# Statistical Mechanics of Recurrent Neural Networks Based on Physical Models

Qu Yuxuan* and Lucas Seah†
*NTU*
(Dated: January 16, 2024)

The Ising model was identified to be the first recurrent neural network. In this paper, we try to explain what the statement means, hence reviewing how the inverse Ising problem is unsupervised learning, and how statistical mechanics can be used to analyze the behaviour of a restricted Boltzmann machine with binary weights, yielding phase transitions.

## I. INTRODUCTION

The proliferation of Artificial Neural Networks, Artificial Intelligence and Machine Learning have seen a rise in research and development in this area. Its usefulness and applicability span a myriad of sectors including analytical chemistry [1], drug discovery and development [2], particle physics [3], astrophysics [4], and understanding our brain. While its development has come a long way to advanced algorithms and robust infrastructures, it is interesting to study its genesis and its close ties with statistical mechanics.

## II. NEURAL NETWORKS FOR MACHINE LEARNING

Neural networks are inspired by the structure and function neural circuits in the human brain [5] [6]. A typical feedforward neural networks (FNN) can be modeled by weighted, directed graphs (the mathematical object). Nodes (neurons) are organised into three layers - input, hidden, and output. The links (synapses) which have weights that determine the strength of interaction betwen nodes and impact information processing. Data points are fed into input nodes of the network. Neuron activation values of layers are calculated from the preceding layers through the weights, which are represented by a feature matrix and the activation function. The processed information is fed forward through the network, producing an output[7].

Machine learning employs algorithms and statistical models to computer programs on a data set to learn patterns and make predictions about the data set. There are several approach to machine learning — supervised learning, unsupervised learning, reinforcement learning. We're primarily focused on unsupervised learning. It is learning from unlabeled data, finding patterns or underlying structures within the data [8].

A type of neural network we are interested in are recurrent neural networks (RNN). Unlike FNNs, RNNs are bidirectional which can be represented by an undirected graph. This means that output from some nodes are able to affect subsequent input to the same nodes [9][10]. In particular, we are interested in a type of unsupervised learning that is called auto-associative self-supervised feature learning. What it does is that a neural network is trained on its own output data after forgetting its original feature matrix used to generate those data, so as to reconstruct the feature matrix and reproduce the same output data statistics [11]. This is the fundamental task of neural networks known as auto-encoders [12].

## III. ISING MODEL AS AN RNN FOR UNSUPERVISED LEARNING

In this section, we introduce the Ising problem for spin glass models, summarize and discuss some techniques to solve the Ising problem, and explain why the Ising/spin glass models can be considered RNNs. Then, we introduce the inverse Ising problem and explain how it is essentially unsupervised learning.

### A. The Ising problem of Spin Glass Models

The original Ising problem [13] was a problem in many body physics entailing the solution of the magnetization (per spin), $m$, of a lattice of $N$ spins, each admitting values -1 and 1, given that neighbouring spins interact with the same coupling constants, $J$, every spin experiences an external magnetic field $H$ and the system has an inverse temperature $\beta$.

If we view the spins as neurons, magnetization of each spin as neuron activation values, and $J$ as the weights of the synapses between the neurons, and let $H = 0$, then there is a close correspondence between the Ising model and the neural network. Crucially, it is the fact that the Ising model reaches equilibrium given $J$ and $\beta$ which allows the calculation of magnetization [14]. However, if there is only a single $J$ for all the synapses, all neurons end up with the same magnetization, and it would not be a really useful neural network. Therefore, spin glass models are needed to capture the complexity of neural networks.

We introduce the general spin model [15] by comparing its Hamiltonian (Equation 2) with that of the original Ising model (Equation 1) in the absence of an external

* quyu0001@e.ntu.edu.sg
† seah0222@e.ntu.edu.sg

field.

$$\mathcal{H} = -\frac{1}{2}\sum_i \sum_{j\in\partial i} J s_i s_j \tag{1}$$

$$\mathcal{H} = -\sum_a J_a \prod_{i=1}^{N} s_i \tag{2}$$

Note that $\partial i$ is the set of indices of spins that are immediate neighbours of spin $i$. The general spin glass Hamiltonian is highly complex. There can be arbitrarily many interactions between spins up to all spins in the spin glass. If only unique pair-wise interactions are allowed, then the spin glass model fits the description of a general RNN (with no hidden nodes). If we further restrict the model to nearest neighbour interactions only, then the original Ising model is recovered.

### B. Solving the Ising Problem of Spin Glass Models

There are various ways to solve the spin glass model. We look at two methods.

The replica method uses the important property [16] originating from spin glass physics that the free energy of the spin glass due to *a particular realization* of $\{J_{ij}\}$ converges quickly to the quenched average free energy $\langle -(\ln Z)/\beta\rangle_d$ in the thermodynamic limit ($N \to \infty$). Note that $Z$ is the partition function while the $\langle\bullet\rangle_d$, known as the disorder average, is the average over *all possible realizations* of $\{J_{ij}\}$. Unfortunately, the quenched average free energy is notoriously difficult to calculate, hence the replica trick is needed. It can be shown [17] that

$$\lim_{N\to\infty} \frac{\langle \ln Z\rangle_d}{N\beta} = \lim_{n\to 0}\lim_{N\to\infty} \frac{\ln\langle Z^n\rangle_d}{nN\beta} \tag{3}$$

In the literature, $(\ln\langle Z^n\rangle)/\beta$ is known as the annealed average free energy of a composite system consisting of $n$ replicas of the original system, hence the name "replica method". The quenched and annealed averages describe very different physical pictures [18]. The quenched average is an average of free energy over all the realizations of $\{J_{ij}\}$, but the annealed average treats each $J_{ij}$ to be fluctuating just like each spin value. Then, it can be shown [19] that the free energy is a function of many order parameters:

$$F = F\left((r_{\rho\sigma})^{(n,n)}_{(\rho,\sigma)=(1,1)}, (\hat{q}_{\rho\sigma})^{(n,n)}_{(\rho,\sigma)=(1,1)}, (m_\rho)^n_{\rho=1}\right) \tag{4}$$

where $\rho, \sigma$ are indices over the replica systems. The first order approximation is always obtained through the replica symmetric ansatz [20], which assume that the order parameters are the same over all indices such that there are only three order parameters $\hat{q}, r, m$. Such solutions can become unstable or unphysical when the temperature is low. In such a case, there has to be replica symmetry breaking [21] in the free energy expression, yielding higher order approximations. However, the reason why the replica trick works is enigmatic. There is no clear physical intuition behind it, and neither is its mathematical derivation truly rigorous [17].

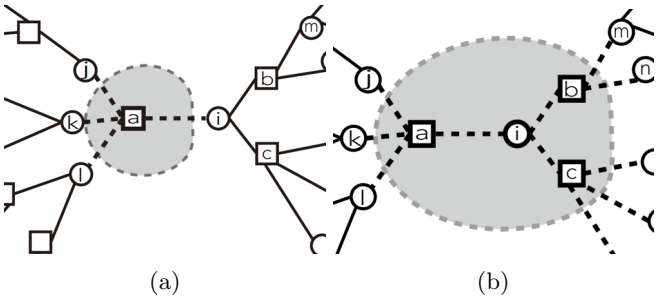



Figure 1: Figures showing cavities in regions of a factor graph. (a) shows a cavity in a small region while (b) shows one in a large region. Figures obtained from [15].

The cavity method [15], equivalent to the variational Bethe approximation from Physics and belief propagation from Computer Science [22], solves the magnetization from a self-consistent equation through iterations of the so-called message passing equations. The cavity method is formalized over factor graph representations of the problem.

Firstly, notice that the partition function of the spin glasses can be written as

$$\begin{aligned} Z &= \sum_{\{s_i\}_{i=1}^N} \exp\left(-\sum_a J_a \prod_{i=1}^N s_i\right) \\ &= \sum_{\{s_i\}_{i=1}^N} \prod_a \exp\left(-J_a \prod_{i=1}^N s_i\right) \end{aligned} \tag{5}$$

A factor graph is a representation for the product of functions. Each function is represented by a square node or the so-called a factor node. All variables are represented by circular nodes or the so-called variable nodes. If a function takes a variable as an input, there is an unweighted link between the corresponding factor node and variable node. Therefore, the factor graph of spin glass is a complete bipartite graph with $N$ variable nodes and arbitrarily many factor nodes.

The cavity method proceeds by replacing small or large regions of a factor graph (Figure 1) by a cavity. Small regions are single factor nodes while large regions are single variable nodes and its neighboring factor nodes. A cavity allows one to talk about the magnetization of each surrounding variable node when the nodes in the cavity region doesn't exist and if the variable nodes are weakly correlated. If we then consider placing the nodes in the cavity region back, we can calculate the shift in free energy due to this change. By piecing up the shifts in free energy for all variable and factor nodes [23] we can find an expression of the total free energy, $F$, of the

system.

$$F = \sum_i \Delta F_i + \sum_a \Delta F_a - \sum_a |\partial a| \Delta F_a \tag{6}$$

$$\Delta F_i = -\frac{1}{\beta} \ln \left( \sum_{\lambda=\pm} \prod_{b \in \partial i} \Lambda^{\lambda}_{b \to i} \right) \tag{7}$$

$$\Lambda^{\pm}_{b \to i} = \cosh(\beta J_b) \left( 1 \pm \tanh(\beta J_b) \prod_{j \in \partial b \backslash i} m_{j \to b} \right) \tag{8}$$

$$\Delta F_a = -\frac{1}{\beta} \ln \left[ \cosh(\beta J_a) \left( 1 + \tanh(\beta J_a) \prod_{i \in \partial a} m_{i \to a} \right) \right] \tag{9}$$

where $m_{i \to a}$ is the magnetization of the $i$-th variable node when the $a$-th factor node is in a cavity, $\partial b \backslash i$ is the set of all variable nodes neighbouring the $b$-th factor node except the $i$-th variable node, $\Delta F_i$ and $\Delta F_a$ are respectively the shifts in free energy by adding the $i$-th variable node and the $a$-th factor node, and the last term in Equation 6 removes over-counting.

It can be shown [24] that $m_{i \to a}$ can be calculated through Equation 10 by iterating over Equations 11 and 12 until a fixed point is reached. At this fixed point, Equations 11 and 12 become a self-consistent equation, and interestingly the fixed point will also correspond to a stationary point of $F$.

$$m_{i \to a} = \tanh \beta h_{i \to a}, \quad \hat{m}_{a \to i} = \tanh \beta u_{a \to i} \tag{10}$$

$$h_{i \to a} = \frac{1}{\beta} \left( \sum_{b \in \partial i \backslash a} \beta u_{b \to i} \right) \tag{11}$$

$$u_{a \to i} = \frac{1}{\beta} \tanh^{-1} \left[ \tanh(\beta J_a) \prod_{j \in \partial a \backslash i} \tanh(\beta h_{j \to a}) \right] \tag{12}$$

where $h_{i \to a}$ is the cavity local field, $u_{a \to i}$ is the cavity bias, and $\hat{m}_{a \to i}$ is the conjugate magnetization. Equations 11 and 12 are known as the message passing equations of the cavity fields. They are called message passing equations because they can be reduced to the message passing equations of belief propagation in Computer Science. Interestingly, they provide a very good physical picture for what is happening in a spin glass model. Equation 12 is saying that the magnetization of the variable nodes neighbouring the factor node $a$ tells how the interaction encoded in the factor node should be. Equation 11 is saying that the factor nodes neighbouring the variable node $i$ creates an effective local field at $i$, telling how the variable node should be magnetized. Hence, the variable nodes effectively communicate with each other recursively until an equilibrium is reached. This is exactly how a recurrent neural network should behave!

Finally, for every fixed point the free energy can be computed. With the fixed point that minimizes the free energy, the magnetization of each variable node can be calculated to solve the Ising problem: $m_i = \tanh\left(\sum_{b \in \partial i} \beta u_{b \to i}\right)$.

### C. The Inverse Ising Problem and Unsupervised Learning

The inverse Ising problem [25] is formulated with a teacher-student scenario. $L$ samples of the teacher spin model's spin values is collected. Then, the student Ising model is tasked to reconstruct the weights of the teacher model from the samples, so as to reproduce spin values with the same statistics as the samples. It is clear that this fits closely the description of auto-associative self-supervised feature learning in Section II.

In general, for continuous weights, the learning is done through gradient descent/ascent: $\delta J_{ij} \propto \langle s_i s_j \rangle_s - \langle s_i s_j \rangle_T$, where $\delta J_{ij}$ is a small nudge at each iteration step of gradient descent/ascent, $\langle \bullet \rangle_s$ is the sample average, and $\langle \bullet \rangle_T$ is the thermal average. To perform gradient descent/ascent, the Ising problem has to be solved at each iteration step. By extending the spin glass model to consider external fields ($H_i$), $\langle s_i s_j \rangle_T$ can be found through the fluctuation-dissipation theorem: $\langle s_i s_j \rangle_T = \left[ \frac{\partial m_j}{\partial H_i} + m_i m_j \right]_{\forall i, H_i = 0}$.

## IV. SIMPLEST MODEL OF UNSUPERVISED LEARNING

Having established the inverse Ising problem, it is a natural question to ask what are the factors affecting the efficacy of learning. Here we consider a simple model to investigate the roles played by the parameters $L$ and $\beta$.

We introduce an RNN architecture known as the restricted Boltzmann machine (RBM). An RBM can be represented by a bipartite graph where there are only two layers in the neural network — visible and hidden nodes. As the simplest model of unsupervised learning, we consider RBMs with only a single hidden node, $N$ visible nodes, and binary weights. This implies that there is a one-to-one correspondence between the visible nodes and the weights. Thus, the Hamiltonian of our simple model can be written as:

$$\mathcal{H}\left((s_i)_{i=1}^N, s_0\right) = -\sum_{i=1}^{N} s_i J_i s_0, \quad s_i, s_0, J_i \in \{1, -1\} \tag{13}$$

where $s_0$ is the spin value of the hidden node. *Independent* samples produced from the teacher RBM would only contain the spin values of the visible nodes and the student RBM is tasked to reconstruct the weights of the teacher RBM at the same temperature. Since this RBM has discrete binary weights, gradient descent/ascent no longer makes sense. Instead, a Bayesian learning framework can be adopted — the set of predicted weights, $\{\hat{J}_i\}$

should maximize the posterior probability:

$$\Pr\left(J_i \middle| \{(s_i)_{i=1}^N\}_{a=1}^L\right) = \frac{1}{Z}\prod_{a=1}^{L}\cosh\left(\frac{\beta}{\sqrt{N}}\sum_{i=1}^{N}J_i s_{i,a}\right) \tag{14}$$

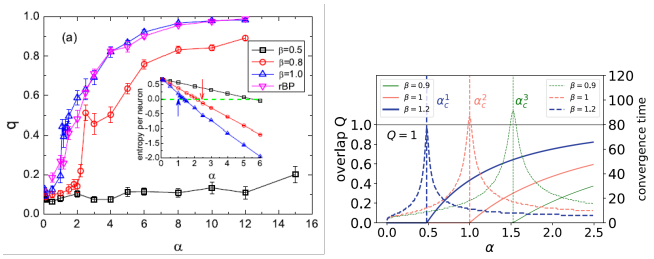
which can be derived by marginalizing the joint conditional probability $\Pr\left(\{(s_i)_{i=0}^N\}_{a=1}^L \middle| J_i\right)$ obtained from Equation 13 through the canonical ensemble and applying Bayes' theorem [26, 27]. Then, notice that Equation 14 can be easily cast into a factor graph representation, with $L$ factor nodes indexed by $a$ and $N$ variable nodes indexed by $i$ (Caution: the weights are now the variable nodes instead!). With appropriate approximations in the thermodynamic limit, the cavity method described by Equations 10 - 12 can be implemented with little modifications [26, 28].

Finally, the weights can be reconstructed through $\hat{J}_i = \arg\max_{J_i}\left(\frac{1+m_i J_i}{2}\right) = \mathrm{sgn}(m_i)$, where $m_i$ is the magnetization of the $i$-th variable node.

### A. Temperature, Data Density, and Phase Transition

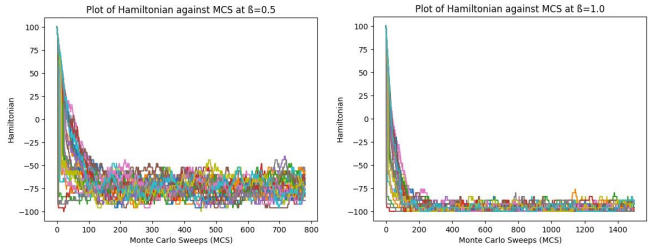
The efficacy of learning can be analyzed through the overlaps $q = \left\langle \left| \overline{J_i^{\text{true}} \hat{J}_i} \right| \right\rangle_{d,s}$ and $Q = \overline{m_i{}^2}$, where $\langle\bullet\rangle_{d,s}$ is a disorder average over a random set of realizations while $\overline{\bullet}$ is the average over all variable nodes. $q$ measures how well the student RBM's predicted weights overlap with the teacher RBM. A $q$ near zero indicates that the student RBM is performing not much better than random guessing. The modulus appearing in the expression for $q$ emphasizes that having a very low accuracy in predicting the weights is in fact favourable. A student RBM with a perfectly wrong feature matrix will yield exactly the same output spin value statistics as a perfectly correct student RBM. This is the result of $Z_2$ symmetry present in the model with no external fields. On the other hand, $Q$ is closely related to the Edwards-Anderson order parameter, $\hat{q}_{\rho\sigma}$, from the replica method.

Figure 2 depicts how $q$ increases with the data density $\alpha = L/N$ after a critical $\alpha_c$ governed by temperature. Better performance with increasing $\alpha$ is expected as learning is always more effective with more data. Interesting questions are then what effects $\beta$ has and why $Q$ exhibits the same behaviour. We only attempt to answer the former here. Figure 3 shows how the Gibbs sampling of the teacher RBM is done with the Metropolis algorithm. At higher temperature, the equilibrium Hamiltonian has large fluctuations, so the sample generated carry more random noise. This offers intuitive explanation of why high temperature negatively affects $q$.




(a) (b)

Figure 2: Phase transition in learning. (a) is a plot of $q$ against $\alpha$ at different $\beta$ for the blue, orange and black lines. Plot from [28]. The inset depicts entropy per neuron which we do not discuss here. (b) shows a plot of $Q$ against the same with different $\beta$ for the solid lines from [27]. Dashed lines show convergence time of message passing which we do not discuss here.




(a) (b)

Figure 3: Hamiltonian values of an ensemble of random walks against Monte Carlo sweeps in Metropolis algorithm used for Gibbs sampling of teacher RBM. (a) is at $\beta = 0.5$ and (b) is at $\beta = 1$.

## V. CONCLUSIONS AND POINTS OF FURTHER INTEREST

In this paper, we have demonstrated that spin glass models are RNNs and the inverse Ising problem is a form of unsupervised learning. We also discussed how $\alpha$ and $\beta$ affects the effectiveness of learning.

However, we did not discuss the connection between $Q$ and $q$, which lies in the physics of replica symmetry breaking [18, 21]. Replica symmetry breaking together with Goldstone's theorem also yield other interesting concepts [29]. With regards to phase transitions, the reader may also be interested in the analysis of neural networks with other tools like renormalization groups [30].

Lastly, we note that the model we have treated here is way too simple compared to modern neural networks used for, as an example, deep learning. Understanding the inner-workings of complex neural networks is ongoing work. Some recent work are available in that regard [31].

[1] B. Debus, H. Parastar, P. Harrington, and D. Kirsanov, Deep learning in analytical chemistry, TrAC Trends in Analytical Chemistry **145**, 116459 (2021).

[2] J. Jiménez-Luna, F. Grisoni, and G. Schneider, Drug discovery with explainable artificial intelligence, Nature Machine Intelligence **2**, 573 (2020).

[3] J. Shlomi, P. Battaglia, and J.-R. Vlimant, Graph neural networks in particle physics, Machine Learning: Science and Technology **2**, 021001 (2020).

[4] D. George and E. Huerta, Deep neural networks to enable real-time multimessenger astrophysics, Physical Review D **97**, 044039 (2018).

[5] D. E. Rumelhart, J. L. McClelland, P. R. Group, *et al.*, Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations, (1986).

[6] C. A. Charu, *Neural networks and deep learning: a textbook* (Spinger, 2018).

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).

[8] K. P. Murphy, *Machine learning: a probabilistic perspective* (MIT press, 2012).

[9] S. Dupond, A thorough review on the current advance of neural network structures, Annual Reviews in Control **14**, 200 (2019).

[10] A. Tealab, Time series forecasting using artificial neural networks methodologies: A systematic review, Future Computing and Informatics Journal **3**, 334 (2018).

[11] M. A. Kramer, Nonlinear principal component analysis using autoassociative neural networks, AIChE journal **37**, 233 (1991).

[12] P. Schneider and F. Xhafa, Chapter 8 - machine learning: Ml for ehealth systems, in *Anomaly Detection and Complex Event Processing over IoT Data Streams*, edited by P. Schneider and F. Xhafa (Academic Press, 2022) pp. 149–191.

[13] J. Yeomans, *Statistical Mechanics of Phase Transitions* (Clarendon Press, 1992).

[14] J. Schmidhuber, Annotated history of modern ai and deep learning, arXiv preprint arXiv:2212.11279 (2022).

[15] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) Chap. 2.

[16] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) p. 66.

[17] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) pp. 66–67.

[18] T. Castellani and A. Cavagna, Spin-glass theory for pedestrians, Journal of Statistical Mechanics: Theory and Experiment **2005**, P05012 (2005).

[19] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) pp. 67–72.

[20] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) p. 73.

[21] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) Chap. 9.

[22] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) Chap. 3.2.2.

[23] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) Chap. 2.2.

[24] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) Chap. 2.3.

[25] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) Chap. 3.3.

[26] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) Chap. 10.

[27] H. Huang, in *Statistical mechanics of neural networks* (Springer, 2021) Chap. 11.

[28] H. Huang and T. Toyoizumi, Unsupervised feature learning from finite data by message passing: discontinuous versus continuous phase transition, Physical Review E **94**, 062310 (2016).

[29] A. A. Fedorenko, Replicon modes and stability of critical behaviour of disordered systems with respect to the continuous replica symmetry breaking, Journal of Physics A: Mathematical and General **36**, 1239 (2003).

[30] S.-H. Li and L. Wang, Neural network renormalization group, Physical review letters **121**, 260601 (2018).

[31] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Statistical mechanics of deep learning, Annual Review of Condensed Matter Physics **11**, 501 (2020), https://doi.org/10.1146/annurev-conmatphys-031119-050745.