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

TODO!

Vabandust, et sissejuhatust veel hetkel ei ole. Minu jaoks on alati sissejuhatus kõige raskem osa olnud kirjatükkides, ja pole olnud kordagi, kus sissejuhatus poleks jäänud absoluutselt viimaseks osaks. Sissejuhatus annab ülevaate kogu minu tööle, mistõttu arvan, et seda on paslik koostada siis, kui on millest ülevaade teha.

Kui sissejuhatuse eesmärk hetkel on lihtsalt töö läbimõtlemine ja kirjutamisoskuse demonstreerimine, siis ehk piisab hetkel sisukorrast ja esimesest peatükist, mis tasapisi edeneb. Kui mitte, siis mõistagi tuleb mul see sissejuhatus lihtsalt ära teha kiiremas korras.

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 and concrete examples.

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 $p_Y(y_j) = Pr\{Y = y_j\}$ $(j = 1, ..., m)$.

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:

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

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 of x_i of a random variable X can be quantified by $-log_2 p(x_i)$. 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 $p \in [0, 1]$. When $p \in \{0, 1\}$, then the outcome of the coin toss is completely deterministic, meaning there is no uncertainty in the outcome. Accordingly, the entropy is 0 at these points. Conversely, when the coin is fair, we are completely uncertain about the outcome, unable to favour neither heads or tails. Again, the mathematical definition and intuition agree, as the entropy is indeed at its maximum when $p = 0.5$.

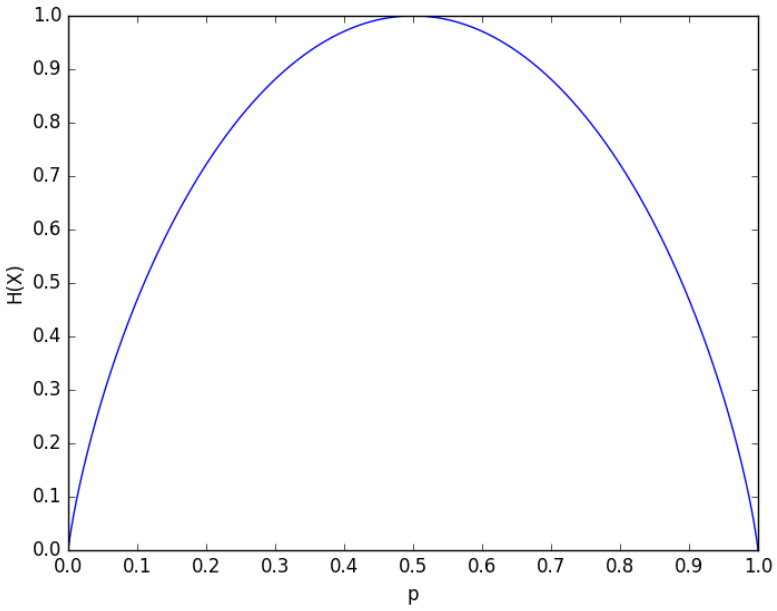

Figure 1: Entropy of X plotted against the value of p.

2.2 Joint and Conditional Entropy

Let the joint distribution of the random variables X and Y be $p(x_i, y_j) = Pr\{X = x_i, Y = y_j\}$ $(i = 1, ..., n;\ j = 1, ..., m)$. 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.

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 other [CT06]:

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

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 the distance between two probability mass functions. Importantly, the relative entropy is non-negative, with inequality exactly when the 2 probability distributions are equal [CT06], again corresponding to our intuitive notion of distance. Indeed, when the two probability mass functions are equal, 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 indeed 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]. Because mutual information is symmetric, the converse statement would also hold, meaning that 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].

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 ensebmble of "input" random variables carry about some "output" 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 pair 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 sources 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 **??** the mutual information between the pair (Y, Z) and X is defined as

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

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 output 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 the 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 computed using equation 6. The discussion of calculating the specific PID terms is developer further in the next section.

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

TODO? PID applications?

3.2 Calculating PID terms

TODO? Should I mention here that the section is based on the PID neural goal functions paper and omit the references after every passage?

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$^+$13] were able to provide such a definition for unique information. Their insight was that if a variable contains unique information, there must be a way to exploit it. In other words, there must exist a situation such that an agent having access to unique information has an advantage over another agent who does not possess this knowledge. Given such a situation, the agent in posession of unique information can prove it to others by designing a bet on the output variable, such that on average, the bet is won by the designer. [WPK$^+$15]

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 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 $p(X, Z)$ and as a result, the results of the bet are *not* dependent on the full distribution $p(X, Y, Z)$, but rather only on its marginals. [WPK$^+$15].

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 trivially derived:

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

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

does this minimization always produce 0 synergy in 12 ??

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 Related work

5 Elementary cellular automata

5.1 Problem description

5.2 Experimental setup

5.3 Results

5.4 Discussion

6 Ising model

6.1 Problem description

6.2 Experimental setup

6.3 Results

6.4 Discussion

7 Discussion

7.1 Limitations

7.2 Future work

References

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

[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.

[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.

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