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

1 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.

2 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 $p(x_i, y_j) = Pr\{X = x_i, Y = y_j\}$ $(i = 1, ..., n;\ j = 1, ..., m)$.

2.1 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:

$$H(X) = -\sum_{i=1}^{n} p_X(x_i) \log_2 p_X(x_i) \tag{1}$$

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*.

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 $-\log_2 p_X(x_i)$. Viewed from this angle, the definition of entropy can be rewritten as an expectation of the random variable $-\log_2 p(X)$:

$$H(X) = \mathbb{E}\left[-\log_2 p(X)\right] = \mathbb{E}\left[\log_2 \frac{1}{p(X)}\right].$$

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].

2.2 Joint and Conditional Entropy

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

$$H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i, y_j) \tag{2}$$

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

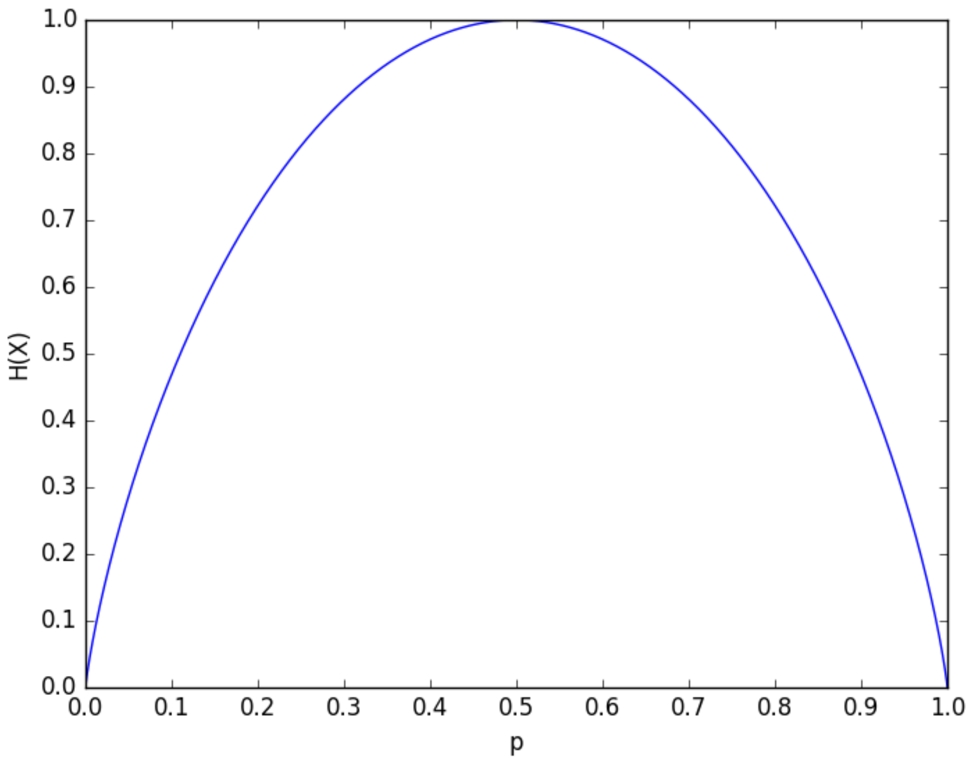

Figure 1: Entropy of X plotted against the value of p.

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

$$H(Y|X) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 p(y_j|x_i) \tag{3}$$

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]:

$$\begin{aligned}
H(X,Y) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log_2 p(x_i,y_j) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log_2 p_X(x_i)p(y_j|x_i) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log_2 p_X(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log_2 p(y_j|x_i) \\
&= -\sum_{i=1}^{n} p_X(x_i)\log_2 p_X(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_j)\log_2 p(y_j|x_i) \\
&= H(X) + H(Y|X)
\end{aligned} \tag{4}$$

2.3 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

$$D(p||q) = \sum_{i=1}^{n} p(x_i)\log_2 \frac{p_X(x_i)}{q_X(x_i)} \tag{5}$$

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.

2.4 Mutual information

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

$$MI(X;Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p_X(x_i)p_Y(y_j)} \tag{6}$$

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:

$$\begin{aligned} MI(X;Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p_X(x_i)p_Y(y_j)} \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 p_X(x_i) + -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i|y_j) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p_X(x_i) - \left(-\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i|y_j) \right) \\ &= H(X) - H(X|Y) \end{aligned} \tag{7}$$

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.

2.5 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].

2.6 Transfer Entropy

TODO!?

3 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.

3.1 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 3 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?

3.2 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.

3.3 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.

4 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 from a fairly ordered solid state to a relatively less organized liquid state very abruptly 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.

4.1 Background

Before introducing the Ising model, a short overview of a phycsical mechanism that it is modelling - ferromagnetism - is in order. This is done in the following 2 paragraphs, both of 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.*

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.

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 $\boldsymbol{s}$. $\beta = \frac{1}{k_B T}$ is the *inverse temperature* of the system, where T is the temperature and k_B is the Boltzmann constant.

Thus, the probability of a configuration $\boldsymbol{s}$ depends on 2 unrelated quantities: the internal energy of the configuration under discussion, and the temperature. Moreover, two important observations that stem from equation 18 are important to be worth stating explicitly. First, the lower the energy of the configuration, the higher its probability. Second, the higher the temperature T (or equivalently, the lower the inverse temperature β), the more diffuse the proabilities become. 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.

From equation 19, it can be seen that the spins in the Ising model directly interact with only their nearest neighbors. If the interaction energy ϵ is positive, the configurations where neighboring spins have aligned have higher probability. Indeed, when $\epsilon > 0$,

the spins can be thought of as trying to align, as this would decrease the value of the energy function, which in turn would increase the probability of the configuration. Conversely, a negative interaction strength favors configurations where neighboring spins are pointing in the opposite directions.

More often than not, the model is simplified even further. In particular, the internal magnetic field interacting with the lattice is omitted, and the interaction strength between pairs of nearest neighbors is 1. After incorporating these assumptions into the model, the energy of a configuration $\boldsymbol{s}$ simplifies to

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

Intuitively, we can think of it that the spins are trying to line up intrinsically, but the temperature of the system prohibits them from doing so. When the temperature decreases, the local interactions of the spins win out, the spins align, and the lattice becomes magnetized.

Given a temperature, one can find the expected magnetization by the following formula

$$\langle M \rangle = \sum_{\boldsymbol{s}} M(\boldsymbol{s}) P(\boldsymbol{s}) \tag{21}$$

However, 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.

To overcome this problem, the expectation given by equation 21 must be approximated. This can be done using the Metropolis algorithm [MRR$^+$53] - an instance of a more general class of algorithms called the Markov Chain Monte Carlo algorithms. The Hastings algorithm works by iteratively drawing uncorrelated samples from the Boltzmann distribution according to their probabilities. An important charateristic of the algorithm, relevant to the methods section of this thesis, is that the initial samples that are drawn are not representative of the distribution. Thus, the algorithm has a warm-up period from which samples should be disgarded.

4.2 Related Work

It has been shown that the mutual information peaks at the phase transition.

4.3 Experimental Setup

4.4 Results

5 Neural Networks

5.1 Background

5.2 Related Work

5.3 Experimental Setup

5.4 Results

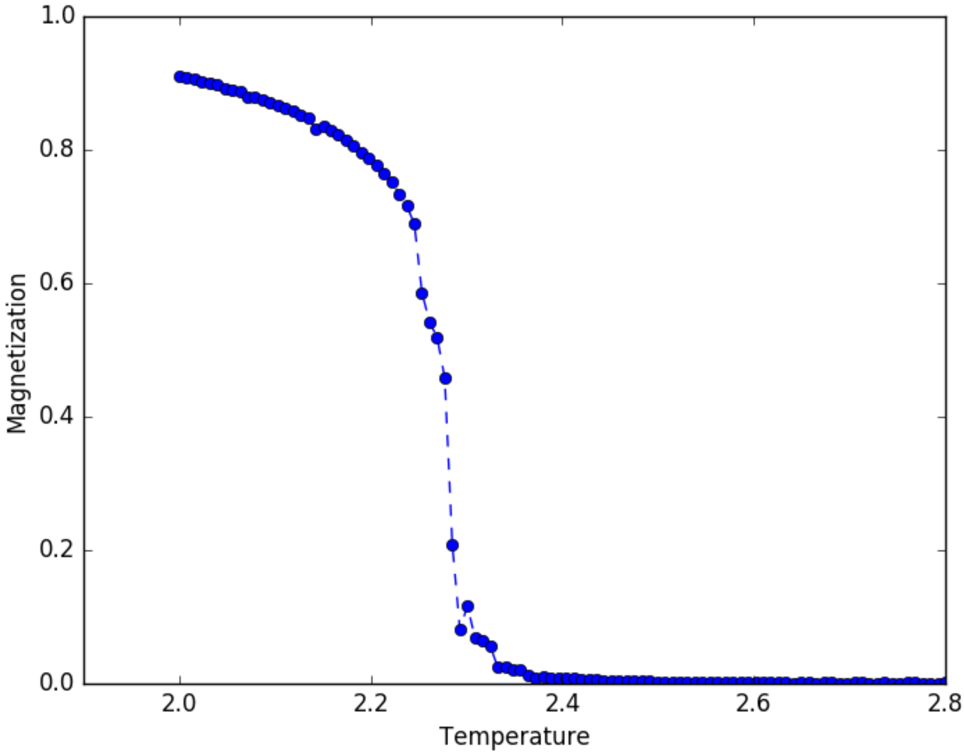

Figure 2: Magnetization

We saw that in the Ising model synergy peaks before the phase transition.

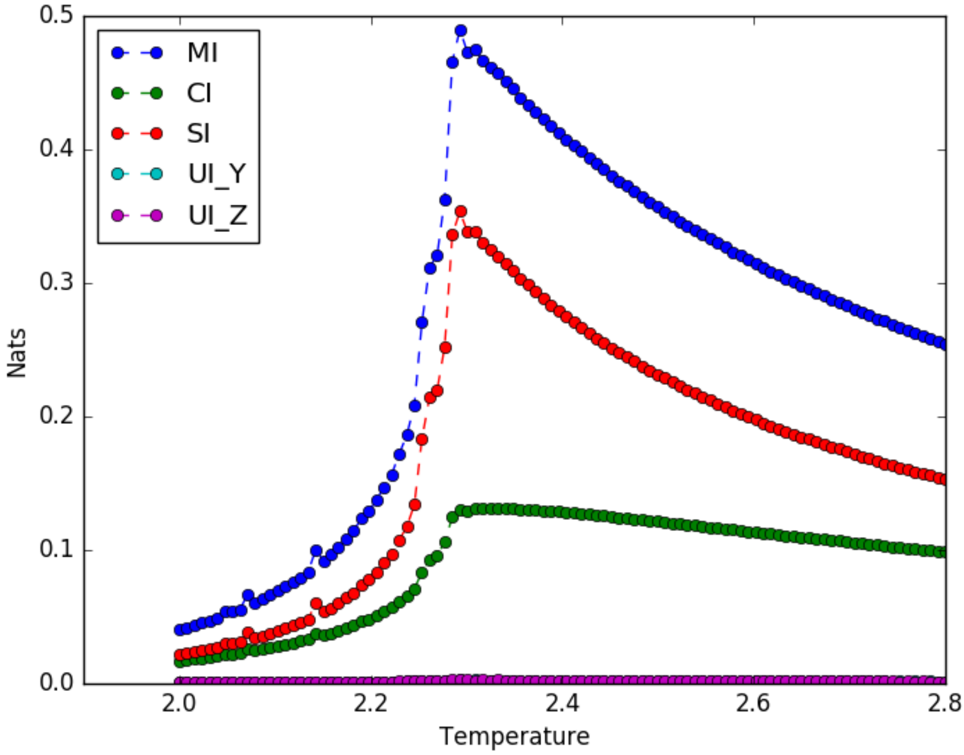

Figure 3: Ising 128x128 PID 2 nbs

6 Discussion

6.1 Limitations

6.2 Future work

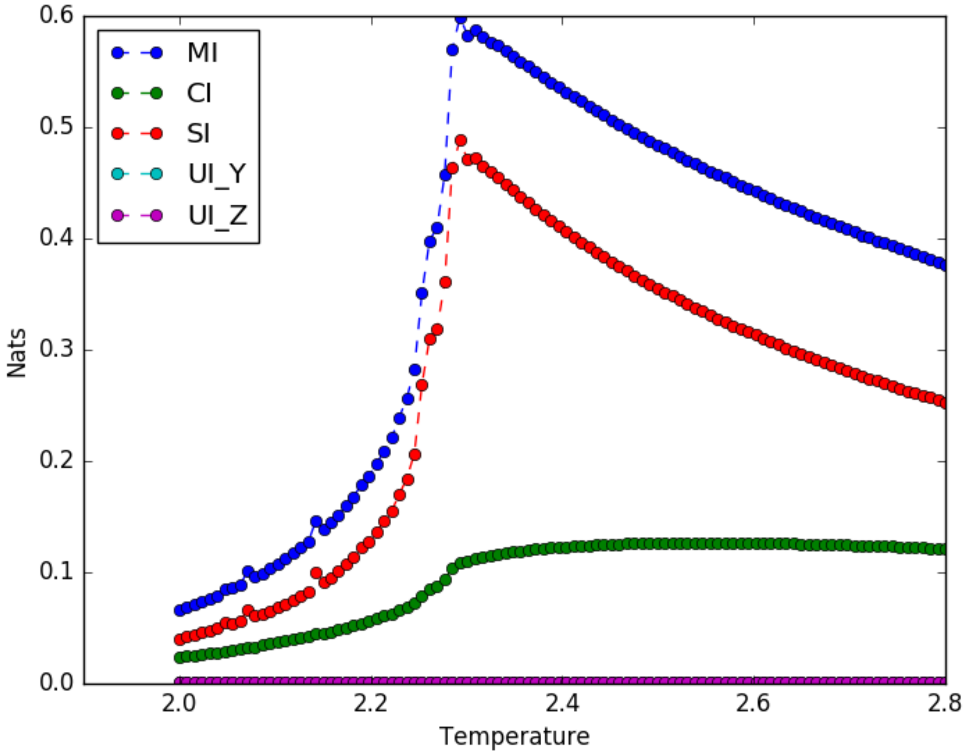

Figure 4: Ising 128x128 PID 4 nbs

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].

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

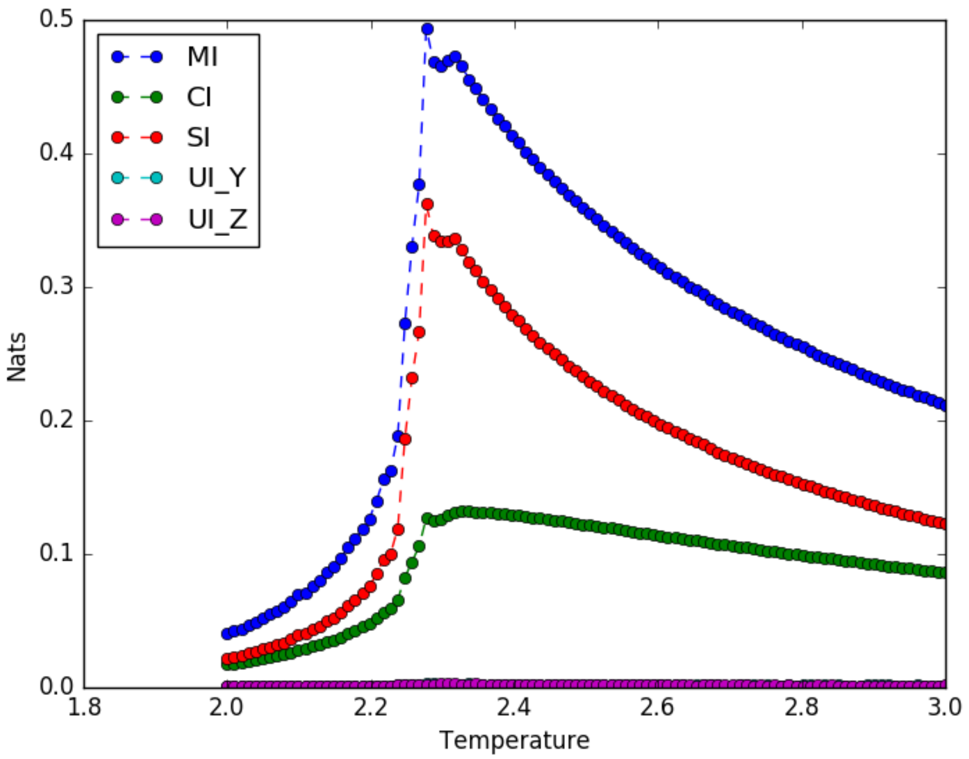

Figure 5: Ising 64x64 PID 2 nbs

[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.

[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.

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

[MRR$^+$53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calcu-

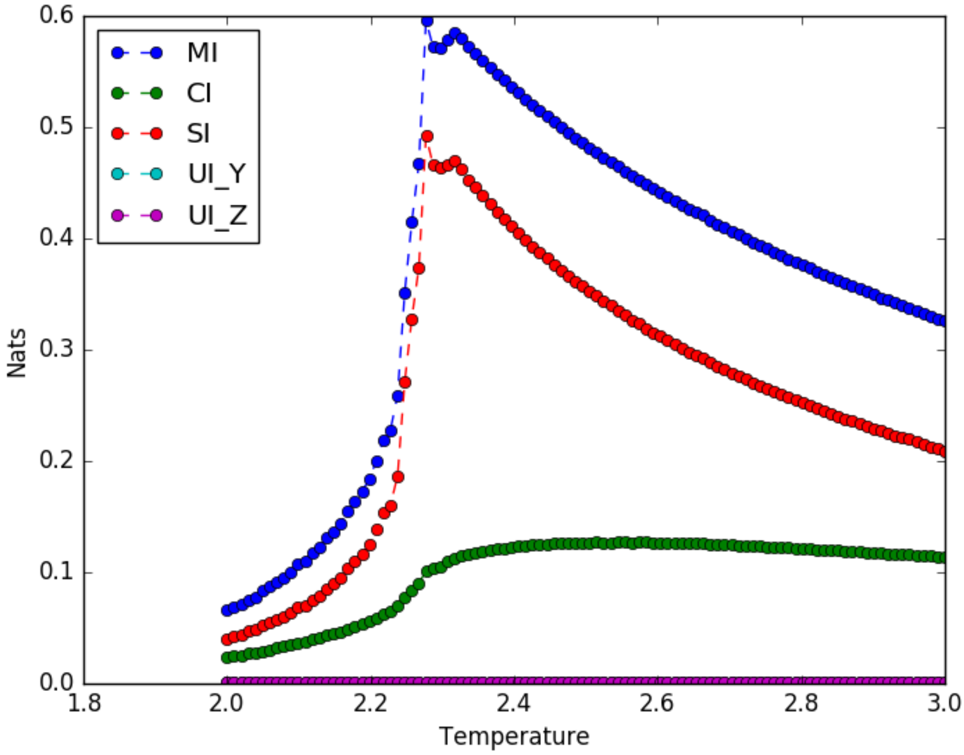

Figure 6: Ising 64x64 PID 4 nbs

lations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[Mye15] Myers. *Introductory Solid State Physics, 2Nd Edition*. T&F India, 2015.

[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.

[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.

[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.

[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