to the world of computer simulations and developed a particle-in-cell (PIC) code for simulating a plasma wave accelerator. After finishing high school, I moved to Belgrade, where I obtained my bachelor's degree. In the first year, my publication "New Approaches for Boosting to Uniformity" was awarded the best technical paper at Belgrade University. During my bachelor's, I continued my interest in computational physics. I did a summer internship at Helmholtz-Zentrum Dresden Rossendorf in a group of Dr. Michael Bussmann, where I worked on an improved version of an integration algorithm for their PIC code.

I continued my master's degree in Belgrade. Still, I did most of my thesis work at the Institute of Physics under the supervision of Milica Milovanović. The title of my master's thesis is: "Lattice-like structures in Lowest Landau Level" where we were trying to construct an effective Hamiltonian for a bosonic system on a square lattice that will support fractional quantum Hall effect (FQHE) at $\nu = \frac{1}{2}$. I started my Ph.D. in November 2018, in the Quantum Matter Theory Group at the Lorentz Institute of Leiden University, under the supervision of Prof. Dr. J. Zaanen and Prof. Dr. K.E. Schalm. During my Ph.D., I taught a couple of courses as a teaching assistant, "Classical Electrodynamics" and "Theory of General Relativity". Alongside the research, I have attended several schools during my Ph.D., DRSTP Schools in High Energy and Condensed Matter Physics, in Brazil and the Netherlands. I have also presented my work at several Physics@Veldhoven conferences.

As of January 1st, 2023, I have started working as a Quantitative Developer for an energy trading company Northpool B.V.

Acknowledgements

This thesis is a culmination of all the work done and events that happened since I moved to the Netherlands. First, I would like to thank Jan Zaanen and Koenraad Schalm for their guidance on this journey and all the knowledge they shared during that time.

Even though there is a single name on this thesis, people in my research group have helped me a lot through numerous discussions, code debugging, and brainstorming; thank you a lot, in no particular order, Vladimir Ohanesjan, Aravindh Swaminathan Shankar, Floris Balm, Nicolas Chagnet, Tereza Vakhtel, Philippe Sabella-Garnier, Kevin Grosvenor, Vincenzo Scopelliti and Aurelio Romero-Bermudez. Also, thank you for making these years in Leiden very enjoyable.

I would also like to express my gratitude to my friends, with whom I have also discussed some of the problems that I've encountered while working on this thesis, thank you Marko Kuzmanović, Daniel Szilagyi, Ognjen Stanisavljević and Dušan Drobnjak.

I also want to thank the support staff, the secretariat, and the IT of the Lorentz Institute and LION for their help. I would also like to thank the surfSARA, Alice cluster, and Xmaris personnel for the access to their compute resources, without which none of the results in this thesis would be possible.

And finally, from the bottom of my heart, I would like to thank my mom, dad, and brother for all the support they gave me. Without you, none of this would be possible. Puno vam hvala!

Stellingen

Behorende bij het proefschrift

Numerical Exploration of Statistical Physics

1. Integrable theories can have operators that thermalize.

   [*Chapter 2*]

2. Multiple matter fields gauged with the same Z_2 field exhibit "registry" order parameter in the Higgs phase.

   [*Chapter 3*]

3. Entanglement entropy of the lattice gauge theories with matter fields at criticality does not grow with the number of additional matter fields.

   [*Chapter 4*]

4. Initialization along the edge of chaos is a necessary but not sufficient condition for optimal trainability.

   [*Chapter 5*]

5. Generative models can be used to improve calculation of entanglement entropy in lattice gauge field models.

   M. Medvidovic, J. Carrasquilla, L. E. Hayward, B. Kulchytskyy, *Generative models for sampling of lattice field theories* arXiv:cond-mat/2012.01442

6. The recent development of machine learning techniques can significantly improve our current optimization algorithms. But caution must be applied and methods understood rather than mindlessly used.

7. Advancements in computational physics would be vastly accelerated by making codes of research papers publicly available.

8. The accuracy of the ground-state energy is not sufficient evidence of the proper representation of a ground state in gapless systems.

9. Considering machine learning beyond being merely a profoundly intricate minimization challenge is fruitless.

Aleksandar Bukva
Leiden, 10$^{\text{th}}$ October 2023