UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Sten Sootla

Analysing information distribution in complex systems

Bachelor's Thesis (9 ECTS)

Supervisor: Raul Vicente Zafra, PhD
Supervisor: Dirk Oliver Theis, PhD

Tartu 2017

Contents

Introduction

In Chapter 1, the basics of information theory and partial information decomposition are covered. The chapter ends with an overview of the numerical estimator for PID. The subsequent 3 chapters each introduce a specific complex system and the results of measuring information distribution in them, while they are naturally evolving. In the final, concluding chapter, a summary of the contributions of this thesis is given, alongside suggestions for further work.

1 Background

1.1 Classical information theory

In order to understand partial information decomposition, which is the mathematical framework that is used in this thesis to analyse complex systems, a solid understanding of basic information theory is essential. This section fills that gap, giving a brief overview of the fundamental concepts of information theory. Where appropriate, the rather abstract definitions are further elaborated on by providing the reader with intuitive explanations, concrete examples and practical applications.

In the following discussion, when not specified otherwise, it is assumed that X is a discrete random variable with possible realizations from the set $\{x_1, x_2, ..., x_n\}$ and a probability mass function $p_X(x_i) = Pr\{X = x_i\}$ $(i = 1, ..., n)$. Similiarly, Y is a discrete random variable with possible realizations from the set $\{y_1, y_2, ..., y_m\}$ and a probability mass function $p_Y(y_j) = Pr\{Y = y_j\}$ $(j = 1, ..., m)$. Furthermore, let the joint distribution of the random variables X and Y be $p(x_i, y_j) = Pr\{X = x_i, Y = y_j\}$ $(i = 1, ..., n;\ j = 1, ..., m)$.

1.1.1 Entropy

The most fundamental quantitiy of information theory is *entropy*, being a basic building block of all the other information theoretic functionals introduced in this thesis. The entropy of the random variable X is defined by Shannon [Sha48] as follows:

$$H(X) = -\sum_{i=1}^{n} p_X(x_i) \log p_X(x_i) \tag{1}$$

If the base of the logarithm is 2, the units the entropy is measured in are called *bits*. Another common base for the logarithm is Euler's number $e \approx 2.718$, in which case the units of measurment are called *nats*. As in this definition, the base of the logarihm is also omitted in subsequent discussion for both generality and consistency with [CT06].

Intuitively, entropy can be thought of as the average amount of uncertainty of a random variable. It is indeed an *average*, as the uncertainty of a single relatization x_i of a random variable X can be quantified by $-\log p_X(x_i)$. Viewed from this

angle, the definition of entropy can be rewritten as an expectation of the random variable $-\log p(X)$:

$$H(X) = \mathbb{E}\left[-\log p(X)\right] = \mathbb{E}\left[\log \frac{1}{p(X)}\right].$$

To see why this intuition should correspond to the mathematical definition, it is instructive to look at a concrete example, inspired by [CT06]. Suppose we have a binary random variable X, defined as follows:

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0, & \text{with probability } 1-p. \end{cases}$$

Essentially, this random variable encodes a coin toss, where the probability of flipping heads is p and the probability of flipping tails is $1-p$. If $p = 0.5$, the coin is considered to be unbiased, otherwise it is called biased.

Using equation 1, it is straightforward to calculate the entropy of X, given some specific value of p. Figure 1 graphs the value of $H(X)$ against every possible $p \in [0, 1]$. If $p \in \{0, 1\}$, the outcome of the coin toss is completely deterministic, meaning there is no uncertainty in the result whatsoever. Accordingly, the entropy is 0 for these values of p. Conversely, when the coin is fair, we are completely uncertain about the outcome, unable to favour neither heads or tails. Again, the mathematical definition agrees with the intuition, as the entropy is indeed at its maximum when $p = 0.5$.

Due to the fundamentality of the measure, the usage of entropy is ubiquitous throughout science and engineering. For example, in finance, it is extensively used in portfolio selection theory to measure the diversity and risk of the portfolio [ZCT13]. In civil engineering, it is is a key ingredient in structural optimization design [PLY06] - a subfield of optimization that is conserned with improving the design of structures with respect to various specifications (safety, cost, weight etc.). A rather interesting example of application of entropy comes from cognitive neuroscience, where it has been used to characterize different states of consciousness in the brain [CHLH$^+$14].

1.1.2 Joint and Conditional Entropy

The *joint entropy* [CT06] of the pair (X, Y) is defined as

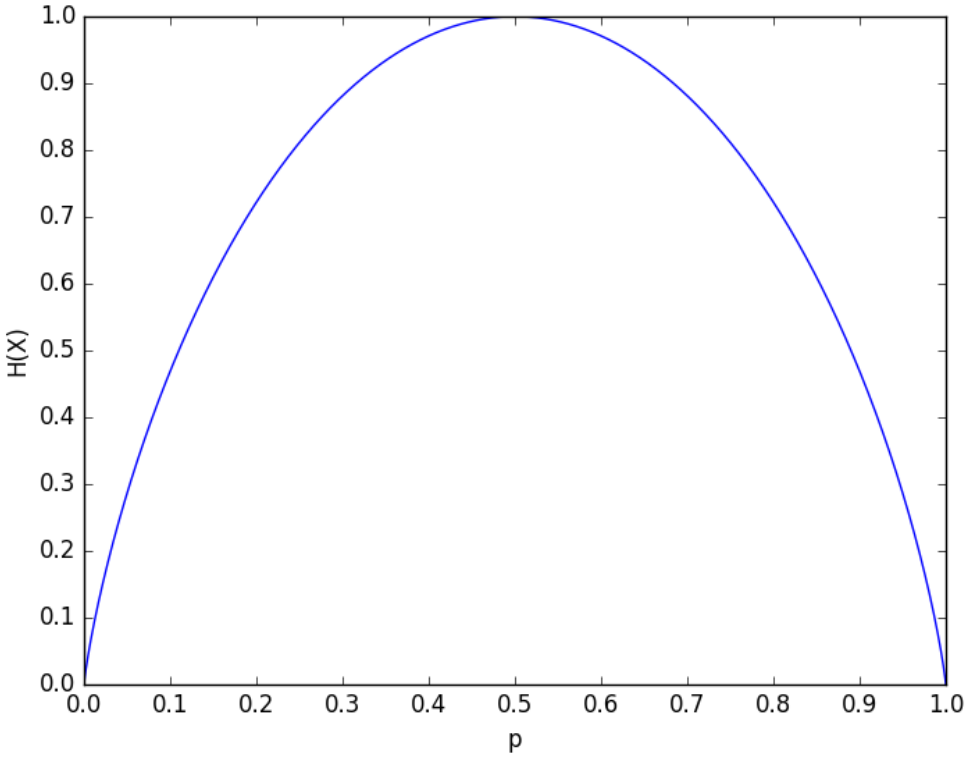

Figure 1: Entropy of X plotted against the value of p.

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log p(x_i, y_j) \tag{2}$$

This is a direct generalization of entropy to multiple variables. Joint entropy for more than 2 random variables can be defined analogously.

The *conditional entropy* [CT06] of the pair (X, Y) is defined as

$$H(Y|X) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log p(y_j|x_i) \tag{3}$$

Conditional entropy can be thought of as the amount of uncertainty one has about a random variable Y, given that X has already been observed. As a special case, if X and Y are independent, observing X does not reveal anything about Y, and $H(Y) = H(Y|X)$.

The entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other [CT06]:

$$\begin{aligned} H(X,Y) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log p(x_i,y_j) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log p_X(x_i)p(y_j|x_i) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log p_X(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log p(y_j|x_i) \\ &= -\sum_{i=1}^{n} p_X(x_i)\log p_X(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log p(y_j|x_i) \\ &= H(X) + H(Y|X) \end{aligned} \tag{4}$$

1.1.3 Kullback-Leibler distance

Let $p_X(x)$ and $q_X(x)$ be two probability mass functions over the support of the random variable X. The *relative entropy* or *Kullback-Leibler distance* [CT06] between $p_X(x)$ and $q_X(x)$ is defined as

$$D(p||q) = \sum_{i=1}^{n} p(x_i)\log\frac{p_X(x_i)}{q_X(x_i)} \tag{5}$$

The above quantity is called a distance, because it can be thought of as measuring how far two probability distributions are from each other. Importantly, the relative entropy is non-negative, with inequality exactly when the 2 distributions are equal [CT06], again corresponding to our intuitive notion of distance. Indeed, when the two distributions are the same, the logarihm in equation 5 evaluates to 0, which in turn yields a relative entropy of 0.

However, it must be stressed that since the Kullback-Leibler distance it is not symmetric and does not satisfy the triangle inequality, it is not a formal distance in the mathematically rigorous sense.

1.1.4 Mutual information

The *mutual information* [CT06] between the random variables X and Y is given by

$$MI(X;Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p_X(x_i) p_Y(y_j)} \tag{6}$$

An attentive reader might notice that the mutual information is the Kullback-Leibler distance between the joint distribution $p(x, y)$ and the product distribution $p_X(x)p_Y(y)$.

Because the mutual information is just a special case of Kullback-Leibler distance, all the properties that hold for relative entropy must also hold for mutual information. In particular, mutual information must be non-negative and 0 exactly when the random variables X and Y are independent. The latter statement must hold, because if X and Y are independent, then $p(x, y) = p_X(x)p_Y(y)$ by definition.

Considering mutual information as a special case of Kullback-Leibler distance, it can be intuitively seen as measuring how far the two random variables X and Y are from being independent. Indeed, when the two are completely independent, one would expect that they contain no information about each other, and this is exactly the conclusion that was reached mathematically directly from equation 6.

The picture of mutual information as a distance between two probability distributions yields a straightforward answer to the question: "when is there no information between two random variables?" However, it does not help in answering the orthogonal question: "when is the information maximized?" To answer the latter, the following identity from [CT06], which relates mutual information directly to entropy, is of importance:

$$
\begin{aligned}
MI(X;Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p_X(x_i)p_Y(y_j)} \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log p_X(x_i) + -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log p(x_i|y_j) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_X(x_i) - \left(-\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log p(x_i|y_j) \right) \\
&= H(X) - H(X|Y)
\end{aligned}
\tag{7}
$$

Intuitively, using identity 7, mutual information between random variables X and Y can be thought of as the reduction in the uncertainty of X due to the knowledge of Y [CT06]. Thus, it is maximized when knowing Y completely determines X, yielding $H(X|Y) = 0$.

Because mutual information is symmetric, the amount of information X has about Y is always equal to the amount of information Y has about X.

1.1.5 Conditional mutual information

Let Z be a discrete random variable. The *conditional mutual information* [CT06] of the random variables X and Y given Z is defined by

$$
MI(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{8}
$$

Intutively, the conditional mutual information measures the reduction in the uncertainty of X due to the knowledge of Y, given that Z has already been observed.

Another useful property that will become important in the discussion on partial information decomposition is the *chain rule for information* [CT06], which allows to express the mutual information between a random vector and a random variable in terms of mutual informations between univariate random variables:

$$
\begin{aligned}
MI(X;Y,Z) &= H(Y,Z) - H(Y,Z|X) \\
&= H(Y) + H(Z|Y) - H(Y|X) - H(Z|Y,X) \\
&= H(Y) - H(Y|X) + H(Z|Y) - H(Z|Y,X) \\
&= MI(X;Y) + MI(X;Z|Y)
\end{aligned}
\tag{9}
$$

When the information between two random variables is measured in a system with many other dependent variables, conditional mutual information is used to eliminate the influence of the other variables, in order to isolate the two variables of interest [WB10]. For example, it has been used to analyse the functional connectivity of different brain regions in schizoprenic patients [SAG+10].

1.1.6 Transfer Entropy

TODO!?

1.2 Partial information decomposition

Mutual information measures the amount of information two random variables, or more generally, two random vectors have about each other. However, it is often worthwhile to ask how much information does an ensemble of input (source) random variables carry about some output (target) variable.

A trivial solution would be to measure the mutual information between the whole input ensemble considered as a single random vector and the output. However, this would not capture the interactions between the input variables themselves. Moreover, by considering the input ensemble as a single unit, knowledge about how the interactions between specific individual units and the output differ is lost.

This section introduces *partial information decomposition (PID)* - a mathematical framework for decomposing mutual information between a group of input variables and single source variable.

1.2.1 Formulation

The simplest non-trivial system to analyse that has an ensemble of inputs and a single output is a system with *two* inputs. Given this setup, one can ask how much information does one input variable have about the output that the other does not, how much information do they share about the output, and how much information do they jointly have about the output such that both inputs must be present for this information to exist.

More formally, let Y and Z be two random variables that are considered as sources to a third random variable X. By equation 7, the mutual information between the pair (Y, Z) and X is defined as

$$MI(X : Y, Z) = H(X) - H(X|Y, Z).$$

The partial information decompositon framework decomposes this mutual information into *unique*, *redundant* and *complementary information* terms.

Unique information quantifies the amount of information that only one of the input variables has about the output variable. The unique information that Y has about output X is denoted as $UI(X : Y \setminus Z)$. Similarly, $UI(X : Z \setminus Y)$ denotes the unqiue information that Z has about the target X.

As an example, consider Table 1, inspired by [GK12], which depicts the joint distribution of the random vector (X, Y, Z). From the table, it can be seen that the output variable X has 4 equiprobable states, each of which is uniquely specified by the two inputs Y and Z. There is unique information in both Y and Z, because they contain different information about the output X that is not present in the other input. Indeed, input Y is able to differentiate between the sets $\{0, 1\}$ and $\{2, 3\}$, while Z discriminates between $\{0, 2\}$ and $\{1, 3\}$.

Y	Z	X	Pr
0	1	0	1/4
0	3	1	1/4
2	1	2	1/4
2	3	3	1/4

Table 1: Example of unique information.

Shared information quantifies the amount of information both inputs share about the output variable. It is also sometimes called *redundant* information, because if both inputs contain the same information about the output, it would suffice to observe only one of the input variables. The shared information is denoted as $SI(X : Y, Z)$.

Table 2, again inspired by [GK12] gives a toy example of shared information. The output variable X has 2 equiprobable states, each of which is again uniquely specified by the two inputs Y and Z. However, in this example, it would actually suffice to observe only one of the inputs Y or Z to uniquely determine the ouput. In other words, one of the input variables is redundant, since the two inputs share all their information about the output.

Y	Z	X	Pr
0	0	0	1/2
1	1	1	1/2

Table 2: Example of shared sinformation

Complementary or *synergetic* information quantifies the amount of information that is only present when both inputs are considered jointly. The complementary information is denoted as $CI(X : Y, Z)$.

Table 4 depicts the `XOR`-gate - the canonical example for illustrating the concept of synergy [GK12]. As before, the output X is fully specified by the two inputs Y and Z. However, in this case *both* inputs Y and Z must be present for the

Y	Z	X	Pr
0	0	0	1/4
0	1	1	1/4
1	0	1	1/4
1	1	0	1/4

Table 3: Synergy

output to be fully determined. Indeed, given a specific value of either Y or Z, there remain two equiprobable values for X.

It is generally agreed ([WB10], [BRO+13], [HSP12], [GK12]) that mutual information can be docomposed into the four terms just described as follows [WPK+15]:

$$MI(X : Y, Z) = SI(X : Y; Z) + UI(X : Y \setminus Z) + UI(X : Z \setminus Y) + CI(X : Y; Z) \tag{10}$$

The same sources also agree on the decomposition of information that a single variable, either Y or Z, has about the output X:

$$\begin{aligned} MI(X : Y) &= UI(X : Y \setminus Z) + SI(X : Y, Z) \\ MI(X : Z) &= UI(X : Z \setminus Y) + SI(X : Y, Z) \end{aligned} \tag{11}$$

It is important to note that we have not actually obtained a way to actually calculate the PID terms yet, but have only stated several logical relationships that such a decomposition should satisfy. The only computable quantities at the moment are the mutual information terms at the left hand side of equations 10 and 11, which can be calculated using equation 6. The discussion of computing the specific PID terms is developed further in the next section, which is heavily inspired by an intuitive overview of [BRO+13], provided by [WPK+15].

TODO! Why does the PID only work with 2 inputs?

TODO? PID applications?

1.2.2 Calculating PID terms

It turns out that the current tools from classical information theory - entropy and various forms of mutual information - are not enough to calculate any of the terms of the PID [WB10]. Indeed, there are only 3 equations (10, 11) relating to the 4 variables of interest, making the system underdetermined. In order to make the problem tractable, a definition of at least one of the PID terms must be given [BRO+13].

Taking inspiration from game theory, [BRO+13] were able to provide such a definition for unique information. Their insight was that if a variable contains unique information, there must be a way to exploit it. In other words, there must exist a situation such that an agent having access to unique information has an advantage over another agent who does not possess this knowledge. Given such a situation, the agent in posession of unique information can prove it to others by designing a bet on the output variable, such that on average, the bet is won by the designer.

In particular, suppose there are two agents - Alice and Bob - Alice having access to the random variable Y and Bob having access to the random variable Z from equation 10. Neither of them have access to the other player's random variable, and both of them can observe, but not directly modify, the output variable X. Alice can prove to Bob that she has unique information about X via Y by constructing a bet on the outcomes of X. Since Alice can only directly modify Y and observe the outcome X, her reward will depend only on the distribution $p(X, Y)$. Similarly, Bob's reward will depend only on the distributon $p(X, Z)$. From this, it follows that the results of the bet are *not* dependent on the full distribution $p(X, Y, Z)$, but rather only on its marginals.

Under the assumption that the unique information depends only on the marginal distribution, a set of probability distributions can be defined which respect the margnials of P such that the unique information stays constant for every element in this set.

$$\begin{aligned}\Delta_P = \{Q \in \Delta : Q(X = x, Y = y) = P(X = x, Y = y) \\ \text{and } Q(X = x, Z = z) = P(X = x, Z = z) \text{ for all } x \in X, y \in Y, z \in Z\}\end{aligned}$$

From the fact that unique information is constant on Δ_P and equation 11, shared information will also be constant on Δ_P. Thus, only synergy varies when considering arbitrary distribution Q from Δ_P.

Now, if we would find a distribution $Q_0 \in \Delta_P$ such that the synergy vanishes, we could find unique information, since from the chain rule for information 9 and decompositions 10 11, the following identities can be derived:

$$\begin{aligned}MI(X : Y|Z) = UI(X : Y \setminus Z) + CI(X : Y, Z) \\ MI(X : Z|Y) = UI(X : Z \setminus Y) + CI(X : Y, Z)\end{aligned} \tag{12}$$

From 12 it can indeed be seen that when synergy is 0, the mutual information and unique information terms coincide. However, [BRO+13] prove that such a distribution $Q_0 \in \Delta_P$ only exists for specific measures of unique, shared and complementary information. They define these measures as follows:

$$\widetilde{UI}(X : Y \setminus Z) = \min_{Q \in \Delta P} MI_Q(X : Y|Z) \tag{13}$$

$$\widetilde{UI}(X : Z \setminus Y) = \min_{Q \in \Delta P} MI_Q(X : Z|Y) \tag{14}$$

$$\widetilde{SI}(X : Y; Z) = \max_{Q \in \Delta_P} MI_Q(X : Y) - MI_Q(X : Y|Z) \tag{15}$$

$$\widetilde{CI}(X : Y; Z) = MI(X : Y, Z) - \min_{Q \in \Delta_P} MI_Q(X : Y, Z) \tag{16}$$

"For this particular choice of measures it can be shown that there is always at least one distribution $Q_0 \in \Delta P$ for which the synergy vanishes, as was desired above." - Neural goal functions paper.

We can see that finding the partial information decomposition terms amount to solving various optimization problems. It is important to note that only one of the the problems needs to be solved, because when one of the terms is found, the remaining ones can be calculated using PID equations.

1.2.3 Numerical Estimator

[BRO+13] show that the optimization problems involved in the definitions of $\widetilde{UI}$, $\widetilde{SI}$ and $\widetilde{CI}$ are convex optimization problems on convex sets.

1.3 Ising Model

In nature, many systems have the property of abruptly transitioning from one state to a completely different state due to some change of external conditions that they are influenced by. Such a phenomenon, where a system does not change its state smoothly, but rather does it in an all-or-nothing fashion, is called a *phase transition*. A large class of phase transitions, which are of great practical importance, can be thought of as shifts from an ordered state to a disordered one, or vice versa. A canonical example of this phenomenon comes form condensed matter phycics, where matter transitions quickly from a fairly ordered solid state to a relatively less organized liquid state when temperature passes a specific threshold.[Bar13].

In this chapter, one of the simplest models that undergoes a phase transition - the Ising model - is analysed in terms of partial information decomposition. Of particular interest is the behaviour of PID terms in relation to the phase transition.

1.3.1 Ferromagnetism

Before introducing the Ising model, a short overview of a phycsical mechanism that it is modelling - ferromagnetism - is in order. This is done in this subsection, which are based on [JNW10].

Electrons in a material have magnetic moments, caused by their spins, which can be in either one of two states. These small magnetic properties of individual electrons do not usually yield a global net magnetization of the material, because the electron in the atoms often come in pairs of opposite spin states, cancelling each other out. However, in ferromagnets, there are many unpaired electrons, which line up with each other, producing a region called a *domain*. While the magnetic field is strong within the domain, the material is still unmagnetized because the many domains themselves are oriented randomly with respect to one another. A characteristic property of a ferromagnetic materials is that even a rather weak external magnetic field can can cause the magnetic domains to line up with each other. When this happens, the material is said to be magnetized. Importantly, in the case of a ferromagnet, the material will remain magnetized even if the influencing external field is removed.

The stability of the magnetization is also dependent on the temperature of the substance. Intuitively, at high temperatures, the atoms in the substance become agitated and start to vibrate. This thermal oscillation breaks the alignment of the spins and the material demagnetizes. This is yet another example of a phase

transition in which an ordered, magnetized system abruptly changes it state to an unordered one. The critical temperature at which this transition happens is called the *Curie temperature*.

1.3.2 Model

The Ising model, first conceived by Wilhelm Lenz in 1920 [Nis05], is a mathematical model of ferromagnetism. The model abstracts away the rather complex details of atomic structures of magnets, consisting simply of a discrete lattice of cells or sites, denoted as s_i, each of which has an associated binary value of either -1 or +1. Conceptually, the lattice can be thought of as a physical material, where the sites roughly represent the unpaired electrons of its atoms. The binary value of each site intuitively corresponds to the direction of the electron's spin. A value of -1 means that the spin is considered to point "down", otherwise it is said to be pointing "up". A given set of spins, denoted as $\boldsymbol{s}$ (without the subscript), is called the *configuration* of the lattice. [Hua87]

The magnetization of a configuration $\boldsymbol{s}$ of an Ising model with a lattice of N sites is given by

$$M(\boldsymbol{s}) = \frac{1}{N} \sum_{i=1}^{N} s_i \tag{17}$$

From equation 17, it can bee seen that the absolute magnetization is small when the number of "up" spins is roughly the same as the number of "down" spins. On the other hand, if all the spins point in the same direction, the absolute magnetization is at it maximum, having a value of 1. This is indeed analogous to the mechanism in play in physical ferromagnets, as described earlier.

The dynamics of the Ising model stem from the fact that the specific spin configurations of the Ising model are random variables. The probability of a configuration $\boldsymbol{s}$ at thermal equilibrium is given by the Boltzmann distribution:

$$P_\beta(\boldsymbol{s}) = \frac{e^{-\beta E(\boldsymbol{s})}}{\sum_{\boldsymbol{s}} e^{-\beta E(\boldsymbol{s})}} \tag{18}$$

where $E(\boldsymbol{s})$ denotes the *energy* associated with the configuration $\boldsymbol{s}$, and $\beta = \frac{1}{k_B T}$ where T is the temperature and k_B is the Boltzmann constant. Thus, β is proportional to the inverse temperature of the system.

The probability of a configuration $\boldsymbol{s}$ depends on 2 unrelated quantities: the internal energy of the configuration under discussion, and the temperature. Two observations that stem from equation 18 are of importance. First, the lower the energy $E(\boldsymbol{s})$ of a configuration $\boldsymbol{s}$, the higher its probability. Second, the higher the temperature T (or equivalently, the lower the parameter β), the more diffuse the distribution becomes. The latter mathematical property models the physical fact that at high temperatures, the thermal oscillation of the atoms break the alignment of the spins, demagnetizing the material.

Energy is a central quantitiy that is associated with almost any model in physics. In the Ising model, energy is given by the Hamiltonian

$$E(s) = -\sum_{\langle ij \rangle} \epsilon s_i s_j - H \sum_{i=1}^{N} s_i \tag{19}$$

where the first sum is over all different neighboring spins, ϵ is the interaction strength between adjacent spins, and H denotes the strength of an exetrnal magnetic field. The latter two quantities are given constants that are specified by the properties of the magnetic material and the external environment of the system, respectively.

Often, the model is simplified even further. In particular, the external magnetic field interacting with the lattice is omitted, and the interaction strength between pairs of nearest neighbors is fixed to be equal to the Boltzmann constant k_B, so they cancel each other out and β becomes exactly the inverse temperature $\frac{1}{T}$. After incorporating these assumptions into the model, the energy of a configuration $\boldsymbol{s}$ simplifies to

$$H(\boldsymbol{s}) = -\sum_{\langle ij \rangle} s_i s_j \tag{20}$$

From equation 20, it can be seen that the spins in the Ising model directly interact with only their nearest neighbors. Moreover, since there is a minus sign in front of the sum, lower energy (and thus, a higher probability) is achieved when neighboring spins take on the same value, as this yields a positive product. It can be intuitively thought as if the spins are intrinsically trying to align with their neighbors, and the temperature of the system quantifies the amount of prohibition that prevents them from doing so.

There are many questions one could ask about the dynamics of the Ising model, but perhaps the most interesting and most extensively studied is the following:

how does the magnetization of the lattice change with temperature? Since for a fixed value of β, the lattice configurations are random variables, an *expectation* of the magnetization must be found, using the following formula:

$$\langle M \rangle_\beta = \sum_{\boldsymbol{s}} M_\beta(\boldsymbol{s}) P_\beta(\boldsymbol{s}) \tag{21}$$

By saying that there is a phase transition in the Ising model, what is meant that there exists a critical temperature T_c such that for temperatures $T > T_c$, the absolute value of the expected magnetization given by equation 21 is 0 (or quickly approaches zero if T is near T_c). Conversely, if $T < T_c$, the absolute magnetization is 1. Ernst Ising himself proved that there is no sponatneous magnetization and therefore, no phase transition, in the 1-dimensional model [Isi25]. In 1944, Lars Onsanger showed [Ons44] that the 2-dimensional Ising model with a square lattice in the absence of an external magnetic field indeed undergoes a phase transition, and furthermore gave the exact value of the parameter β at which the swift order-disorder transition takes place. For higher dimension, no analytic solution exists.

2 Related Work

Barnett et al [BLH+13] argue that "In a system comprising a large number of interacting elements with an order-disorder phase transition, it is easy to argue that mutual information between elements must peak at an intermediate order: for a highly ordered system, there is little indeterminacy about the state of individual elements and hence mutual information between elements will be small, while for a highly disordered system, elements will behave near independently and again mutual information between elements will be small. It also seems reasonable ... to expect that the peak will occur at the phase transition, where susceptibility peaks."

Indeed, it has been shown that in various dynamical complex systems, mutual information peaks at the critical point where the system undergoes an order-disorder transition. For example, such is the case for random boolean networks [LPZ08], swarms models [WCD07], as well as in real world systems, such as [HB09]

In the case of the Ising model, it has been both analytically

Lizier showed that transfer entropy peaks before the phase transition. [BLH+13]

Ising analytic [MKN+96]

Ising numeric [WTV11]s

3 Methods

3.1 Numerical simulation of the Ising model

In theory, finding the expected magnetization for a given temperature T is trivial. According to equation 21, one simply has to enumerate all lattice configurations, multiply their probabilities by their magnetizations, and sum the products. The problems arise in the very first part - enumerating all the configurations. The number of possible configurations of a lattice of size N is 2^N, meaning that the number of configurations increases exponentially in the size of the lattice. Therefore, the sum in equation 21 is intractable for even rather modest sized lattices. This is of course a more general problem that is not only present when calculating the expected magnetization, but rather appears in any task where one has to deal with expectations in the Ising model. For example, in this thesis, we would like to find the expected mutual information between the sites for each temperature point. Unable to do so anlytically, one must resort to simulating the dynamics of the model.

Because enumerating all possible configurations is intractable, a more clever solution must be found. One way to *approximate* the average quantities is to draw many samples (spin configurations) from the Boltzmann distribution and calculate the quantities of interest on these configurations, taking their mean in the end. If the configurations are drawn in proportion to their probabilites given by equation 18, the mean of the quantity of interest will become closer to the true expectation as the number of samples increases. Glauber dynamics [MRR$^+$53], a method of sampling a given probability distribution via a Markov chain `http://www4.ncsu.edu/~dyeun/pub/techrep-csma12.pdf`, allows one to iteratively draw samples from the Boltzmann distribution according to their probabilities.

The Glauber dynamics method works as follows. First, an initial lattice configuration is chosen arbitrarily (as discussed later, some clever tricks in choosing the initial configuration can be done, however). Then, iteratively, a site is chosen uniformly at random from the lattice, and the spin associated with this site is flipped. The new configuration is either accepted or rejected. The probability of acceptance is given by equation 22, where $\boldsymbol{s}_n$ denotes the new lattice configuration with a flipped spin, T denotes the temperature and $\Delta E(\boldsymbol{s} \to \boldsymbol{s}_n) = E(\boldsymbol{s}) - E(\boldsymbol{s}_n)$ is the difference between the energies of the two configurations. For a more compact overview, a pseudocode for a single iteration of this method is given in algorithm 9.

$$P(\boldsymbol{s} \to \boldsymbol{s}_n) = \frac{1}{1 + e^{\frac{\Delta E(\boldsymbol{s} \to \boldsymbol{s}_n)}{T}}} \tag{22}$$

Algorithm 1: A single iteration of Glauber dynamics.

1 **Input:** A lattice configuration $\boldsymbol{s}$
2 Choose a random site from the lattice;
3 Flip the spin associated with the chosen site to obtain a configuration $\boldsymbol{s}_n$;
4 Calculate the energy difference: $\Delta E(\boldsymbol{s} \to \boldsymbol{s}_n) = E(\boldsymbol{s}) - E(\boldsymbol{s}_n)$;
5 Generate a random number x uniformly at random from the interval [0, 1];
6 **if** $x \leq P(\boldsymbol{s} \to \boldsymbol{s}_n)$ **then**
7 **return** $\boldsymbol{s}_n$ ▷ accept the new configuration $\boldsymbol{s}_n$ by returning it
8 **else**
9 **return** $\boldsymbol{s}$ ▷ reject $\boldsymbol{s}_n$ by returning $\boldsymbol{s}$

There are two noteworthy additions to the naive algorithm 9 that must be discussed. First, to uncorrelate the samples, many potential spin flips are considered before a sample is actually drawn, meaning that not every lattice configuration returned by algorithm 9 is considered as a sample, but rather every L-th. The parameter L is a called the *lag*. This procedure is illustrated in algorithm 4. Notice that indeed, all the intermediate configurations on line 3 are discarded, only the L-th configuration is eventually returned. Second, in order to avoid biasing the initial samples towards the initial random configuration, the very first samples are discarded. The number of initial discarded samples is referred to as the *burn in period*. The entire Glauber dynamics method, along with the random initialization, burn-in period and lag is illustrated in algorithm 9

Algorithm 2: An update of the Glauber dynamics where one unit of time is considered to be N spin-flip attempts.

1 **Input:** A lattice configuration $\boldsymbol{s}$ and lag L
2 **for** $i = 1...L$ **do**
3 $\boldsymbol{s}$ = Run algorithm 9 on input $\boldsymbol{s}$;
4 **return** $\boldsymbol{s}$;

TODO! Faster calculation of energy difference

Algorithm 3: The full Glauber dynamics method.

Input: Burn-in period B, lag L, and the number of samples to draw N
Initialize a random lattice configuration $\boldsymbol{s}$;
for $i = 1...B$ **do**
    s = Run algorithm 4 on input s and L;
samples = [ ] ▷ List to save the sampled configurations to
for $i = 1...N$ **do**
    s = Run algorithm 4 on input s and L;
    save configuration $\boldsymbol{s}$ to *samples*;
return *samples*

3.2 Experimental setup

To estimate the PID terms in the Ising model, a 2-dimensional Ising model with a square lattice of size 128x128 was simulated. A single simulation consisted of a burn-in period of 10^4 updates, following 10^5 updates from which the samples were gathered. As in [BLH$^+$13], each update comprised of N potential spin-flips according to Glauber transition probabilities, where N is the size of the lattice. In other words, the model was simulated according to algorithm 9, where $B = 10^4$, $L = 128x128$ and $N = 10^5$. This procedure was performed at 102 temperature points spaced evenly over the interval [2.0, 2.8], which encloses the phase transition at $T = 2.268$

The obtained 10^5 lattice configurations were subsequently used to construct the probability distributions that the PID estimator takes as input. In particular, 100 sites were chosen uniformly at random, and for each of these sites, the relative frequency its configuration along with 4 of its immediate neighbors was measured at each temperature point. Thus, for each temperature point, 100 joint probability distributions of 5 random variables that can take on 32 different values was generated. An illustrative example of one such distribution is given by table 2, where the first random variable X represents center site, and the 4 other random variables represent its immediate neighbors. For example, the second row of the table tells us that the configuration where the center site has a downward spin, along with all its neighbors except the left, appears 30 of the time (out of a total 10^5 configurations that were simulated) at some specific temperature T_i.

The PID estimator works with probability distributions of 3 random variables, where one of them is thought of as input and the remaining as outputs. Thus, it must be decided how are neighboring sites paritioned into input and outputs, and how are 2 outputs chosen out of 4 random variables. Answering the first question,

the neighbors were considered as sources and the center site was chosen as a target random variable. As for the second question, to different approaches were taken. First, only 2 neighbors were chosen without repetitions uniformly at random out of the possible set of 4 neighbors. Second, the 4 neighbors of a site were randomly partitioned into 2 disjoint pairs, such that each pair was now a 2- dimensional random vector.

C	U	R	D	L	Pr
-1	-1	-1	-1	-1	0.1
-1	-1	-1	-1	1	0.05
-1	-1	-1	1	-1	0.03
-1	-1	-1	1	1	0.25

Table 4:

This procedure was performed for 8 runs at 102 temperature points spaced evenly over the interval [2.0, 2.8]. In the very first run, each configuration was initialized randomly for each temperature point, and the lattice configuration that was arrived at after the burn in period of 10^4 updates was saved. For subsequent 7 runs, the very first lattice configuration for temperature point T_i was chosen to be equivalent to the saved lattice configuration from the very first run at temperature point T_i. To gather the 10^5 lattice configurations for all temperature points for 8 different runs, simulations were run for 8 days on 41 computing nodes in parallel in the EENet computer clustser.

good ising slides

glauber dynamics

4 Results

To be confident that the Ising model simulations behave as expected, the average absolute magnetization of the 8 runs was measured. The resulting plot can be seen in figure 2. The phase transition is clearly present, and the critical temperature is around the theoretically correct value of $T_c = 2.68$. At temperatures $T > T_c$, the magnetization of the Ising model is near 0, while at temperatures $T < T_c$, the absolute magnetization quickly approaches 1. This agrees with the physical observations of ferromagnetisms as well as with previous practical and theoretical works, thus validating that the simulation of the Ising model is done correctly.

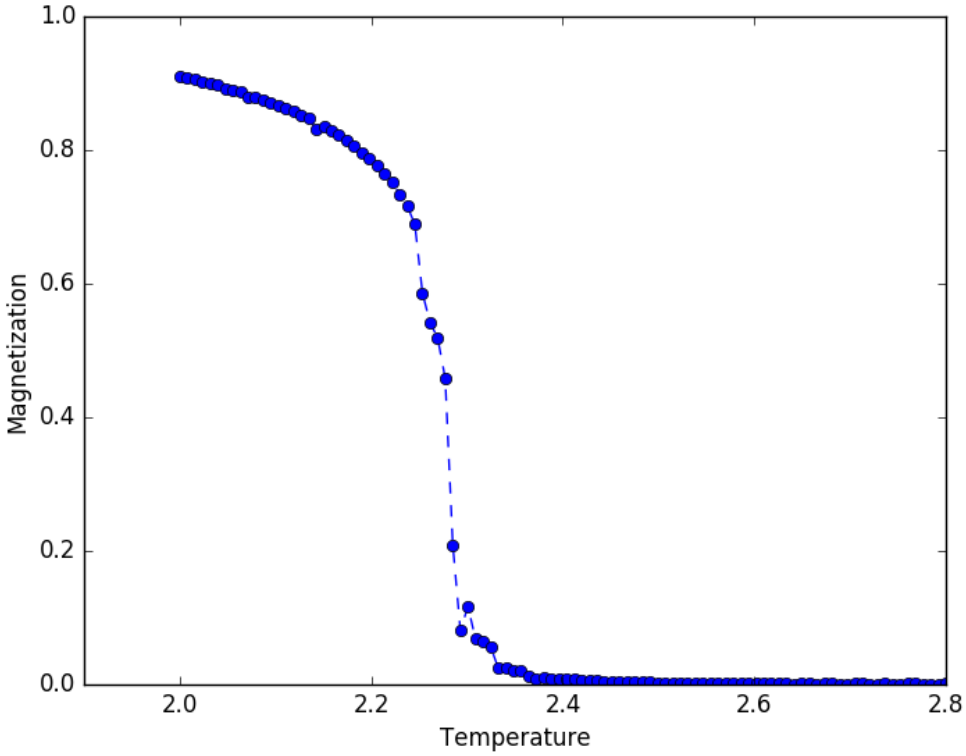

Figure 2: Average absolute magnetization of a 128x128 lattice evaluated at 102 temperature points spaced evenly over the interval [2.0, 2.8].

In figure 3, the information theoretic functionals of the Ising model can be observed, where the mutual information is measured between 100 randomly chosen sites and their 2 random neighbours. Notice that the information is given in nats, meaning that the base of the logarithm in equation 6 is taken to be e.

From the figure, it can be seen that mutual information peaks at the phase transition, which agrees with the previous work (for example [BLH$^+$13]), further validating that the method of estimating the information theoretic terms used in this thesis works as expected. In addition, since in our experiment, the mutual information was measured between a site and 2 of its neighbors, as opposed to measuring it between 2 neighboring sites only, it would be reasonable to expect that in our experiments, the mutual information is higher, because two neighbors should have more information about their center site than a single neighbor has. This is indeed the case - in [BLH$^+$13], the I_{pw} measure peaks at the value of just under 0.3, considerably less than 0.5, which is the case in our experiment.

Nats!?

Looking at the partial information decomposition of the Ising model in figure 3, what pops out right away is that the nonzero terms peak exactly at the phase transition, just like mutual information itself. Visually, the shared information curve follows the mutual information graph almost exactly, with the expection of being shifted downwads about 1.5 nats at every temperature point. The synergetic information term is more interesting. It still peaks at the phase transition, but it behaves differently from the mutual information, being quite a bit flatter. Both of the unique information terms are rather unintering, as they are near zero throughout the entire observational period of the Ising model.

Shared information is the most dominant term in the partial information decomposition in figure 3, meaning that there is an unproportional amount of redundancy between the neighbors with respect to their center site at all temperature points. Intuitively, this is to be expected, as the neighbors of the center site are directly influencing it to take on the same value as them, and vice versa. Thus, reasoning by transitivity, a neighboring site A tries to orient its spin to be parallel to the center spin, and similarly, the center site tries to align its spin such that it points in the same direction as the spin of another neighbor B. Because A and B are actively trying to make their spins parallel to each other through the influence of the center site, it is reasonable to assume that they have a lot of redundancy between them.

The complementary information ...

What is also interesting and noteworthy is that both redundant and synergetic information peak at the phase transition.

The unique information terms are not ...

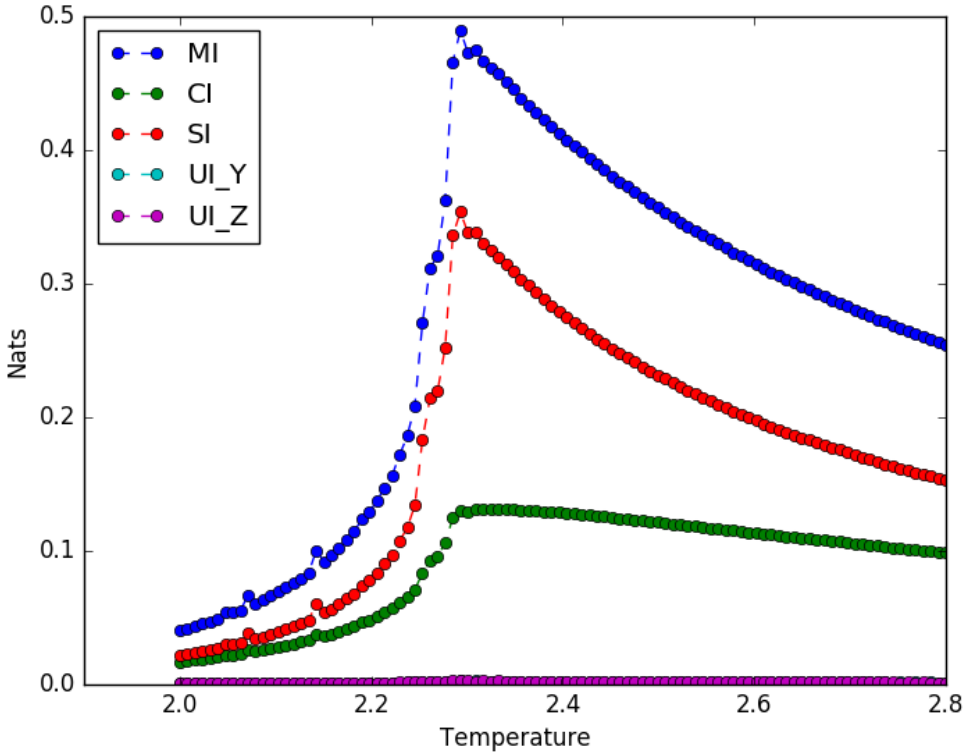

Figure 3: Average mutual information and PID terms of a 128x128 lattice Ising model.

5 Discussion

5.1 Implications of the results

5.2 Limitations

5.3 Future work

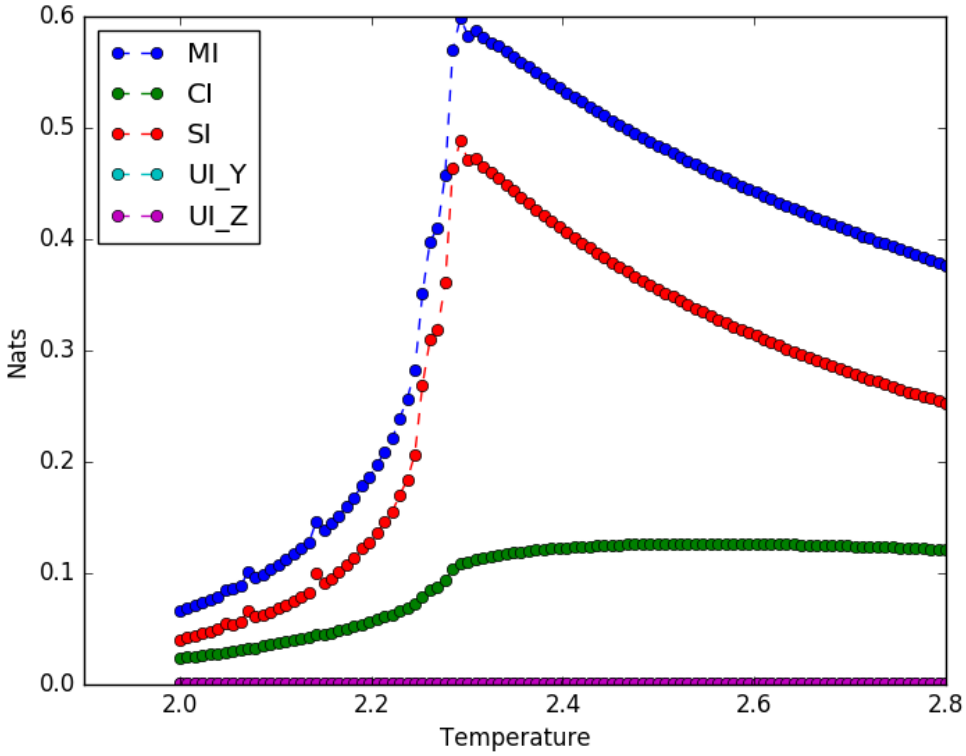

Figure 4: Ising 128x128 PID 4 nbs

Conclusion

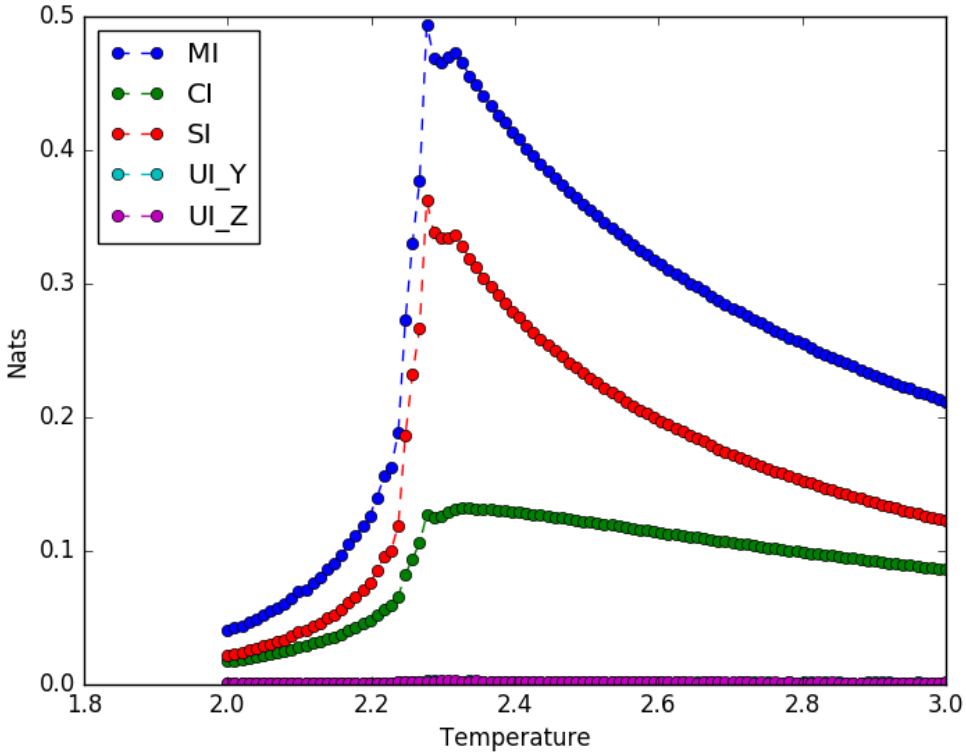

Figure 5: Ising 64x64 PID 2 nbs

References

[Bar13] Lionel Barnett. A commentary on Information flow in a kinetic Ising model peaks in the disordered phase. `http://users.sussex.ac.uk/~lionelb/Ising_TE_commentary.html`, 2013. [Online; accessed 06-April-2017].

[BLH+13] L Barnett, J T Lizier, M Harré, A K Seth, and T Bossomaier. Information flow in a kinetic ising model peaks in the disordered phase. *Phys Rev Lett*, 111(17):177203–177203, Oct 2013.

[BRO+13] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *CoRR*, abs/1311.2852, 2013.

[CHLH+14] Robin Carhart-Harris, Robert Leech, Peter Hellyer, Murray Shanahan, Amanda Feilding, Enzo Tagliazucchi, Dante Chialvo, and David

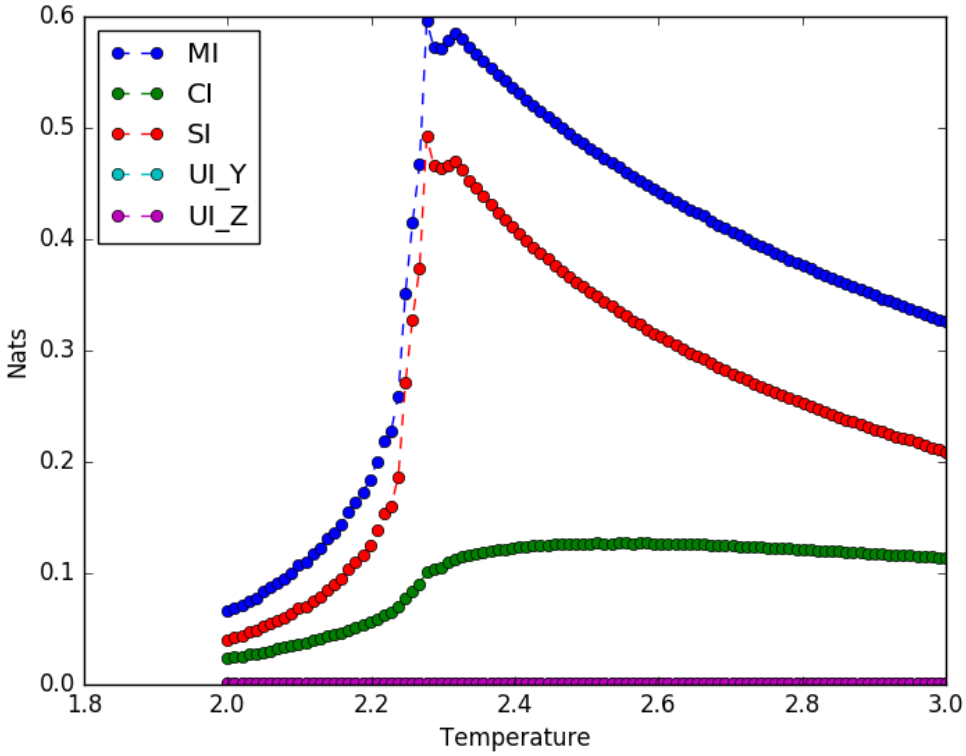

Figure 6: Ising 64x64 PID 4 nbs

Nutt. The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience*, 8:20, 2014.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[GK12] Virgil Griffith and Christof Koch. Quantifying synergistic mutual information, 2012.

[HB09] M. Harré and T. Bossomaier. Phase-transition–like behaviour of information measures in financial markets. *EPL (Europhysics Letters)*, 87(1):18009, 2009.

[HSP12] Malte Harder, Christoph Salge, and Daniel Polani. A bivariate measure of redundant information. *CoRR*, abs/1207.2080, 2012.

[Hua87] Kerson Huang. *Statistical Mechanics. (Second Edition).* John Wiley & Sons, 1987.

[Isi25] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, 1925.

[JNW10] Bruce Jacob, Spencer Ng, and David Wang. *Memory Systems: Cache, DRAM, Disk*. Morgan Kaufmann, 2010.

[LPZ08] Joseph T. Lizier, Mikhail Prokopenko, and Albert Y. Zomaya. The information dynamics of phase transitions in random boolean networks. In *Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems (ALife XI*, pages 374–381. MIT Press, 2008.

[MKN+96] Hiroyuki Matsuda, Kiyoshi Kudo, Ryoku Nakamura, Osamu Yamakawa, and Takuo Murata. Mutual information of ising systems. *International Journal of Theoretical Physics*, 35(4):839–845, 1996.

[MRR+53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[Nis05] Martin Niss. History of the lenz-ising model 1920-1950: From ferromagnetic to cooperative phenomena. *Archive for History of Exact Sciences*, 59(3):267–318, 2005.

[Ons44] Lars Onsager. Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Phys. Rev.*, 65:117–149, Feb 1944.

[PLY06] L. Peiyu, J. Lijie, and W. Yongqing. Application of maximum entropy in engineering structual optimization. In *2006 7th International Conference on Computer-Aided Industrial Design and Conceptual Design*, pages 1–5, Nov 2006.

[SAG+10] R. Salvador, M. Anguera, J. J. Gomar, E. T. Bullmore, and E. Pomarol-Clotet. Conditional mutual information maps as descriptors of net connectivity levels in the brain. *Front Neuroinform*, 4:115, 2010.

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[WB10] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010.

[WCD07] R. T. Wicks, S. C. Chapman, and R. O. Dendy. Mutual information as a tool for identifying phase transitions in dynamical complex systems with limited data. *Phys. Rev. E*, 75:051125, May 2007.

[WPK+15] Michael Wibral, Viola Priesemann, Jim W. Kay, Joseph T. Lizier, and William A. Phillips. Partial information decomposition as a unified approach to the specification of neural goal functions. 2015.

[WTV11] Johannes Wilms, Matthias Troyer, and Frank Verstraete. Mutual information in classical spin models. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(10):P10011, 2011.

[ZCT13] Rongxi Zhou, Ru Cai, and Guanqun Tong. Applications of entropy in finance: A review. *Entropy*, 15(11):4909–4931, 2013.

Non-exclusive licence to reproduce thesis and make thesis public

I, Sten Sootla (date of birth: 17th of January 1995),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Analysing information distribution in complex systems

supervised by Raul Vicente Zafra and Dirk Oliver Theis

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu, dd.mm.yyyy