

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

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Background</b>	<b>4</b>
2.1	Classical information theory . . . . .	4
2.1.1	Entropy . . . . .	4
2.1.2	Joint and Conditional Entropy . . . . .	6
2.1.3	Kullback-Leibler distance . . . . .	6
2.1.4	Mutual information . . . . .	7
2.1.5	Conditional mutual information . . . . .	8
2.2	Partial information decomposition . . . . .	8
2.3	Numerical estimator . . . . .	8
<b>3</b>	<b>Elementary cellular automata</b>	<b>9</b>
3.1	Problem description . . . . .	9
3.2	Related work . . . . .	9
3.3	Experimental setup . . . . .	9
3.4	Results . . . . .	9
3.5	Discussion . . . . .	9
<b>4</b>	<b>Ising model</b>	<b>10</b>
4.1	Problem description . . . . .	10
4.2	Related work . . . . .	10
4.3	Experimental setup . . . . .	10
4.4	Results . . . . .	10
4.5	Discussion . . . . .	10
<b>5</b>	<b>Neural networks</b>	<b>11</b>
5.1	Problem description . . . . .	11
5.2	Related work . . . . .	11
5.3	Experimental setup . . . . .	11
5.4	Results . . . . .	11
5.5	Discussion . . . . .	11
<b>6</b>	<b>Conclusion</b>	<b>12</b>

# 1 Introduction

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

Lubage mul tuua näide elust. Kaks aastat tagasi kohtusin tüdrukuga, kes on nüüdseks minu ekstüdruk. Kui minult oleks palutud kirjutada sel hetkel ülevaade oma ülejäänud elust, siis suure osa kirjaruumist oleks kahtlemata hõlmanud see tüdruk. Õnneks ei alustanud ma tol hetkel ülevaate andmist oma tulevikust, sest nüüd oleme juba kuid lahus. Mõistagi muudab see resultaati minu elulooraamatu sisu märgatavalt. Hea, et ei hakanud elulooraamatusse ennatlikult tulevikust kirjutama, muidu oleks see olnud mahavisatud aeg. Veelgi enam, õnnelikel aegadel kirjutatud tulevikuplaanide lugemine ei oleks hetkel kindlasti kõige tervislikum ajaviide, mil suhe juba purunenud.

Kui sissejuhatuse eesmärk nii varajases staadiumis 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 sissejuhatatus lihtsalt ära teha kiiremas korras :/

In Chapter 1, the basics of classical 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 it.

Chapter 2 - Elementary Cellular Automata Chapter 3 - Ising model Chapter 4 - Artificial Neural Networks

In the final, concluding chapter, a summary of the contributions of this thesis is given, alongside suggestions for further work.

## 2 Background

### 2.1 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 subsection 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  $\{x_1, x_2, \dots, x_n\}$  and a probability mass function  $p_X(x_i) = \Pr\{X = x_i\}$  ( $i = 1, \dots, n$ ). Similarly,  $Y$  is a discrete random variable with possible realizations from the set  $\{y_1, y_2, \dots, y_m\}$  and a probability mass function  $p_Y(y_j) = \Pr\{Y = y_j\}$  ( $j = 1, \dots, m$ ).

#### 2.1.1 Entropy

The most fundamental quantity 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:

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

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 measurement 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 realization 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  $-\log_2 p(X)$ :

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

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  $p = 0.5$ .

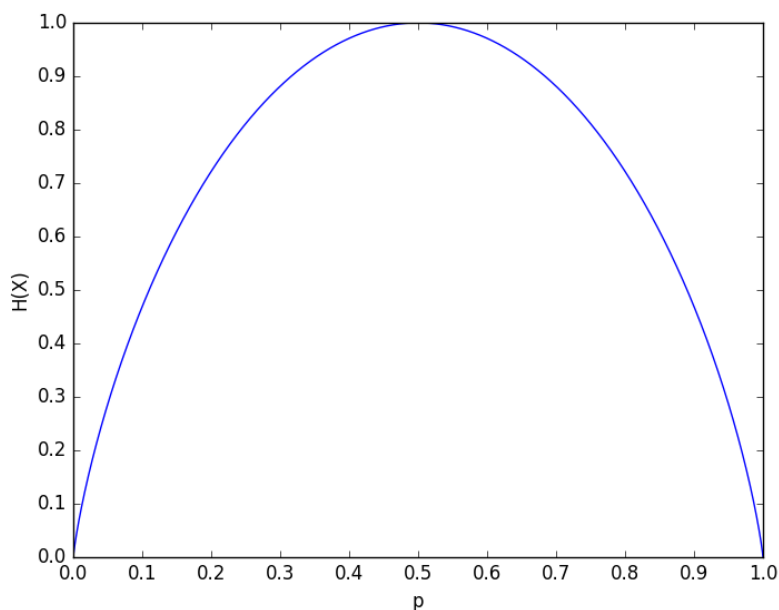


Figure 1: Entropy of  $X$  plotted against the value of  $p$ .

### 2.1.2 Joint and Conditional Entropy

Let the joint distribution of the random variables  $X$  and  $Y$  be  $p(x_i, y_j) = \Pr\{X = x_i, Y = y_j\}$  ( $i = 1, \dots, n; j = 1, \dots, m$ ). The *joint entropy* [CT06] of the pair  $(X, Y)$  is defined as

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

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

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

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  $H(Y) = H(Y|X)$ .

### 2.1.3 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

$$D(p||q) = \sum_{i=1}^n p(x_i) \log_2 \frac{p_X(x_i)}{q_X(x_i)} \quad (4)$$

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 logarithm in equation 4 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.

### 2.1.4 Mutual information

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

$$MI(X; Y) = \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p_X(x_i)p_Y(y_j)} \quad (5)$$

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 5.

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, which relates mutual information directly to entropies, is of importance:

$$MI(X; Y) = H(X) - H(X|Y)$$

### 2.1.5 Conditional mutual information

The *conditional mutual information* [CT06] of the discrete random variables  $X$  and  $Y$  given  $Z$  is defined by

$$MI(X; Y|Z) = H(X|Z) - H(X|Y, Z) = \mathbb{E}_p(x, y, z) \log_2 \frac{p(X, Y|Z)}{p(X|Z)p(Y|Z)}.$$

## 2.2 Partial information decomposition

## 2.3 Numerical estimator



## **3 Elementary cellular automata**

### **3.1 Problem description**

### **3.2 Related work**

### **3.3 Experimental setup**

### **3.4 Results**

### **3.5 Discussion**

## 4 Ising model

### 4.1 Problem description

### 4.2 Related work

### 4.3 Experimental setup

### 4.4 Results

### 4.5 Discussion

## 5 Neural networks

### 5.1 Problem description

### 5.2 Related work

### 5.3 Experimental setup

### 5.4 Results

### 5.5 Discussion

## 6 Conclusion

## References

- [CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.
- [Sha48] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

**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