Learning efficiency in the Inverse Ising Problem

by

Benjamin Sheldan

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Bachelor of Science

in the
Department of Physics
Faculty of Science
supervised by
Dr. David Sivak

SIMON FRASER UNIVERSITY
Spring 2018

Abstract

In recent years, the amount of data available on biological systems such as genetic regulatory networks and neural networks has increased exponentially, thanks to improvements in experimental methods such as drop-seq [1], which enables biologists to simultaneously analyze RNA expression in thousands of cells. To keep pace with the available data, modern machine learning requires efficient methods for using this data to develop predictive models about the natural world. Using a canonical statistical physics example, the Inverse Ising problem, we ask how physical factors such as temperature affect the learning efficiency. In a network governed by a Hamiltonian with spin-spin interactions, we construct a linear system of equations based on equilibrium observations of spin states, and use linear algebra to solve for the underlying spin-spin couplings. We show that there exists an optimal temperature T_{opt} for which learning is most efficient. Furthermore, we discuss several physical correlates for the scaling of T_{opt} with network size for a simple uniform-coupling network and discuss the extension to more general distributions of couplings. The Fisher information, which depends strongly on the variance of the spin-spin alignment, is shown to predict this scaling most accurately.

Keywords: Ising Model; Machine Learning; Fisher Information, Optimization

Acknowledgements

I would like to thank Dr. Thomson for sharing his code and ideas on creating an inference loop as a linear inverse problem. I would also like to thank Dr. Sivak for his help and ideas throughout the semester, and the rest of the Sivak research group for their help.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction - a simple model for biological systems

In this chapter, we will outline some of the general concepts and goals related to modeling biological networks, using a simplified model of reality. Biological systems are complicated. In classical physics we are used to studying deterministic systems—such as a ball falling under the influence of gravity—where outcomes can be predicted accurately using kinematic equations developed in the classical framework. The picture becomes vastly more complicated when many-body interactions are introduced. For example, genetic regulatory networks, synapses in the brain, and many other biological systems are complex due to the nature of system size and number of complex interactions. The data available to biologists on these systems have increased exponentially in recent years with techniques such as drop-seq which have allowed researchers to profile thousands of individual cells via RNA-seq [1]. However, the large amount of corresponding data produced by these cell profiling techniques requires new computational methods in order to process the information and develop predictive models in real time.

In this work, we develop a method to estimate network parameters as a linear inverse problem. We explain the model used to describe biological systems, its connections to a simple magnetic spin system, well known in statistical physics, and explore how environmental factors such as temperature effect the efficiency of learning.

1.1 Goals in modeling biological networks

Some of the biggest medically related questions in biophysics today involve genetic regulatory networks. For example, in the past few decades research has shown that stem cells have the ability to transform into other cell types [2], making them a desired tool in medical research for the possibility of growing organs. However, further questions can be asked about cells in general. Knowing that a cell identity is determined by protein levels, and genetic regulatory networks (GRNs) are responsible for protein production, we can ask whether it

is possible to 'reprogram' GRNs in order to turn skin cells into liver cells, or whether it possible to turn off cancer cells. All of this requires that connections between the genes and proteins in the body are known so that we can influence them ourselves. In general these networks of cells will be extremely complicated but the main processes can be summarized in two key points. First, cell identity is determined by protein levels. Shifting these protein levels can influence cell identity. Also, production of one protein can suppress production of another or promote production of another. In general therefore, genes will be described by a continuous variable that exists between an 'off' state where no protein production occurs, and an 'on' state with a maximum that will depend on the absolute scale of protein production.

1.2 Our network Hamiltonian and its properties

In order to study these networks, we introduce a simplification from the continuous variable of gene expression to a binary 'on' or 'off' variable, representing high and low expression levels. Furthermore, we represent the influence of one gene on the expression of another with a linear coupling term, negative for repression and positive for activation. The resulting Hamiltonian describing the energy of such a network is [3],

$$\mathcal{H}(\sigma, \mathbf{h}) = -\sum_{i,j} \mathbf{J}_{i,j}\sigma_i\sigma_j - \sum_i \mathbf{h}_i\sigma_i \ , \tag{1.1}$$

where $\mathbf{J}_{ij}$ is a matrix consisting of the linear interaction terms, σ_i are the states of each node, and $\mathbf{h}_i$ is the field at that location in the network. i and j run from 1 to m, the number of nodes in the network. A microstate of the network is specified by the 'spin' state (up [+1] or down [-1]) of each node. Both the field and coupling term are real parameters $\mathbf{J}_{ij}, \mathbf{h}_i \in \mathbb{R}$. The applied external field can be continuous or discrete and of different magnitude at different locations i. In a genetic regulatory network, the influence of spin i on spin j need not be identical to the influence of j on i, but we restrict our attention to equilibrium systems that obey detailed balance [4], hence requiring symmetric effects and hence symmetric coupling coefficients $\mathbf{J}_{i,j} = \mathbf{J}_{j,i}$. Future work could consider the possibility of relaxing the constraint that $\mathbf{J}_{ij} = \mathbf{J}_{ji}$, in which case the ensemble would not be governed by Boltzmann statistics. The magnitude of a self-interaction term J_{ii} merely shifts the energy scale of the system, so without loss of generality we set all self-interaction coefficients to zero. Thus the distribution of states is governed by the Boltzmann distribution [4],

$$P(\sigma_k) = \frac{\exp[-\beta \mathrm{E}(\sigma_\mathrm{k})]}{Z} \ , \tag{1.2}$$

for partition function $Z \equiv \sum_i \exp[-\beta E_i]$ summed over all configurations, and inverse thermal energy $\beta \equiv (k_\mathrm{B}T)^{-1}$ In all numerical simulations we set $k_\mathrm{B} = 1$. Figure 1.1 depicts the

complexities exhibited in real biological systems, modeled with (1.1), with varying magnitudes and numbers of connections between nodes.

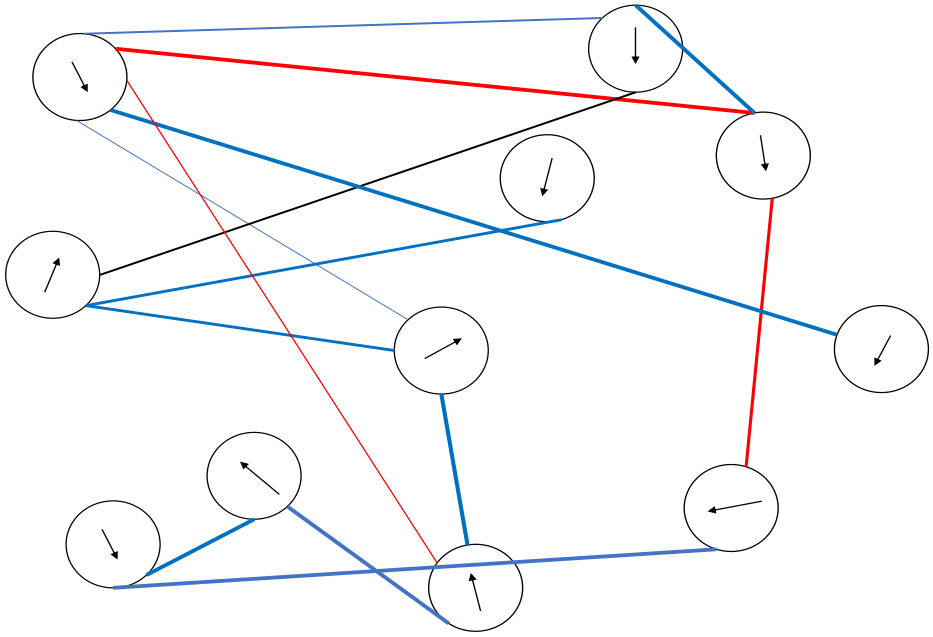

Figure 1.1: **Biological network complexity**. In general, biological networks will vary in numbers of connections, magnitude and sign, representedhere by varying colors of connections. Each generalized 'spin' state is expressed by a continuous variable represented by an angle between 0 and π with respect to the vertical.

1.3 Comparison with magnetic systems

With the introduction of the above simplifications, the network Hamiltonian is now identical to the form used in the Ising model [5]. The Ising model was originally developed to model phase transitions in magnetic material where the microscopic interactions involve the interactions of atomic spins, which are quantized as we know from quantum mechanics. However, the main difference is that interactions terms are usually only considered between nearest neighbors,

$$\mathcal{H}(\sigma, \mathbf{h}) = -\sum_{\langle i,j \rangle} \mathbf{J}_{i,j}\sigma_i\sigma_j - \sum_i \mathbf{h}_i\sigma_i \ , \tag{1.3}$$

and the network structure (chain, lattice, etc.) determines the number of couplings. Ferromagnetic material is modeled with $J_{ij} > 0$, anti-ferromagnetic material with $J_{ij} < 0$, and non-interacting material with $J_{ij} = 0$. In general, biological networks models should allow for all three cases to be present in a mixture.

Using this analogy between genetic regulatory networks (and other types of networks such as neurons, ecosystems, economies, etc.) and magnetism, we use statistical physics to

infer the microscopic interactions in our system from observations of the system behavior. This is known as the Inverse Ising problem, the problem of estimating coupling parameters J_{ij} from equilibrium samples of the system. In statistical physics, for a magnetic system we would predict the total magnetization using the Ising model,

$$M = \left\langle \frac{N_\uparrow - N_\downarrow}{N} \right\rangle , \tag{1.4}$$

where N is the total number of particles and the arrows correspond to spin up or down states and measure this. In a magnetic spin system, the two energy scales are that of the magnetic dipole (to leading order), $(s)\mu \cdot B$, and thermal energy, $k_B T$. Henceforth, we will continue to refer to the two energy scales of the spin network as the internal spin energy and thermal energy. In real biological systems, the thermal energy scale would quantify the size of the general stochastic fluctuations in the system. In genetic regulatory networks, stochastic fluctuations would come from interactions with other proteins and biological material that is not specifically modeled in the genetic network as well as small-number fluctuations in the explicitly modeled degrees of freedom.

Chapter 2

Inference as a Linear Inverse problem

There are currently a number of methods used to solve the Inverse Ising problem and estimate the coupling parameters. One example is the Monte Carlo procedure [6], which takes an initial guess for the system parameters, uses this estimate to calculate the magnetization $\overline{M_i} = \langle \sigma_i \rangle$ and the spin-spin correlations $\langle \sigma_i \sigma_j \rangle - \overline{M_i M_j}$, compares with the true values, and updates the estimates accordingly. The main drawback of this method is simply the long computational time required. Another example is that of pseudo-likelihood [7]. This is done using the same principles of maximizing the log-likelihood function. Mean field theory and small correlation expansion are useful tools as maximizing the likelihood depends on the partition function, and the number of states scale exponentially with network size. In this analysis we use the same principles of maximizing the likelihood function and estimate the energy of the system using multinomial counting and Boltzmann statistics. By estimating the energy of the system and observing the equilibrium configurations of the system spin states we are able to solve for the coupling parameters. In this chapter we set up the linear system of equations, introduce our measure of efficiency, and discuss how to calculate error on the parameters inferred in simulations.

2.1 Linear system of equations

There are several steps to constructing a system of equations. By taking the logarithm of the Boltzmann distribution (1.2) we can construct an estimate for the energy of the system,

$$\langle \widehat{\Delta E_k} \rangle = -T(\ln n_k + \ln Z) \ , \tag{2.1}$$

where the probability is represented by counting states and normalized by the partition function. Setting the original system Hamiltonian (with unknown parameters) equal to the energy estimate obtained with the frequency of states observed $\ln(\mathrm{n_k})$ thus leads to a system

of equations,

$$-T\ln(n_k) = -\sum_{i,j}\mathbf{J}_{i,j}\sigma_i^k\sigma_j^k + T\ \ln Z\ , \tag{2.2}$$

which permits solving for the coupling parameters. Constructing a vectorized version

$$(J_{1,1}, J_{1,2},\ ..., J_{m,m}, -\ln Z) \tag{2.3}$$

of the coupling matrix $\mathbf{J}_{ij}$ turns the problem of inference into a linear inverse problem,

$$-T\ln\mathbf{n} = \mathbf{S}\mathbf{J}_\mathbf{v}\ , \tag{2.4}$$

which can be solved using linear algebra methods.

There are several important notes about the system. First, in this analysis we constrain the external fields to be zero, $\mathbf{h}_i = 0\ \forall\ i$. However, it would be straightforward to add this term into the energy estimate of the system. As will be discussed later, the field has an important effect on the variance of pair-wise alignment which underlines the physical learning in this type of system. Second, the coefficient matrix $\mathbf{S}$ is constructed using the observed pairwise spins of the system at equilibrium, with an additional column of ones for the unknown partition function. In the following section we discuss the specifics of the mathematics of this system and introduce initial arguments for the existence of a region of optimal temperature for learning.

2.1.1 Mathematical form of the system

Based on the earlier constraints on the coupling matrix, we have $m(m-1)/2$ free parameters to estimate with an additional unknown, the partition function to normalize the energy levels. This is schematically shown in the following matrix.

$$\begin{bmatrix} \sigma_1^1\sigma_2^1 & \sigma_1^1\sigma_3^1 & \dots & \sigma_2^1\sigma_3^1 & \dots & \sigma_{m-1}^1\sigma_m^1 & 1 \\ \sigma_1^2\sigma_2^2 & \sigma_1^2\sigma_3^2 & \dots & \sigma_2^2\sigma_3^2 & \dots & \sigma_{m-1}^2\sigma_m^2 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & 1 \\ \sigma_1^k\sigma_2^k & \sigma_1^k\sigma_3^k & \dots & \sigma_2^k\sigma_3^k & \dots & \sigma_{m-1}^k\sigma_m^k & 1 \end{bmatrix} \begin{bmatrix} J_{12} \\ J_{13} \\ \vdots \\ J_{m-1,m} \\ Z \end{bmatrix} = \begin{bmatrix} \widehat{\Delta E_1} \\ \widehat{\Delta E_2} \\ \vdots \\ \widehat{\Delta E_k} \end{bmatrix}$$

Based on the number of unknowns, we can specify three possible scenarios for this system of equations. When $\text{rank}(\mathbf{S}) < \frac{\text{m(m-1)}}{2} + 1$, the system is under-constrained and as a result one cannot estimate all coupling parameters. The second case, $\text{rank}(\mathbf{S}) = \frac{\text{m(m-1)}}{2} + 1$, results in a uniquely determined system of equations. Finally, when $\text{rank}(\mathbf{S}) > \frac{\text{m(m-1)}}{2} + 1$ the system is over-constrained, which can reduce the solution space, or limit inference if the energy estimates are poor. The number of linearly independent rows arises from the number of accessible microstates of the system. Fully connected binary spin systems have

a maximum of 2^m possible configurations. For any system of greater than 2 spins, the exponential growth dominates compared to the quadratic requirement for system inference. Figure 2.1 shows that the required number of samples for inference is exponentially smaller than the total number of configurations possible for the system in a fully connected network.

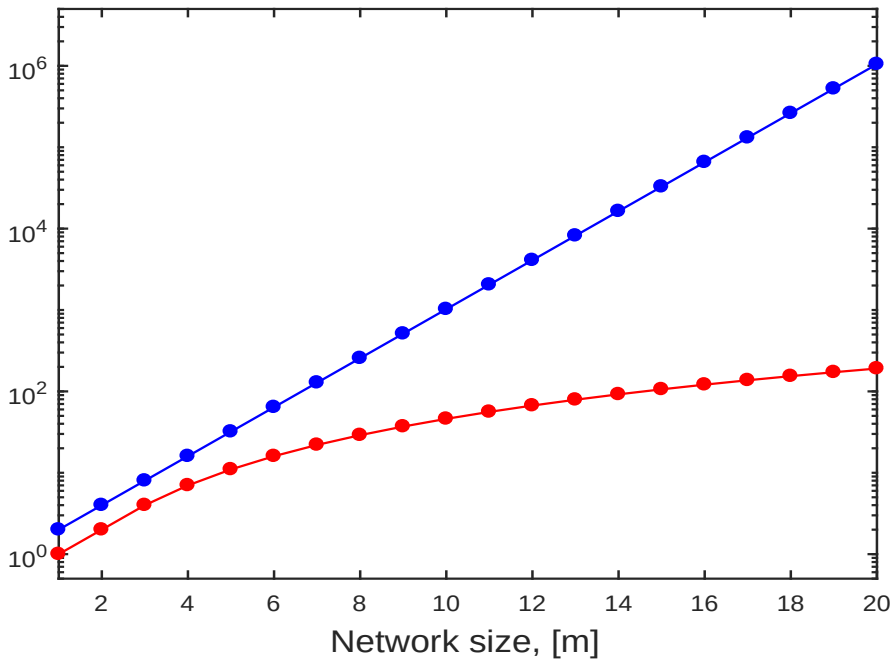

Figure 2.1: **Network configurational space and required states for inference as a function of network size**. Total possible network configurations, 2^m, is plotted in blue and the number of linearly independent states, $m(m-1)/2$, is plotted in red as a function of network size, m.

2.1.2 Asymptotic arguments for inference

Based on the Hamiltonian of the system (1.1), flipping a single spin results in an energy difference of

$$\Delta E = -2 \sum_i \mathbf{J}_{ij} \sigma_i \sigma_j \ , \tag{2.5}$$

and as a result one can estimate coupling parameters only when there is sufficient energy to flip the necessary spins to estimate the resulting energy difference.

Asymptotic arguments point to the existence of an optimal temperature region for inference. At $T = 0$, the only accessible state of the system is the lowest energy level, known as the ground state. As T increases the system remains locked in a small number of low-energy states as governed by the Boltzmann distribution (1.2). In this case, even for a large number of samples of the system at equilibrium, there is not sufficient linearly independent observations of differing states to constrain $\mathbf{J_v}$. Linearly independent observations corre-

spond to observations of the system in different microstates that result in an additional linearly independent row being added to the coefficient matrix of pairwise observed spin states. In this limit, unless there is additional *a priori* knowledge about the network, inference is extremely poor. At the high-temperature limit, there are 2^m states available in a binary spin system, with equal probability $P_i = \frac{1}{2^m}$ for each state. In this limit, there is no problem observing enough states, but the observations are no longer influenced by the couplings. Overwhelming thermal energy causes states to be maximally randomized and in the high-entropy limit, we expect poor inference of energy differences between states and hence of coupling parameters. Using simulations to generate network spin states at equilibrium, preliminary results, shown in Fig. 2.2, agreed with our asymptotic arguments that there is an intermediate temperature that requires substantially fewer observations in order to precisely estimate the coupling parameters.

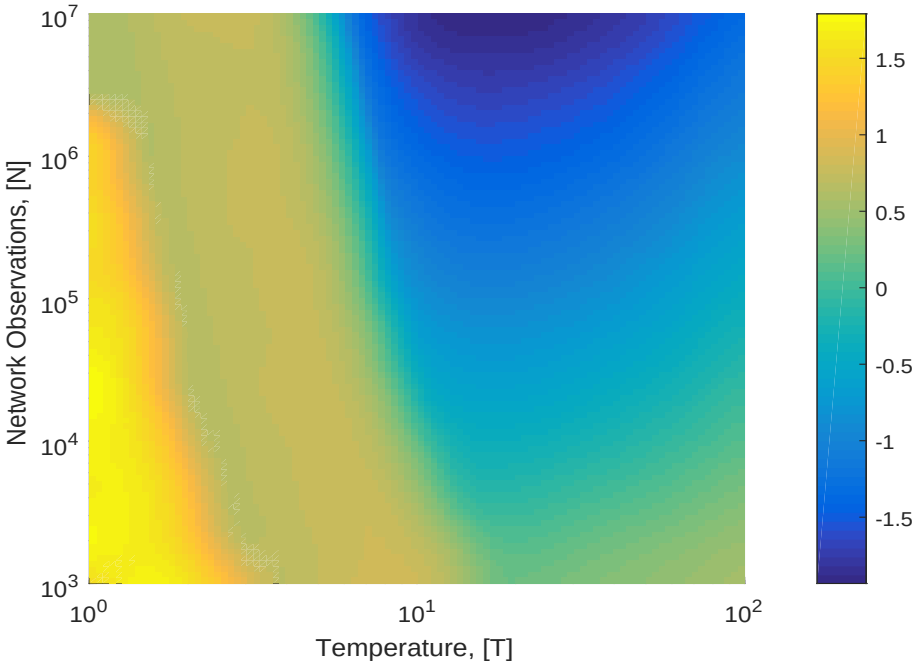

Figure 2.2: **Error on the inferred coupling matrix as a function of temperature and number of equilibrium samples taken for a 10-spin network**. Error is normalized by the number of free parameters $m(m-1)/2$ and displayed on a $\log_{10}$ scale. The number of samples required for inference on the 45 free coupling parameters increases above and below the optimal (lowest-error) temperature region which is depicted in blue.

Here we seek to understand the properties of the optimal temperature region, the physics behind efficient learning at this location, and how the location and size of this region scales with network size. For modeling biological systems we expect that network size will be much greater than 10 nodes, so predicting the optimal temperature scaling with network size will be important for experimental work.

2.2 Inference error

Any increase in network observations should (on average) lead to a decrease in error. From statistical physics, we know that at any finite temperature $T > 0$, there is a non-zero probability of sampling any excited state. These states are suppressed exponentially with their energy above the ground state, but in the many-observation limit ($\lim_{N\to\infty}[Error] \to 0$), as multinomial counting statistics and hence energy estimates improve, the number of required states are eventually observed. We calculate the error in inferring the couplings as,

$$\mathrm{Err} = \frac{\sqrt{\sum_{\mathrm{i,j}}(\mathrm{J}_{\mathrm{ij}}^2 - \mathrm{J}_{\mathrm{ij}}'^2)}}{\mathrm{m(m-1)/2}} \ , \tag{2.6}$$

where $\mathbf{J}_{ij}$ are the set of true couplings used to produce the network spin states in simulations and $\mathbf{J}'_{ij}$ are the set of inferred couplings.

The error depends on the estimates obtained from the system of equations, and in the case of an under-constrained system of equations, initial guesses will have a large impact on the reported error. We estimate the couplings as,

$$J'_{ij} = \begin{cases} J_0 & \mathrm{rank}(\mathbf{S}) = 1 \\ \widehat{J_{ij}} & \mathrm{rank}(\mathbf{S}) > 1 \ , \end{cases} \tag{2.7}$$

where J_0 is some large initial guess that indicates no knowledge about the coupling parameters and will lead to large error, and $\widehat{J_{ij}}$ are the set of estimates from solving the system of equations. However, this method only introduces a large error for the case when only a single microstate is accessed and there are infinite solutions to the system of equations. For an under-constrained system of equations, $1 < \mathrm{rank}(\mathbf{S}) < \mathrm{m(m-1)/2}$, setting free parameters to 0 when lacking sufficient samples to estimate them (a reasonable value given no knowledge about the network *a priori*) leads to error changing non-monotonically with increasing samples, because estimates of coupling parameters with just barely sufficient samples will be quite noisy. This is seen in Fig. 3.1 D.

Future work could implement a more principled Bayesian inference procedure such as,

$$\mathbf{J}''_{ij} = 0 + w_{ij} \cdot \mathbf{J}'_{ij} \ , \tag{2.8}$$

where the couplings are inferred to be zero until there have been enough state observations to implement a reasonable guess based on solving the linear system. The weighting w_{ij} of the parameter inferred from data would be a function of the number of samples, smoothly interpolating between 0 when that coupling is not being estimated and 1 when some threshold number N_0 of state observations have been made, giving high confidence in the output of the linear system solution.

2.3 Definition of inference efficiency

As in many physical scenarios, an important concept is efficiency. How much time, energy, or some other important quantity, is required in order to complete a task? For inferring the model parameters in the Inverse Ising problem, efficiency can be thought of both in terms of computational time and the number of equilibrium samples collected by experimentalists. These are practically equivalent, as the number of equilibrium samples required means more cell profiling and experimental observation time, which also corresponds to more data and pure computational time in solving these systems of equations. We define inference efficiency as the inverse of the number of samples N_0 required for a desired error threshold,

$$\epsilon \propto \frac{1}{N_0(T, \mathbf{h})} \ . \tag{2.9}$$

N_0 will depend on temperature, as seen and argued previously, as well as the external field. Efficiency is naturally important for real networks which may have hundreds, thousands, or greater numbers m of network spins, where computational time becomes of paramount importance.

Chapter 3

Understanding optimal temperature for inference

Initially, we introduced asymptotic arguments 2.1.2, for the existence of an optimal temperature region and used physical arguments involving the number of states accessed based on the thermal energy available to the system. The mathematical justification is that we construct a linear system of equations to solve for the coupling parameters, which is sensitive to fluctuations in the products of spin pairs $\sigma_i\sigma_j$ throughout the network. In this chapter, we will describe elements of the physics behind learning parameters in Ising model networks, in an attempt to physically understand the temperature of optimal inference. We focus on the simpler (and hence easier to analyze) case of uniform coupling constants $J_{ij} = J$ and briefly discuss extensions to a more realistic case with random couping constants drawn from a Gaussian distribution $J_{ij} \sim \mathcal{N}(\mu, \kappa^2)$.

3.1 Generating simulation data and studying linear system results

In this research we worked exclusively with simulated data. An important step, therefore, is the computational process used to generate network spin samples. As described earlier, for simplicity we constrained our network coupling parameters to be uniform. The distribution of states is governed by the Boltzmann distribution (1.2). In Table 3.1 we outline the Metropolis Monte Carlo [8] data generation process for our simulations.

Our earlier asymptotic arguments regarding temperature and increasing number of observed microstates, 2.1.2, suggest that the temperature where $\text{rank}(\mathbf{S}) = \text{m}(\text{m}-1)/2 + 1$ should correlate with the optimal temperature region. Figure 3.1 shows that the temperature where $\mathbf{S}$ becomes full rank is approximately a lower bound on the optimal temperature region. In order to predict the temperature where this condition on the rank of the coefficient matrix occurs, we discuss the phase transition in the configurational space explored by the network and its properties for two types of networks.

Table 3.1: **Algorithm for generating Metropolis Data**

Algorithm Steps
1. Create specific parameters: coupling matrix $\mathbf{J}_{ij}$, field $\mathbf{h}_i$, temperature T, and number N of samples to generate
2. Initialize the system in a random spin state using a random number generator
3. Iterate through each spin in the randomly generated network, calculating the energy and flipping spins until the lowest energy configuration is achieved, and use this as the ground state
4. For $T \neq 0$, use the temperature input and Boltzmann distribution to calculate a probability for a specific spin state to occur
5. For each spin in the network, consider the current energy and the energy if the spin is flipped, compare $e^{-\Delta E/T}$ with a random number in the interval $(0, 1)$ and if the random number is smaller, flip the spin
Note: As $T \to 0$, the Boltzmann probability (1.2) of all excited states goes to zero and spin alignment is unperturbed. As T increases, the probability of excited states increases and the spins are increasingly randomized
6. Repeat steps 2 through 5 for the number of required network observations

To place a lower bound on the optimal temperature region, we know mathematically that we must observe $m(m-1)/2$ linearly independent microstates of the system to obtain a full-rank coefficient matrix. We can predict the temperature where this occurs by looking at the phase transition in configurational space as a function of temperature. Physically, at low temperatures where only a small subset of configurations are explored, one cannot estimate the excited energy levels and therefore cannot determine many of the coupling coefficients. The corresponding linear algebra situation is a rank-deficient matrix and hence under-constrained system of equations. As will be introduced later, 3.2, this transition is described by the critical temperature for the phase transition. At this critical temperature, the system begins to explore all 2^m possible states with non-trivial probability. As a result, the critical temperature will be shown to bound the optimal temperature region.

In summary, the critical temperature describing the phase transition allows one to estimate the temperature at which $\text{rank}(\mathbf{S}) = \text{m}(\text{m}-1)/2 + 1$ and lower bound the region of optimal temperature. In 3.2, we will discuss the mathematical form of the critical temperature and its dependence on network size. Using results from further simulations, we will determine the empirical optimal temperature as a function of network size and compare with the critical temperature. Furthermore, we will test predictions about the scaling of the optimal temperature by studying the entropy and Fisher information of the system.

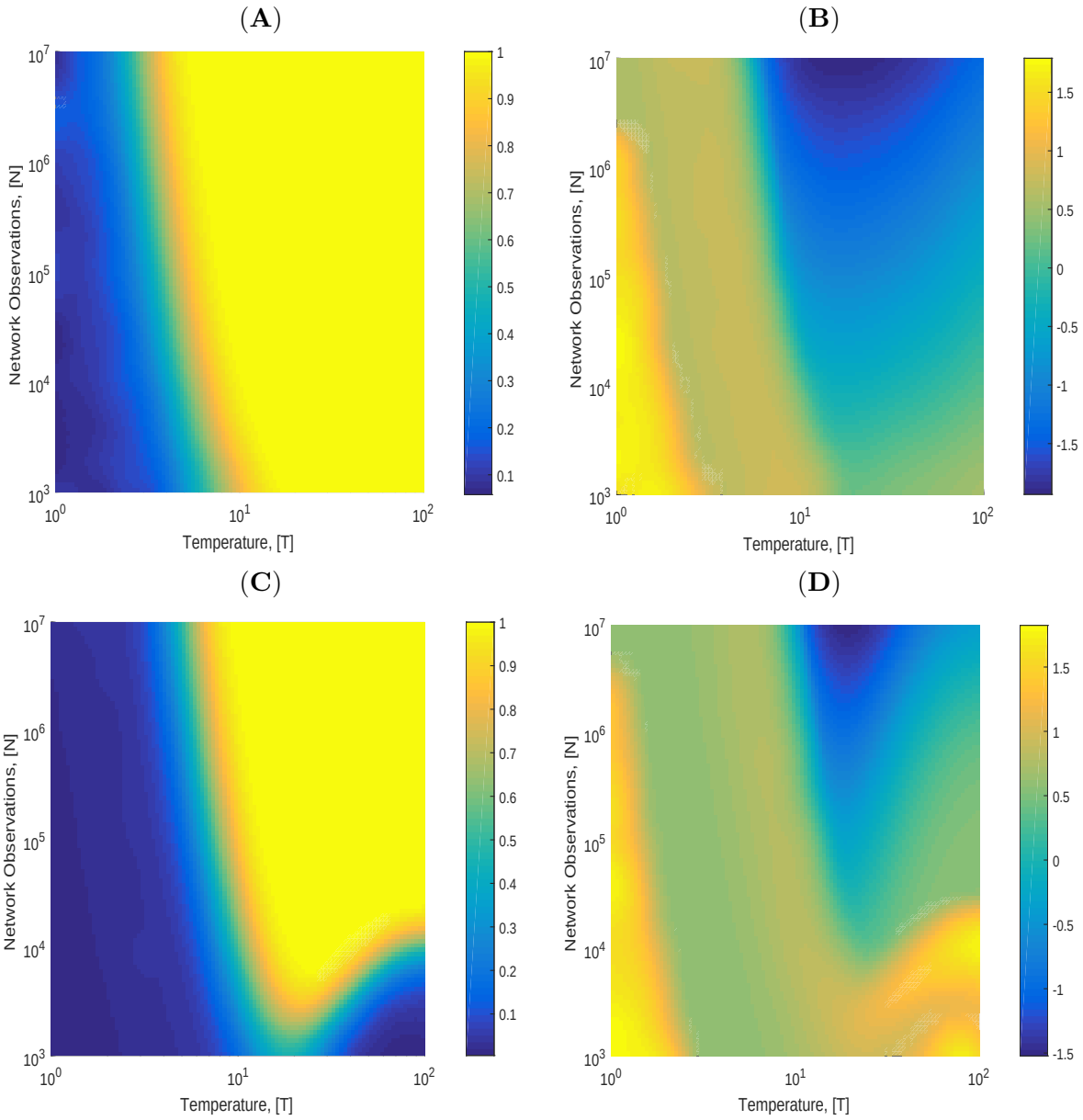

Figure 3.1: **Comparing inference error with rank of the coefficient matrix for networks of 10 and 20 spins.** The transition from empty to full rank occurs where inference error begins to decrease.

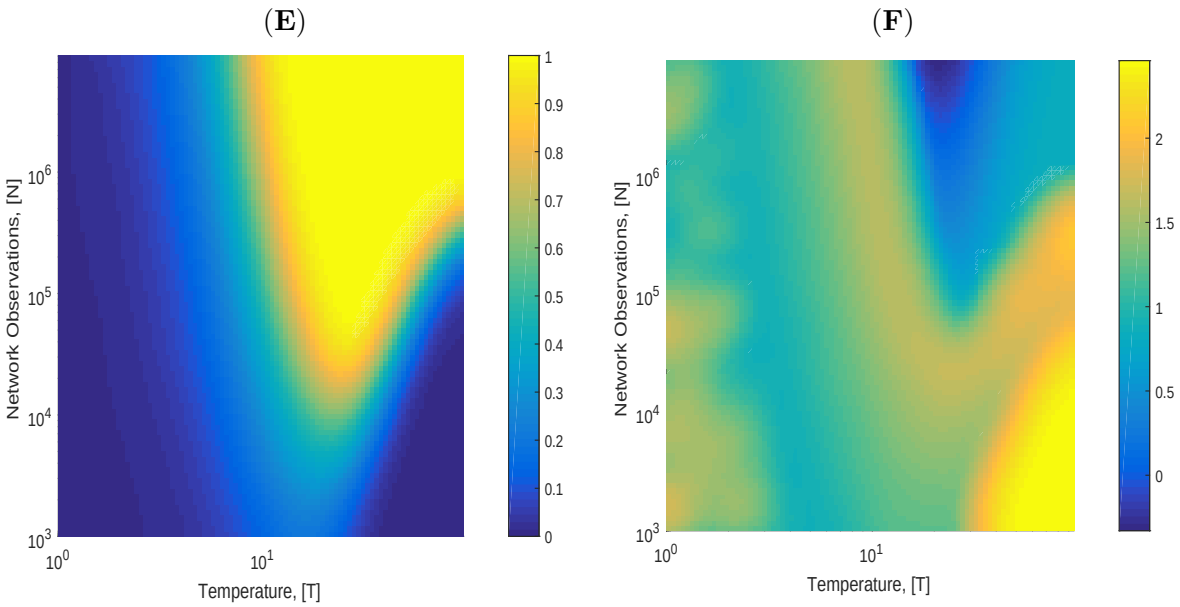

Figure 3.2: **Comparing inference error with rank of the coefficient matrix for a network of 30 spins.** The transition from empty to full rank occurs where inference error begins to decrease.

3.2 Critical temperature for phase transition

The critical temperature defines the transition from the system being locked in a small number of low-energy microstates to exploring a greater range of the 2^m configurational space. In the uniform-coupling network, the largest energy difference is between the ground state and fist excited state, and we can write down an approximate form for the critical temperature considering only these two states.

We first derive the partition function for the uniform-coupling network with $J_{ij} = J$:

$$Z = \sum_{n}^{m} \frac{m!}{n!(m-n)!} \exp\left[-\frac{\mathbf{J}m(m-1)}{T} - \frac{4\mathbf{J}n(m-n)}{T} \right] . \tag{3.1}$$

The energy difference between successive states is,

$$\Delta E_{k,k+1} = 4J(m-1-2k) \ , \tag{3.2}$$

and the largest energy gap occurs for $k = 1$. This energy difference grows as the size of the network increases. This can be visualized by examining the spin state of the network,

$$\uparrow_i \uparrow_{i+1} \uparrow_{i+2} \cdots \uparrow_m \ . \tag{3.3}$$

For a fully connected network, flipping a single spin (m choices) breaks the largest number (m) of connections. Additionally, from the energy gap relation (3.2), successive excited states are closer in energy, requiring smaller shifts in temperature to populate these additional states.

Motivated by this largest energy gap, we derive an expression for the critical temperature considering only the ground state and first excited state. For an effective two-state system the partition function is trivial, giving the probability of sampling the ground state,

$$P_{\mathrm{GS}} = \frac{\Omega_1 e^{-\beta E_1}}{\Omega_1 e^{-\beta E_1} + \Omega_2 e^{-\beta E_2}} \tag{3.4}$$

$$= \frac{1}{1 + \frac{\Omega_2}{\Omega_1} e^{-\beta(E_2 - E_1)}} \tag{3.5}$$

$$= \frac{1}{1 + e^{(1/T - 1/T_c)w_c}} \ . \tag{3.6}$$

Using the partition function, the critical temperature and critical temperature width are,

$$T_c = \frac{4\mathbf{J}(m-1)}{\ln(\mathrm{m})} \quad ; \quad w_c = 4\mathbf{J}m \ , \tag{3.7}$$

where each depend explicitly on the coupling magnitude J and the network size m.

We fit this predicted sigmoidal form of the ground state probability as a function of temperature to numerical simulations of ground state probability as a function of temperature, for several network sizes. Figure 3.4 shows that the simple model works reasonably well for describing the sigmoidal nature of the phase transition, though it misses some features at the edges of the transition. The critical temperature and transition width fit using (3.7) are,

$$T_c = \frac{(4.8 \pm 0.3)\mathbf{J}[(0.380 \pm 0.003)m - 1]}{\ln m} \ , \ w_c = (3.12 \pm 0.06)\mathbf{J}m \ , \tag{3.8}$$

Despite the difference in actual coefficients, the functional form works quite well.

With a functional form describing the critical temperature and both mathematical reasoning and physical intuition for the existence of the optimal temperature region and its lower bound, we examine how the optimal temperature scales with network size. This is of interest because biological systems such as a genetic regulatory network will typically have many more nodes than the networks simulated in Fig. 3.1 and 3.2, though small size reduces computational time and thus facilitates debugging and analysis. Figure 3.3 shows that the optimal temperature for inference increases with network size by comparing error analysis on systems of 20 and 30 spins. We next turn to predicting the scaling of the optimal temperature with network size.

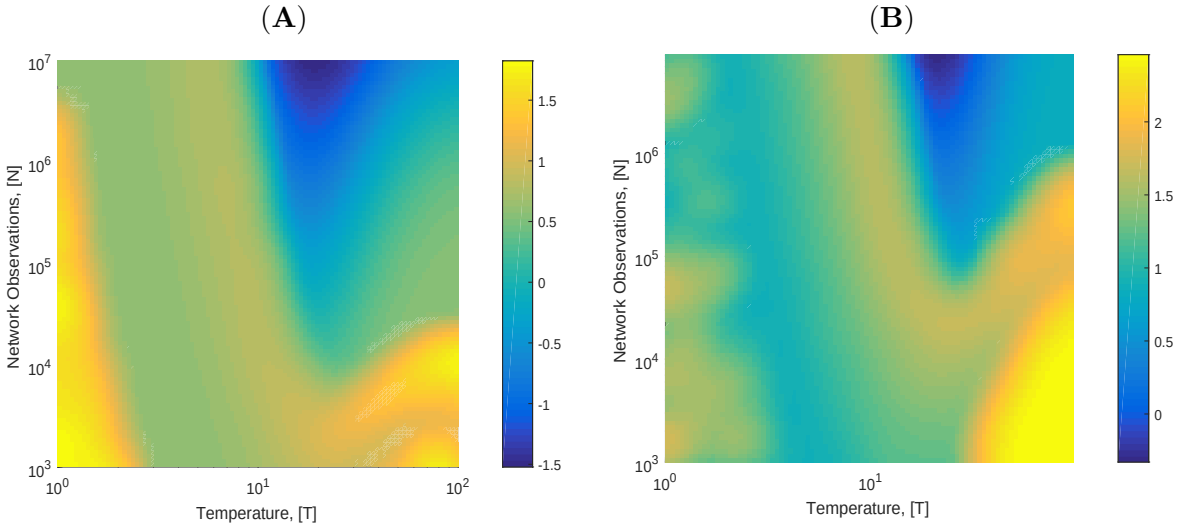

Figure 3.3: **Inference error is minimized at intermediate temperature.** Error plotted as a function of N and T to showcase the optimal temperature region shifting for networks of 20 (A) and 30 (B) spins. Non-monotonic scaling of error with increasing N is discussed in Section 2.2.

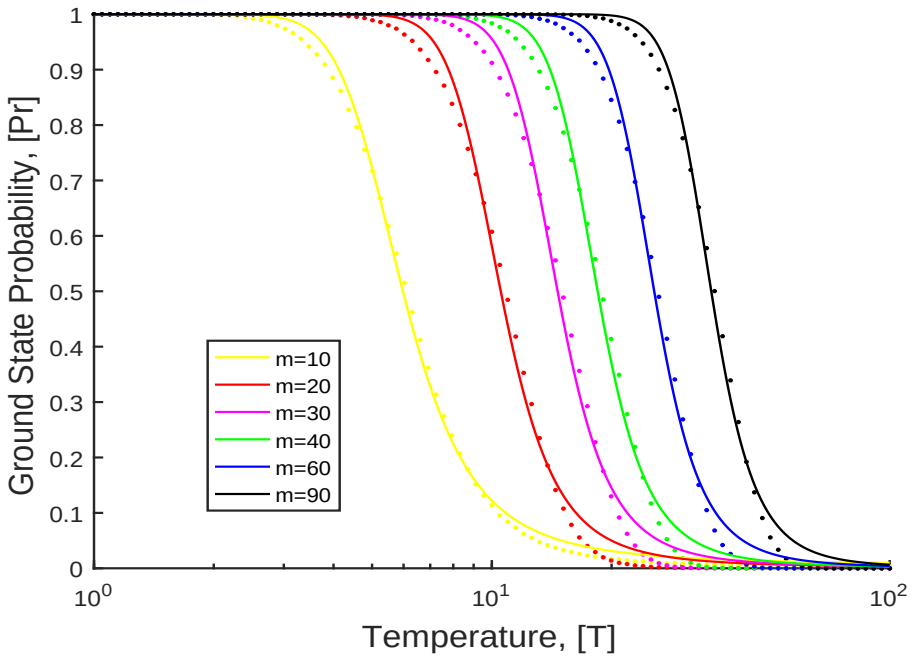

Figure 3.4: **The probability of sampling the ground state as a function of temperature for networks of $m = 10, 20, 30, 40, 60$ and 90 nodes**. Dots indicate empirical data from simulation and smooth curves represent best fits of (3.4) to extract the critical temperature and transition width.

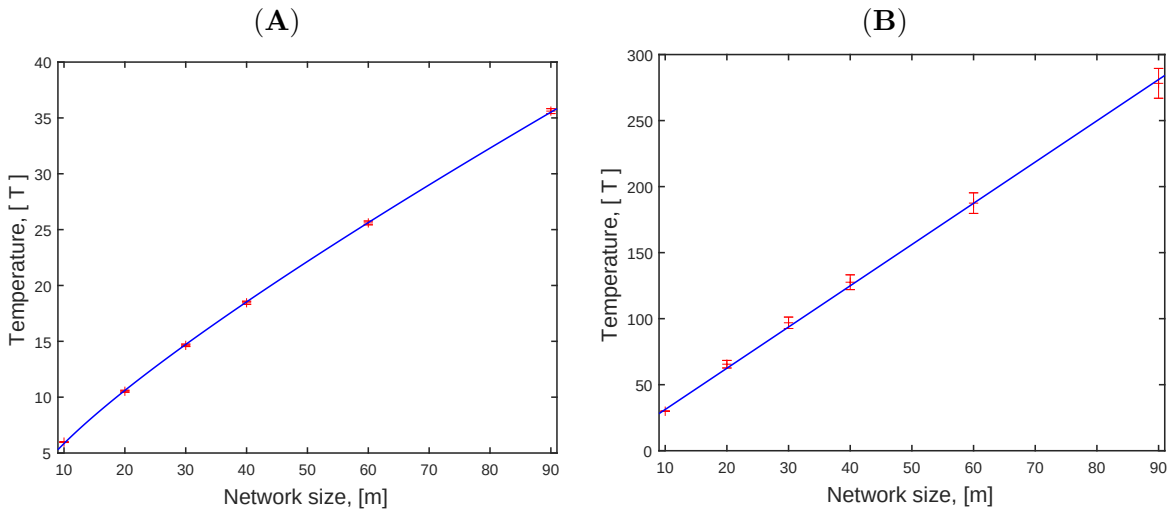

Figure 3.5: **Observed critical temperature and transition width extracted from sigmoidal fits of** (3.4) **for uniform-coupling network**. The form predicted in (3.7) fits well the scaling of critical temperature (A) and transition width (B) with network size m.

3.3 Optimal temperature as a function of network size – Simulation results

We present the initial results for empirical observations of optimal temperature as a function of network size for both network cases, uniform and randomly distributed couplings, in Figs. 3.6, 3.7 and 3.8.

In order to build a predictive model for optimal temperature, the next method of choice was to look at the entropy. Entropy has the general form [9],

$$S = k_{\mathrm{B}} \ln \Omega \ , \tag{3.9}$$

where Ω is the number of possible configurations of the system. By relating the temperature at which the minimum number of states were observed we wanted to test whether this tracked optimal temperature. Furthermore, by investigating the Fisher information which has a closed form for this simple system, we found that it depended on the variance of pairwise alignment and provided another method for tracking the optimal temperature.

3.4 Methods for understanding optimal temperature

3.4.1 Entropy

We began by inspecting the entropy of the network as a function of temperature. This was another motivation for using uniform coupling constants since the partition function can be written down in closed form. Using the usual statistical mechanics identity relating entropy

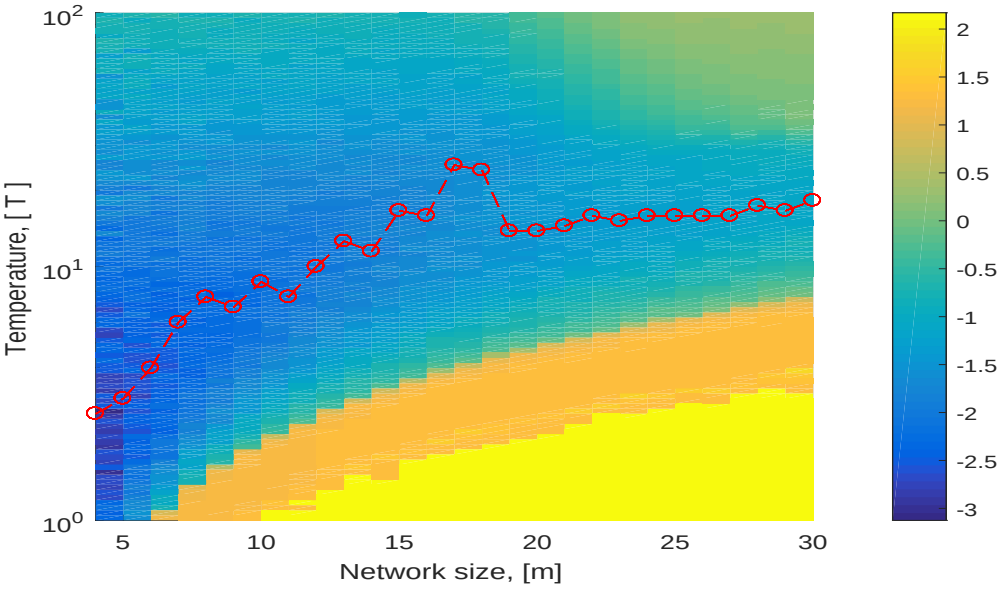

Figure 3.6: **Empirical optimal temperature as a function of network size for uniform coupling constants**. Optimal temperature is highlighted in red. Each error is calculated from 10^7 network observations.

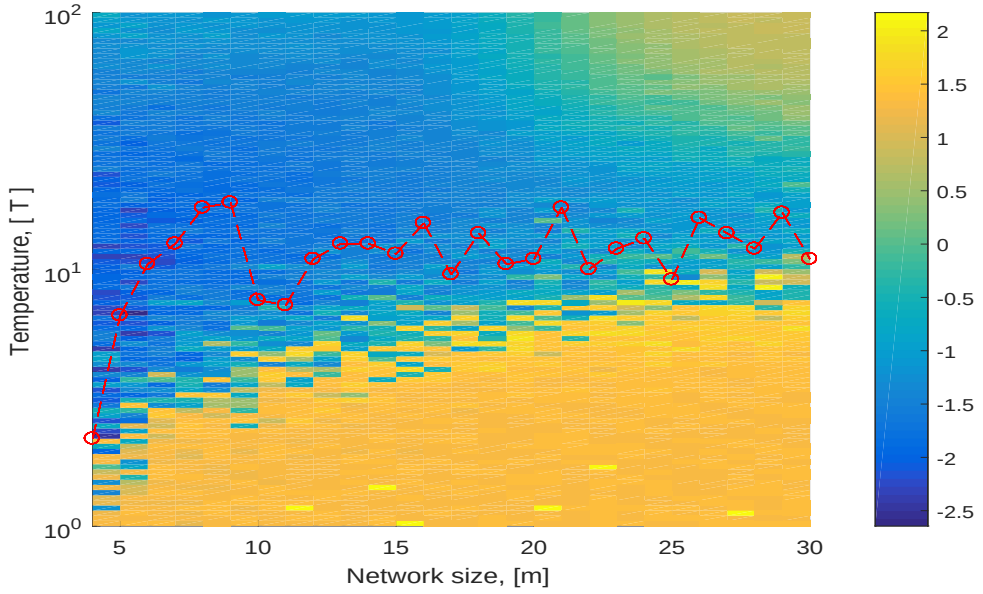

Figure 3.7: **Empirical optimal temperature as a function of network size for Gaussian-distributed coupling constants**. Optimal temperature is highlighted in red. Each error is calculated from 10^7 network observations.

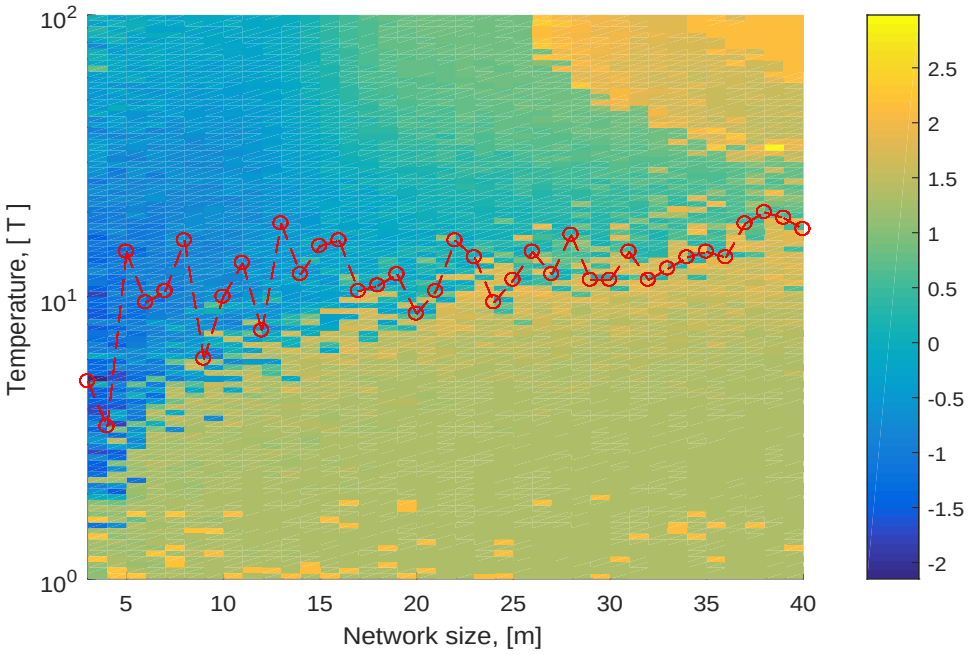

Figure 3.8: **Empirical optimal temperature as a function of increased network size for Gaussian-distributed coupling constants**. Optimal temperature is highlighted in red. Each error is calculated from 10^5 network observations.

to the partition function [10],

$$S = \ln Z + T\frac{\partial \ln Z}{\partial T} \ , \tag{3.10}$$

where k_{B} has been set to 1. This in general takes the form $S = \ln(\Sigma_i \exp(-\beta \mathrm{E_i})) - \Sigma_{\mathrm{i}} \mathrm{E_i}/\mathrm{T}$, which is sigmoidal with temperature, similar to the form that was observed in the ground state probabilities in the previous chapter. This is shown in Fig. 3.9.

First we examine when the entropy equals the logarithm of $m(m-1)/2$, the minimum number of states required to uniquely determine (fully constrain) the solution to the system of equations,

$$S|_{T_s} = \ln\left[\frac{m(m-1)}{2}\right] \ . \tag{3.11}$$

However, at finite T there is not uniform sampling across the observed states, so in general (3.11) is satisfied when less than $m(m-1)/2$ states are observed. Moreover, the $m(m-1)/2$ lowest-energy states will in general not be all linearly independent, preventing full solution of the linear system of equations. These considerations suggest that the entropy should be higher than (3.11) when the linear system of equations first becomes fully constrained, limiting the utility of (3.11).

For Gaussian-distributed coupling constants, we do not know of a simple functional form for the entropy's dependence on temperature, thus the above reasoning (already fraught for an analytic entropy) would have to proceed using the numerically calculated empirical

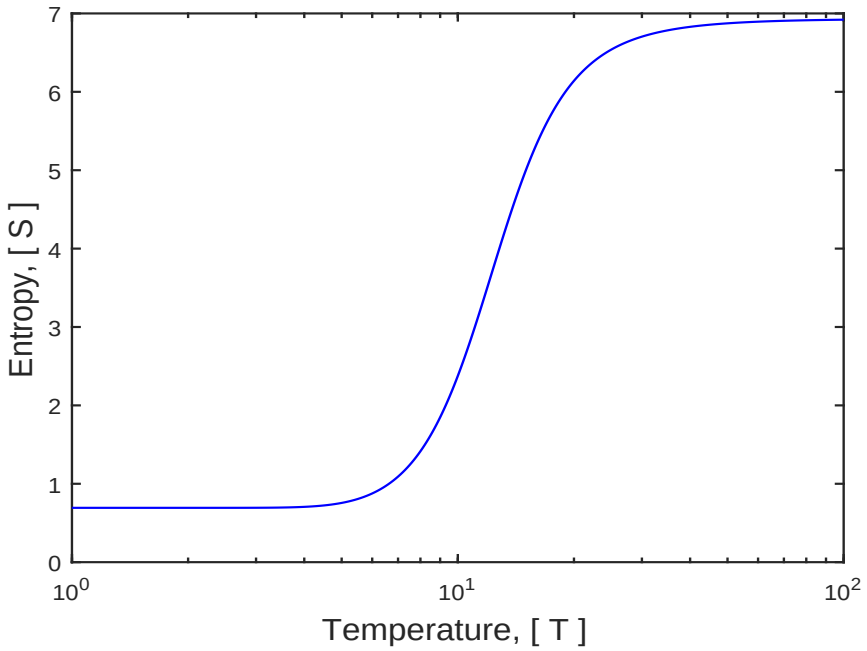

Figure 3.9: **Entropy as a function of temperature for 10-spin network**. Minimum and maximum entropy are $\ln 2$ and $m\ \ln 2$ for, respectively, equal probability of ground states only or equal probability of all states.

entropy,

$$\hat{H} = -\Sigma_i \hat{P}_i \log(\hat{\mathrm{P}}_\mathrm{i}) \ , \tag{3.12}$$

for empirical state probabilities $\hat{P}$.

3.4.2 Fisher information

Fisher information provides an alternate route to estimate the optimal temperature. Fisher information specifically deals with the information carried by a random observable, the empirical state distribution, about an unknown parameter, the network couplings. The definition of Fisher information is [11],

$$\mathcal{I}(\mathbf{J}_{ij}, T) = \left\langle \frac{\partial^2 \ln \mathcal{L}(\mathbf{J}_{\mathrm{ij}}, \mathrm{s_k})}{\partial \mathbf{J}_{ij}^2} \right\rangle \ , \tag{3.13}$$

which depends on the second derivative of the log-likelihood function with respect to the unknown parameter. For our binary spin system, the Fisher information takes the simple closed form,

$$\mathcal{I}(\mathbf{J}_{ij}, T) = \frac{\langle (\sigma_i \sigma_j)^2 \rangle - \langle \sigma_i \sigma_j \rangle^2}{T^2} \ , \tag{3.14}$$

which depends directly on the variance of the spin-spin alignment $\sigma_i\sigma_j$ and is inversely proportional to temperature. Since we are considering a binary system, where the product $\sigma_i\sigma_j$ of two spin states can only be ± 1, the first term in the numerator is always one. However, it will be instructive to leave the Fisher information in this form to compare with other physical quantities which take on similar form. Again, we use asymptotic arguments to explain the form of the Fisher information. At low temperature, the small thermal energy scale does not introduce many fluctuations and the variance of the pairwise alignment goes to 0 as $T \to 0$. At high temperature, the $1/T^2$ dependence dominates, sending the Fisher information to 0. Therefore, there should be a maximal Fisher information at some intermediate temperature. Figure 3.10 shows Fisher information as a function of temperature for several network sizes, showing a maximum as predicted.

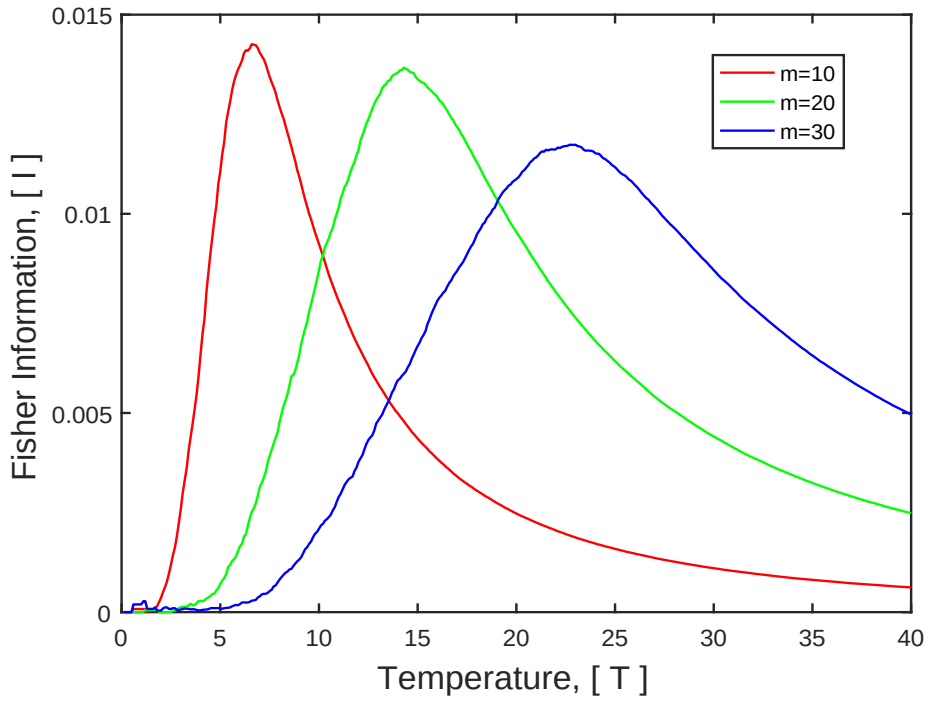

Figure 3.10: **Fisher information for networks of 10, 20 and 30 spins as a function of temperature**. For each network case, Fisher information asymptotes to zero as temperature goes to 0 and ∞.

3.5 Optimal temperature comparisons

To compare with the optimal temperature as estimated using entropy, the phase transition, and Fisher information, we ran simulations to estimate error as a function of temperature for a large number ($N = 10^7$) of network observations, and used the location of the minimum error as the observed optimal temperature. Figure 3.11 shows empirical observations for optimal temperature as a function of network size, compared to the critical temperature T_c,

the temperature T_s producing a certain entropy, and the temperature T_f that maximizes Fisher Information.

The optimal temperature is higher than the critical temperature of the phase transition. This characteristic may be specific to this uniform-coupling network, as the phase transition describes the transition from a single ground state to excited states. In networks with randomly distributed coupling constants, the optimal temperature could occur below this phase transition, provided that there are sufficient low-energy states to uniquely determine the couplings. Our simple entropy calculation overestimates the optimal temperature. This is possibly due to the fact that a single energy level can have multiple spin states which produce linearly independent rows for constraints in the system of equations, and the degeneracy in this network changes the requirements for inference. Empirically, we find that Fisher information is the most accurate at predicting the temperature of most efficient learning, and another magnetic system analogy gives some idea of why.

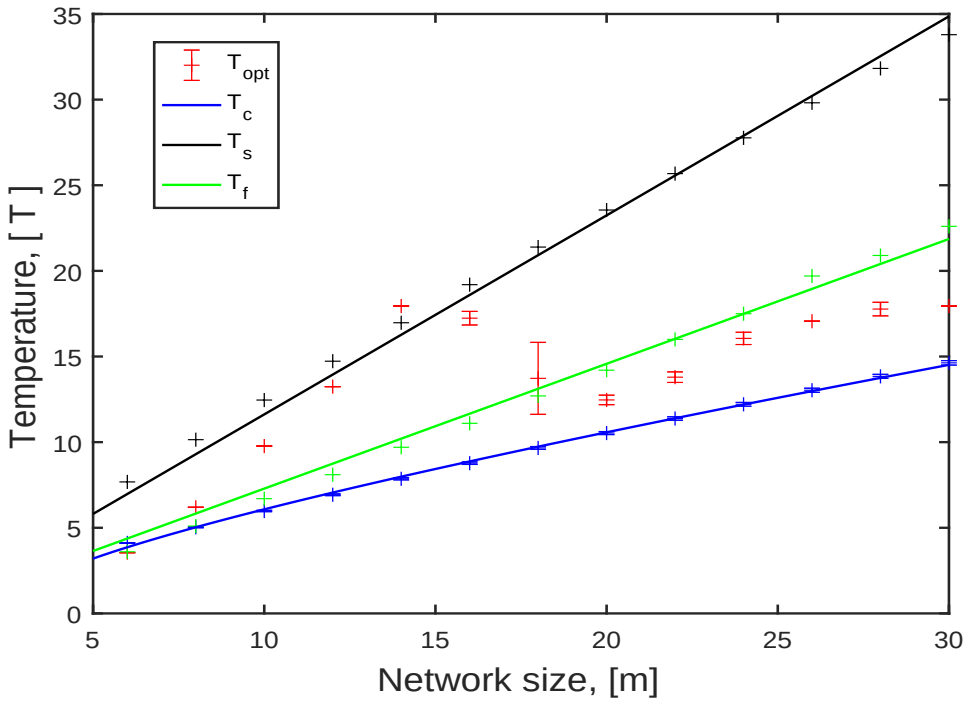

Figure 3.11: **Empirical optimal temperature compared with predictions, using critical temperature for phase transition, intermediate entropy, and maximal Fisher information**. Optimal temperature was calculated by averaging over five data sets. Error bars show the standard error of the mean. Critical temperature was extracted from fits to (3.4), and intermediate entropy and maximal Fisher information were solved numerically using (3.11) and (3.14), respectively.

3.6 Magnetic susceptibility and connection to learning

In systems described by the canonical ensemble in statistical mechanics, a useful thermodynamic quantity is the free energy [12] $F = -k_{\mathrm{B}}T \ln Z$, as useful quantities like the magnetization in magnetic materials can be related to derivatives of the free energy [10], $\overline{M} = \frac{\partial F}{\partial B}$, where B is the applied field. It is well known that the resulting magnetization in certain materials depends on the previous magnetic field history, a phenomena called hysteresis. This hysteresis effect can be calculated by taking the second derivative of the free energy to derive the magnetic susceptibility, a response function [13],

$$\chi = \frac{\partial M}{\partial H} \tag{3.15a}$$

$$\approx \mu_0 \frac{\partial^2 F}{\partial B^2} \tag{3.15b}$$

$$= \mu_0 k_B T \Big[\frac{1}{Z} \frac{\partial^2 Z}{\partial B^2} - \Big(\frac{1}{Z} \frac{\partial Z}{\partial B} \Big)^2 \Big] \tag{3.15c}$$

$$= \frac{\mu_0}{k_B T} (\overline{M^2} - \overline{M}^2) \tag{3.15d}$$

$$\propto \mathrm{var(M)} \ . \tag{3.15e}$$

The susceptibility depends on the variance of the magnetization, and thus the response of the material to applied fields can depend on previous magnetizations. This was derived by taking derivatives with respect to the applied field. For the Fisher information, the only term that survives the second derivative of the log-likelihood function is the partition function,

$$\mathcal{I} = -\Big\langle \frac{\partial^2 \mathrm{lnZ}}{\partial J_{ij}^2} \Big\rangle = -\Big\langle \frac{1}{Z} \frac{\partial^2 Z}{\partial J_{ij}^2} - \Big(\frac{1}{Z} \frac{\partial Z}{\partial J_{ij}} \Big)^2 \Big\rangle \ , \tag{3.16}$$

which is the analogous to the magnetic susceptibility, up to an overall multiplicative factor. In this case, the Fisher information can be thought of as a response function which also depends on a variance,

$$\mathcal{I} \propto \langle (\sigma_i \sigma_j)^2 \rangle - \langle \sigma_i \sigma_j \rangle^2 \propto \mathrm{var}(\sigma_{\mathrm{i}} \sigma_{\mathrm{j}}) \ , \tag{3.17}$$

indicating that the response, the learning process, depends on the equilibrium correlations of the pairwise alignment in the system. This is directly related to our initial assumptions regarding temperature and rank arguments for the coefficient matrix. If there is no variance in the pairwise spin-spin alignment, Fisher information tends to zero as $\langle (\sigma_i \sigma_j)^2 \rangle = \langle \sigma_i \sigma_j \rangle^2 = 1$ and the same network spin state is observed repeatedly, leading to an under-constrained system of equations, and hence no information about the Hamiltonian parameters. In the high-temperature limit there are sufficient linearly independent spin states observed to make the coefficient matrix full rank, however, the pairwise spin statistics are not measurably in-

fluenced by the coupling parameters as thermal energy randomizes the spins. To conclude, the physics of learning information about this simple Ising model system depends on the equilibrium correlations in pairwise spin alignment, and this directly leads to sensitivity in learning the system coupling parameters.

Chapter 4

Concluding remarks and future work

Learning the properties of biological systems such as genetic regulatory networks promises long-term therapeutic benefits such as reprogramming cell fate. Since cell identity is determined by protein levels, and genes are responsible for protein production and suppression, we developed a simple binary spin-state model to represent genes being 'on' and 'off'. This leads to a Hamiltonian representation of the network, the Ising Model, familiar from statistical physics and previously modeled phase transitions in magnetic materials. In order to reprogram this network, we needed to develop a method to solve for the unknown coupling parameters in the system. In this way, we could perturb a system to suppress or produce desired proteins. We developed an approach that constructed a linear system of equations based on equilibrium network spin observations.

However, despite the wealth of biological data, many environmental factors make this learning an exponentially costly process both in experimental research and computational time. Here we examined how to tune these environmental factors, specifically temperature, in order to efficiently learn about the network. We used asymptotic arguments to predict poor learning efficiency at low temperature, due to few states of the system being observed, and at high temperature, being unable to resolve the different energy levels at these high thermal energy scales. Using simulation results (Fig. 2.2), we verified the existence of the optimal temperature region and developed physical understanding for its dependence on network size, in order to predict optimal temperature for larger networks for experimental work. We tested several methods, to predict optimal temperature scaling, including the critical temperature for phase transition, entropy, and Fisher information. We found that Fisher information predicted this scaling most accurately, and this can be understood in the context of another magnetic statistical mechanics process. Magnetic hysteresis, the dependence of magnetization of a material on past history of magnetic fields, can be described by a response function, the magnetic susceptibility, that is proportional to the second derivative of the Helmholtz free energy. In the same way, Fisher information is the second

derivative of the log-likelihood with respect to coupling parameters, and depends on the variance of the pairwise spin-spin alignment. Maximally efficient learning occurs when the Fisher information—which depends on the equilibrium correlations of pairwise alignment—is maximal, in agreement with our original asymptotic arguments in 2.1.2. These insights point the way to tackling more complicated networks.

In a real biological system, we expect that spin couplings will vary in sign and magnitude. Now that we have a more thorough understanding of the emergence of the optimal temperature region, and a method for predicting its scaling with network size, we intend to continue the analysis for more complicated networks, with Gaussian-distributed coupling coefficients. Furthermore, many biological networks (such as synaptic networks in the brain) are sparse [14]. Preliminary data shows that the optimal temperature for sparse networks scales in the same way with density as it does with network size in fully connected networks. One can imagine combining these results to describe networks where there are fully connected sub-networks which in turn weakly couple to other sub-networks. Furthermore, our physical insight that equilibrium correlations in pairwise alignment influence learning points the way to using field perturbations to improve learning. If it possible to control the environment in which learning occurs, we can compare whether it is optimal to introduce field effects or if shifting temperature is more efficient.

Bibliography

[1] E. Macosko *et al.*, *Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets.* Cell, 161:1202-1214, 2015.

[2] S. Sell, *Stem Cells Handbook.* Humana Press; 2nd edition, 2013.

[3] H. Nguyen, R. Zecchina, and J. Berg, *Inverse statistical problems: from the inverse Ising problem to data science.* arXiv:1702.01522, 2017.

[4] W. Kauzmann, *Kinetic Theory of Gases.* Dover Publications, 2012.

[5] H. Stanley, *Introduction to Phase Transitions and Critical Phenomena.* Clarendon Press Oxford, 1987.

[6] F. Ricci-Tersenghi, J. Raymond, and A. Decelle, *A brief introduction to the inverse Ising problem and some algorithms to solve it.* Lecture Slides, Physics Department, Sapienza University, 2004.

[7] J. Berg, *Statistical mechanics of the inverse Ising problem and the optimal objective function.* arXiv:1611.04281, 2016.

[8] D. Frenkel and B. Smit, *Understanding Molecular Simulation.* Academic Press, 2002.

[9] D. Schroeder, *An Introduction to Thermal Physics.* Pearson, 1999.

[10] F. Reif, *Fundamentals of Statistical and Thermal Physics.* Waveland Press, 2009.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Hoboken: Wiley-Interscience, 2nd ed., 2006.

[12] S. Blundell and K. Blundell, *Concepts in Thermal Physics.* Oxford University Press, 2009.

[13] O. Narayan and A. Young, *Free energies in the presence of electric and magnetic fields.* arXiv:cond-mat/0408259, 2005.

[14] J. Snider, *Indistinguishable Synapses Lead to Sparse Networks.* MIT Press Journals, 1989.