Bachelor Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Entropy maximization under entropic constraints

Anna Ibatullina

Supervisor: doc. Ing. Tomáš Kroupa, Ph.D.
Field of study: Open Informatics
Subfield: Software
May 2023

BACHELOR‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name:	**Ibatullina Anna**	Personal ID number:	**498938**
Faculty / Institute:	**Faculty of Electrical Engineering**		
Department / Institute:	**Department of Computer Science**		
Study program:	**Open Informatics**		
Specialisation:	**Software**		

II. Bachelor’s thesis details

Bachelor’s thesis title in English:

Entropy maximization under entropic constraints

Bachelor’s thesis title in Czech:

Maximalizace entropie za entropických omezení

Guidelines:

Entropy maximization under moment or marginal constraints is a ubiquitous convex optimization problem for which many solution techniques exist [1]. However, there are variants of the entropy problem in which the powerful algorithms of numerical optimization cannot be used due to the lack of convexity in constraints. For example, measuring higher-order interactions in complex stochastic systems is based on the quantity called connected information [2], which amounts to solving the entropy maximization problem with constraints on the values of lower-dimensional entropies. This is a difficult non-convex optimization problem for which standard optimization methods fail.

The goal of this bachelor thesis are as follows.

1. To formulate the entropy problem under entropic constraints, with regard to the computation of connected information, and assess the usability of available numerical methods.
2. To approximate this problem using the techniques of information theory [3]. In particular, approximate the original problem as the problem with entropic variables and linear constraints expressing the properties of entropy vectors.
3. To implement an LP-based solver in Julia language for smaller instances of the problem approximation from point 2., and to evaluate the scalability of this technique with respect to the number of variables and dimensions of the state spaces.
4. To verify the quality of the approximation on the real data coming from the measurements of physical systems [2].

Bibliography / sources:

[1] Fang, S.-C., Rajasekera, J. R. & Tsao, H.-S. J. Entropy Optimization and Mathematical Programming. (Springer US). (1997)

[2] Martin, E. A., Hlinka, J. & Davidsen, J. Pairwise network information and nonlinear correlations. Phys Rev E 94, 040301 (2016)

[3] Raymond W. Yeung. Information Theory and Network Coding, Springer (2008)

Name and workplace of bachelor’s thesis supervisor:

doc. Ing. Tomáš Kroupa, Ph.D. Artificial Intelligence Center FEE

Name and workplace of second bachelor’s thesis supervisor or consultant:

Date of bachelor’s thesis assignment: **23.02.2023** Deadline for bachelor thesis submission: **26.05.2023**

Assignment valid until: **16.02.2025**

doc. Ing. Tomáš Kroupa, Ph.D. Supervisor’s signature	Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the bachelor’s thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the bachelor’s thesis, the author must state the names of consultants and include a list of references.

.

Date of assignment receipt	Student’s signature

Acknowledgements

I would like to thank doc. Ing. Tomáš Kroupa, Ph.D. for patiently supervising work on this thesis through the year. I also want to thank my family, all my friends and my boyfriend who were motivating, encouraging, and supporting me.

Declaration

I declare that the presented work was developed independently and I have cited all sources I have used in the bibliography.

Prague, May 23, 2023

Prohlašuji, že jsem předloženou práci vypracovala samostatně, a že jsem uvedla veškerou použitou literaturu.

V Praze, 23. May 2023

Abstract

This thesis is devoted to the maximization of entropy under entropic constraints.
First, the objectives and motivation behind will be discussed. Then we will focus on parts of information theory essential for the project purposes.
After, we will delve into the analysis of the problem posed and the proposed method of its approximation along with the theory behind it. Besides, we will also inspect the possibility and expediency of the stated problem relaxation, as it might be a very powerful technique for the elimination of non-convex constraints.
We will demonstrate the implementation part afterward, as well as the testing approach, and examine the resulting approximation quality using real data measurements.

Keywords: Connected information, Entropy maximization, Stochastic systems, Numerical optimization

Supervisor: doc. Ing. Tomáš Kroupa, Ph.D.
Artificial Intelligence Center, FEE

Abstrakt

Tato práce se věnuje maximalizaci entropie za entropických omezení.
Nejprve budou diskutovány cíle a motivace. Poté se zaměříme na části teorie informace podstatné pro účely projektu.
Dále se ponoříme do analýzy zadaného problému a navržené metody jeho aproximace spolu s teorií, která za ní stojí. Kromě toho také prověříme možnost a účelnost uvedené relaxace problému, neboť by se mohlo jednat o velmi účinnou techniku pro eliminaci nekonvexních omezení.
Následně předvedeme implementační část i testovací přístup a prověříme výslednou kvalitu aproximace pomocí měření reálných dat.

Klíčová slova: Spojená informace, Maximalizace entropií, Stochastické systémy, Numerická optimalizace

Překlad názvu: Maximalizace entropie za entropických omezení

Contents

Figures

Tables

Chapter 1

Introduction

Nowadays we have access to massive amounts of data. As an example, it is possible to get measurements from the human body through sensors, from very advanced and precise to very simple, accessible to the general public.
A lot of network transactions might also be captured and used as throughput or latency measurements. Data can be in diverse forms and come from different sources.
In all fields, it is very important to have a considerable amount of data. Data leads to research, and with the right research, helpful development follows. However, the provided data needs to be correct and relevant in order to not be detrimental to the final result. That's why it is extremely important to analyze data and its structure, to measure and understand the dependencies between its pieces in order to unlock its full potential.

One of the important data methods is the maximum entropy approximation, which is highly instrumental in the analysis of stochastic systems.
The principle of maximum entropy [7] states that the most probable state of the system is the one that retains the most uncertainty, thus, as we discuss further in this thesis, it is the state with the maximum possible entropy. It is widely used in a vast amount of disciplines [3] such as statistics, data compression, communication, machine learning [8], image processing [5] and many more.

In the next chapter, we will study basic measures from information theory. Information theory is a branch of mathematics and computer science that deals with the transmission, processing, and storage of information and data. It is a fundamental field in many areas of applied mathematics and electrical

engineering.
In information theory, entropy is an essential notion that measures randomness within a dataset: higher entropy means the data is more uncertain or unpredictable. Entropy can also be subject to certain constraints, in order to limit and direct these predictions towards ways that are useful and relevant, depending on the scope.

In this thesis, the reason for calculating entropy approximation is to obtain connected information – another information-theoretic quantity that nowadays could be used to analyze the structure and dynamics of large-scale systems and was designed to measure higher-order interactions in stochastic systems.

Computing connected information in large and complex systems can be a challenging task: it often involves dealing with massive and complex datasets, which can be computationally intensive and require powerful hardware and sophisticated algorithms. In this thesis, we will delve into the problem and implementation of connected information computation.

1.1 Motivation

We will look for computationally efficient algorithms for solving the problem of connected information calculation and will further study approximation methods based on advanced techniques of information theory.

1.2 Objectives

The first objective of this thesis is to study computational approaches.
The second is implementing the module in Julia language with the chosen approximation algorithm. Furthermore, we will compare the approximations obtained using the developed module to the results from the “Network inference and maximum entropy estimation on information diagrams” [9] paper.

Chapter 2

Entropy and Connected Information

Information theory is a branch of applied mathematics and electrical engineering that deals with the representation, transmission, and manipulation of information. It was developed in 1948 by Claude Shannon in his seminal paper "A Mathematical Theory of Communication" [10] published in the Bell System Technical Journal. Shannon's work laid the foundation for the field of information theory by defining the concepts of entropy and mutual information and providing a mathematical framework for understanding the limits of data compression and transmission over noisy channels.

Information theory has many important applications in a wide range of fields, including communication systems, data compression, cryptography, artificial intelligence, and statistics.
In this chapter, we will briefly cover the very basics of this field. The definition of Shannon's measures, information diagrams and entropic vectors in this chapter is based on the "Information Theory and Network Coding" book [12].

2.1 Shannon's Information Measures

Information measures form the foundation of information theory. We begin this chapter by introducing some of the essential Shannon's information measures important for this thesis's purposes.

In this thesis, we will employ the following notation. A discrete random vector $X = (X_1, \dots, X_n)$ has a joint probability distribution p and the sample space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$. Let $N = \{1, \dots, n\}$. For any joint probability distribution p of X and nonempty $A \subseteq N$, we define the random vector $X_A = (X_i)_{i \in A}$ with the sample space

$$\mathcal{X}_A = \bigtimes_{i \in A} \mathcal{X}_i.$$

Then, the marginal probability distribution of X_A is p_A such that

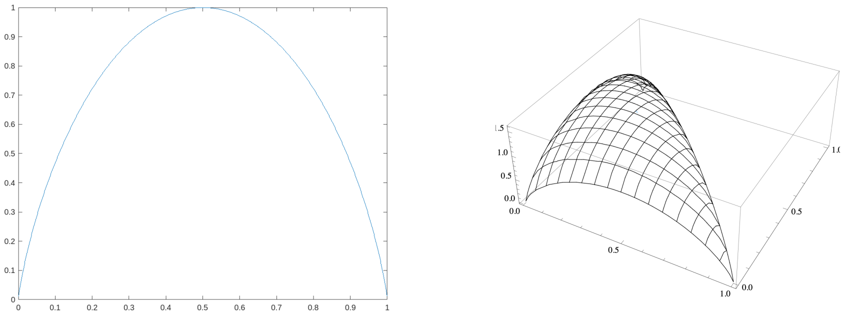
$$p_A(x) = \sum_{y \in \mathcal{X}_{\bar{A}}} p(x, y), \qquad x \in \mathcal{X}_A,$$

where $\bar{A} = N \backslash A$. In case of a few random variables composing the random vector, such as (X, Y, Z), we may occasionally use the alternative and more natural notation p_{XYZ} to denote the corresponding probability distribution.

Definition 2.1. The *entropy* of a discrete random variable X with a probability distribution p is the number

$$H(X) = -\sum_x p(x) \log p(x), \tag{2.1}$$

where the logarithm has base 2 and, as a convention, $0 \log 0 = 0$.



(a) : Graph for a binary variable.

(b) : Graph for a ternary variable.

Figure 2.1: Entropy graph

Definition 2.2. The *joint entropy* of two random variables X and Y with a joint probability distribution p is the number

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y). \tag{2.2}$$

Generalizing, the joint entropy of an n-dimensional random vector with a joint probability distribution p is

$$H(X_1, \dots, X_n) = -\sum_{x_1, \dots, x_n} p(x_1, \dots, x_n) \log p(x_1, \dots, x_n). \tag{2.3}$$

Entropy has always a non-negative value, which is easy to see from the definition. This property is also supported by the idea of entropy. The entropy of a random variable is the amount of information in the variable. Similarly, the joint entropy of n variables is the average amount of information gained from knowing the outcome of those variables.

We will also need another information measure for a better understanding of the problem.

Definition 2.3. The *mutual information* of two random variables X and Y is the number

$$I(X;Y) = \sum_{x,y} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x) \cdot p_Y(y)}, \tag{2.4}$$

where the sum above goes over x, y with $p_{XY}(x,y) > 0$.

To define mutual information for more than two variables we should also familiarize ourselves with conditional mutual information.

Definition 2.4. For three random variables X, Y, and Z, the *conditional mutual information* between X and Y conditioned on Z is the number

$$I(X;Y|Z) = \sum_{x,y,z} p_{XYZ}(x,y,z) \log \frac{p_Z(z) p_{XYZ}(x,y,z)}{p_{XZ}(x,z) \cdot p_{YZ}(y,z)}, \tag{2.5}$$

where the sum above goes over x, y, z with $p_{XYZ}(x,y,z) > 0$.

Then we can inductively define the mutual information for more than two variables as follows:

Definition 2.5. The *mutual information* of $n+1$ random variables is

$$I(X_1; \ldots ; X_{n+1}) = I(X_1; \ldots ; X_n) - I(X_1; \ldots ; X_n | X_{n+1}). \tag{2.6}$$

Moreover, the mutual information can be expressed in terms of entropy as well:

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{2.7}$$

Introduced information measures could be pictured as it is shown on a diagram 2.2 below.

Further details and reasoning behind diagram construction can be found in Raymond W. Yeung's book [12].

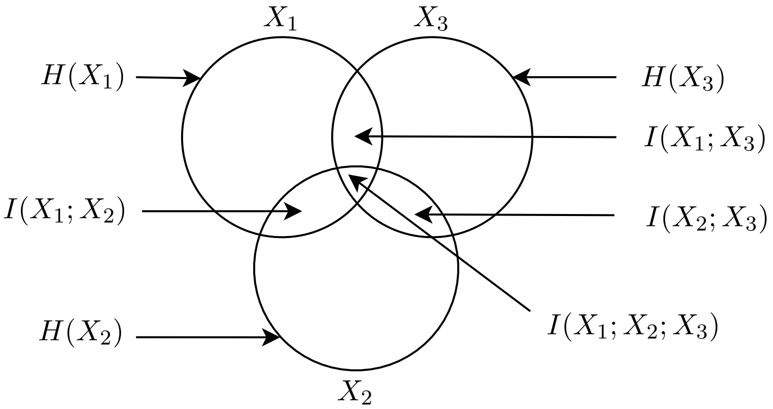


Figure 2.2: Entropy and mutual information for three variables.

While entropy can describe the amount of information within one random variable, mutual information measures the amount of information that one random variable contains about another, in other words, how dependent on each other variables are.
Mutual information, unlike entropy, can have a negative value. This can be easily shown on an example with 3 random variables, where one variable is dependent on two others.

Example 2.6. Let X_1, X_2, X_3 be random binary variables.
Define $X_3 = X_1 \oplus X_2$, where $\oplus$ is the XOR function, and every possible outcome has the same chance of happening, in other words, all possible events have the same probability. The joint distribution of those variables is presented in Table 2.1.

X_1	X_2	X_3	$p(x_1, x_2, x_3)$
0	0	0	1/4
0	1	1	1/4
1	0	1	1/4
1	1	0	1/4

Table 2.1: The joint distribution of X_1, X_2, X_3, where $X_3 = X_1 \oplus X_2$.

From the entropy definition and simple calculations, we can verify the equations:

$$
\begin{aligned}
H(X_i) &= 1 \qquad \forall i \in \{1,2,3\} && (2.8)\\
H(X_i, X_j) &= 2 \qquad \forall i,j \in \{1,2,3\}, i \neq j && (2.9)\\
H(X_1, X_2, X_3) &= 2 && (2.10)\\
I(X_i; X_j) &= 0 \qquad \forall i,j \in \{1,2,3\}, i \neq j && (2.11)\\
I(X_1; X_2; X_3) &= I(X_1; X_2) - I(X_1; X_2|X_3) - 1 && (2.12)\\
I(X_1; X_2|X_3) &= 1 && (2.13)
\end{aligned}
$$

As we can conclude from the information diagram and example, the more independent the random variables, the bigger is the joint entropy value and the smaller is the mutual information value. Mutual information equals zero when no information is gained within interactions, namely when there is no dependency between all variables under consideration. Similarly, negative mutual information indicates redundant interactions, where knowing a subset of variables reduces the information gained from knowing the others.

It also could be seen from equations (2.11) and (2.13) that conditioning on a variable can introduce a dependency. And, as described above, with appearing dependency the mutual information has also a bigger value.

2.2 Entropic vectors and submodular functions

In order to characterize the properties of the problem stated in the next chapters, we will present another possible representation of entropy.

Definition 2.7. Let p be a joint probability distribution of a random vector $X = (X_1, \ldots, X_n)$. The *entropic vector* is a vector h_p with coordinates

$$h_p(A) = H(X_A), \tag{2.14}$$

for all nonempty $A \subseteq N$, and $h_p(\emptyset) = 0$. *Entropy region* Γ_n^* is the set of all entropic vectors, that is,

$$\Gamma_n^* = \{h_p \mid p \text{ is a probability distribution of } (X_1, \ldots, X_n) \text{ on some sample space}\}. \tag{2.15}$$

To describe some of the characteristics of the entropic vectors, we will introduce the following properties. We will denote the set of all subsets as $\mathcal{P}(N)$.

A function $h\colon \mathcal{P}(N) \to \mathbb{R}$ is called

- *grounded* if $h(\varnothing) = 0$,
- *monotone* if $h(A) \leqslant h(B)$ for all $A, B \subseteq N$ with $A \subseteq B$,
- *submodular* if $h(A \cup B) + h(A \cap B) \leqslant h(A) + h(B)$, for all $A, B \subseteq N$.

Definition 2.8. A *polymatroid* $h\colon \mathcal{P}(N) \to \mathbb{R}$ is a grounded, monotone and submodular function.

Let Γ_n be the set of all polymatroids on $\mathcal{P}(N)$. Note that every $h \in \Gamma_n$ is *nonnegative*, that is, $h(A) \geqslant 0$ for all $A \subseteq N$.
Moreover, for every $\alpha_1, \alpha_2 \geqslant 0$ and all $h_1, h_2 \in \Gamma_n$, we obtain $\alpha_1 h_1 + \alpha_2 h_2 \in \Gamma_n$. In other words, the set $\Gamma_n \subseteq \mathbb{R}^{\mathcal{P}(N)}$ is a convex cone. Furthermore, Γ_n is polyhedral since it is defined by finitely-many linear inequalities.

Since every $h \in \Gamma_n$ is a nonnegative function, Γ_n does not contain a nontrivial linear subspace. Therefore Γ_n is said to be a *pointed* convex cone. We call Γ_n a *polymatroid cone (of order n)*.

It can be shown that $n + 2^{n-2}\binom{n}{2}$ monotonicity and submodular inequalities are enough to characterize Γ_n. Specifically, those are minimal submodular inequalities, for all $i, j \in N$, $i \neq j$, and all $A \subseteq N \backslash ij$,

$$h(A \cup i) + h(A \cup j) \geqslant h(A \cup ij) + h(A), \tag{2.16}$$

and the following monotonicity inequalities for all $i \in N$:

$$h(N) \geqslant h(N \backslash i). \tag{2.17}$$

A *Shannon-type inequality* is any inequality which is a nonnegative linear combination of inequalities (2.16)–(2.17). Any Shannon inequality of the form (2.16)–(2.17) is called *elemental* (or *minimal*).

The basic examples of vectors belonging to Γ_n are entropic vectors.

Proposition 2.9. $\Gamma_n^* \subseteq \Gamma_n$ *for every* $n \geqslant 1$.

This proposition was proved by Zhang and Yeung in the “On characterization of entropy function via information inequalities” paper [13].

2.3 Connected Information

We will be dealing with the connected information measure based on entropic constraints, which was investigated in “Network Inference and Maximum Entropy Estimation on Information Diagrams” paper [9]. We will occasionally write $H(p)$ in place of $H(X)$, where p is a probability distribution of X.

Definition 2.10. Let p be a joint probability distribution of a random vector $X = (X_1, \dots, X_n)$. The *connected information* of order $k = 2, \dots, n$ is the number

$$I^k(p) = H(q^{k-1}) - H(q^k), \tag{2.18}$$

where q^k is the maximum entropy probability distribution of $X = (X_1, \dots, X_n)$ among those consistent with all the one through k-variate marginal entropies.

It follows from the properties of maximization that

$$H(q^{k-1}) \geqslant H(q^k). \tag{2.19}$$

Hence, the connected information of any order can have only a non-negative value.

As we can see from the definition, the connected information is derived from the entropies of marginal probability distributions of order k with possibly different k-variate entropies. Applying the idea of entropy, we can conclude that connected information of order k represents the amount of information contained in k-dimensional dependencies. In other words, how much information about k random variables is already obtained knowing the results for $k-1$ variables.

Example 2.11. Let’s consider the same probability distribution as in 2.6 example corresponding to $X_3 = X_1 \oplus X_2$ and calculate connected information for it.

First, we will find $H(q^1)$. Namely, the maximum possible entropy distribution for random vector $X = (X_1, X_2, X_3)$ that satisfies $H(X_i) = 1$. As it follows from entropy properties, the maximum entropy for random variables can be achieved in case they are independent of each other. This state can be illustrated with the information diagram 2.3.

Assuming that, the entropy of random vector X simply equals to the sum of entropies of individual variables, $H(X) = H(X_1) + H(X_2) + H(X_3) = 3$.

To find $H(q^2)$, the probability distribution of random vector X should

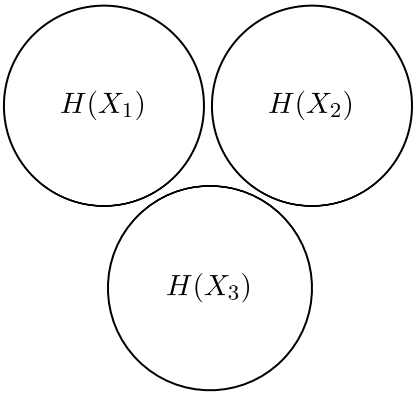


Figure 2.3: Information diagram for three independent random variables.

satisfy not only the equalities with entropies of order 1, but also constraints $H(X_i, X_j) = 2$, where $i \neq j$. We can notice that when random variables are independent those equalities hold. Therefore, $H(q^2) = 3$.

Finally, the only possible value for $H(q^3)$ is 2.

Having all maximum entropies we can now calculate connected information from the definition:

$$I^2(p) = H(q^1) - H(q^2) = 0, \tag{2.20}$$

$$I^3(p) = H(q^2) - H(q^3) = 1. \tag{2.21}$$

Chapter 3

Connected Information approximation

The connected information could be defined differently, respecting other types of constraints for entropy maximization. The problem of calculating it, therefore, can differ based on the definition of connected information. In this chapter, we will formulate an optimization problem for the definition from the previous section and one similar to it to compare and understand the key distinctions. Then we will take a look at possible constraints' relaxations and the reasons those weren't used. Finally, we will discuss chosen linear programming solution to approximate maximum entropy.

3.1 Problem with marginal constraints

First, we will take a look at an alternative definition of connected information.

Definition 3.1. Let p be a joint probability distribution of a random vector $X = (X_1, \ldots, X_n)$. The *connected information* of order $k = 2, \ldots, n$ is the number

$$\hat{I}^k(p) = H(\hat{q}^{k-1}) - H(\hat{q}^k), \tag{3.1}$$

where $\hat{q}^k$ is the maximum entropy probability distribution of $X = (X_1, \ldots, X_n)$ among those consistent with all the one through k-dimensional marginals of the probability distribution p.

Let $\mathcal{P}_k(N)$ be the set of all subsets of N with cardinality smaller than k.

The problem of finding $H(\hat{q}^k)$ could be then stated as follows:

$$\text{Maximize } H(q) \text{ subject to } q \in \hat{\Pi}_p^k, \tag{3.2}$$

where $\hat{\Pi}_p^k = \{q \mid q_A = p_A \text{ for all } A \subseteq \mathcal{P}_k(N)\}$.

Specifically, the nonlinear programming formulation of this problem is given by:

$$\begin{aligned} &\text{Maximize} && H(q) && (3.3)\\ &\text{subject to} && q_A = p_A, \ \ \forall A \subseteq \mathcal{P}_k(N). && (3.4) \end{aligned}$$

The difference between 3.1 and 2.10 definitions is in the feasible set formed by the constraints on probability distributions.

In particular, the problem with marginal constraints is considered to be easier to solve because it is essentially a convex optimization problem in which the constraints are linear equations.

This is a ubiquitous convex optimization problem for which many solution techniques exist [4]. The problem can be solved by standard numerical techniques for convex optimization problems as well, for instance, by Newton's methods. For more details about this method and others see the book "Numerical algorithms: methods for computer vision, machine learning, and graphics" [11].

3.2 Problem with entropy constraints

Returning to the first definition of connected information 2.10, we will define accordingly the problem of finding corresponding maximum entropy approximations.

We will consider the set Π_p^k of all probability distributions q which have the entropies of at most k-dimensional marginals of p,

$$\Pi_p^k = \{q \mid H(q_A) = H(p_A) \text{ for all } A \subseteq \mathcal{P}_k(N)\}. \tag{3.5}$$

Then the problem of entropy maximization under the entropy constraints could be defined as

$$\text{Maximize } H(X_q) \text{ subject to } q \in \Pi_p^k. \tag{3.6}$$

We will refer to this problem as the *primal problem* in this thesis.

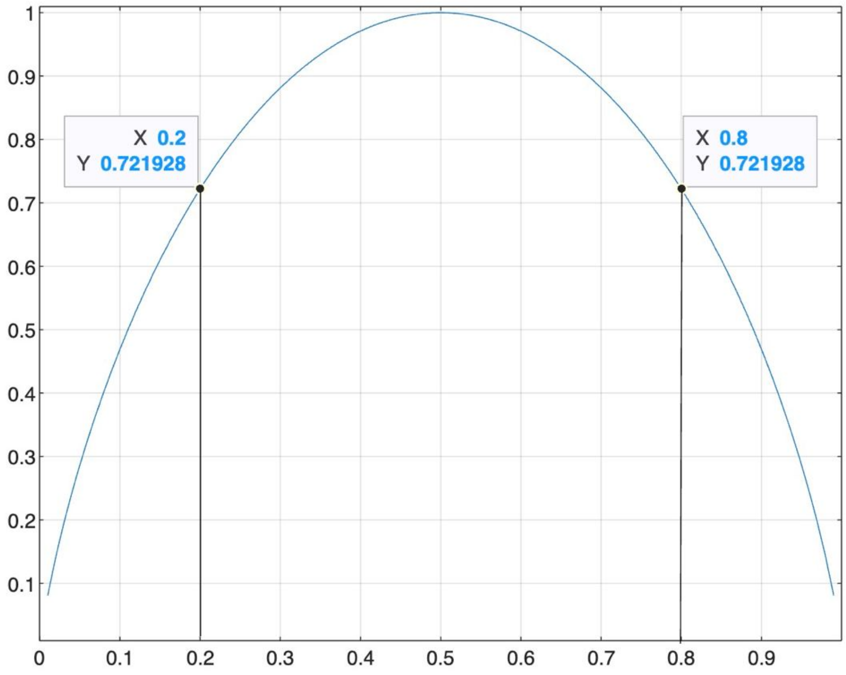


Figure 3.1: Feasible set of the primal problem for a binary variable.

The objective function is the same entropy function as the one from the previous section 3.1. However, the constraints contain entropy functions and couldn't be described as linear. Furthermore, those constraints form a non-convex feasible set and, thereby, increase the computation difficulty.

Moreover, we can notice that $\hat{\Pi}_p^k \subseteq \Pi_p^k$, as the equality of distribution marginals imply the equality of corresponding entropies.

3.2.1 Relaxation of primal problem constraints

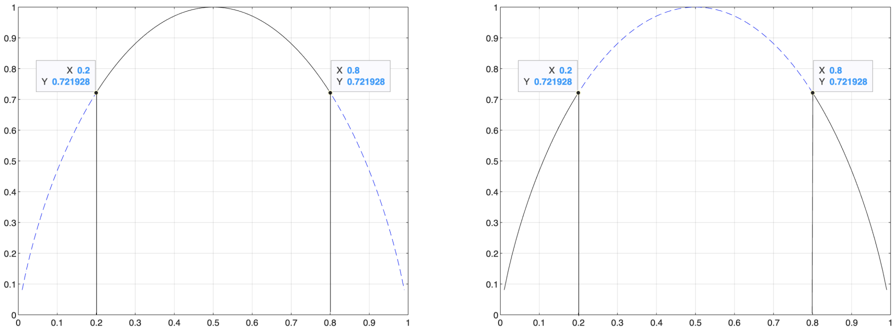
The primal problem is considered difficult mainly because equality constraints form a non-convex set. In some cases, the relaxation of equality constraints to inequality could make the feasible set convex and therefore tremendously alleviate the problem.

To begin with, we will try to transform equality constraints from primal problem 3.5 $H(q_A) = H(p_A)$ to inequality constraints $H(q_A) \geqslant H(p_A)$.

As a result, this relaxation leads to an optimization problem with a convex feasible set, but the optimal value of the resulting problem will be different for most of the cases. Maximum possible entropy, the one that could be acquired from independent random variables, always satisfies the constraints and always is in a feasible set. Therefore this relaxation can not be used for reducing computational complexity.

Another possibility for relaxation would be to transform constraints (3.5) to $H(q_A) \leqslant H(p_A)$ instead. However, those constraints won't simplify computation, because they form a non-convex feasible set as well.

Both constraint relaxations are presented for a random binary variable on the graphs below (3.2).



(a) : Relaxation to $H(q_A) \geqslant H(p_A)$.

(b) : Relaxation to $H(q_A) \leqslant H(p_A)$.

Figure 3.2: Feasible set after constraints relaxation.

3.3 Approximation

After exploring the approaches discussed in the previous section, we have concluded that they are not suitable for the solver implementation. Therefore, we propose another approach that shows itself more applicable to this problem.

It follows from the problem statement (3.5) that

$$\Pi_p^1 \supseteq \cdots \supseteq \Pi_p^n. \tag{3.7}$$

For each $k \in N$ the set Π_p^k is nonempty ($p \in \Pi_p^k$), however, Π_p^k is typically

not a convex set. Since most non-convex problems are extremely difficult to solve, we need to find computationally tractable approximations to (3.6). This is based on the observation that both the objective and the constraints of (3.6) depend only on the values of entropy. Therefore, we will use
the well-known properties of entropic vectors to transform the original problem (3.6) into the optimization problem involving only variables representing entropies.
The potential drawback is that there may be no probability distribution associated with optimal solutions to such a general problem. However,
the resulting optimization problem will be linear and it will provide the upper bound on the value of the optimal solution to (3.6).

Specifically, for each $k \in N$, we consider the linear programming problem with variables $h(A)$ for $A \subseteq N$, where the objective to maximize is the linear function representing a single variable $h(N)$, and where the constraints are

- $h \in \Gamma_n$, (entropic vector)
- $h(A) = h_p(A)$ for all $A \in \mathcal{P}_k(N)$, (constraints from (3.5))
- $h(A) \leqslant \log_2 |\mathcal{X}_A|$ for all $A \in \mathcal{P}(N) \backslash \mathcal{P}_k(N)$, (the upper bound on the entropies of marginals)
- Zhang-Yeung inequalities (in case $n \geqslant 4$).

Let $n \geqslant 4$. In this case, Zhang and Yeung showed that $\overline{\Gamma_n^*} \neq \Gamma_n$ by exhibiting a non-Shannon type inequality. Introduced in “On characterization of entropy function via information inequalities” paper [13], Zhang-Yeung inequality could be formulated as

$$\begin{aligned} &3[h(ik) + h(il) + h(kl)] + h(jk) + h(jl) \\ &\quad - h(i) - 2[h(k) + h(l)] - h(ij) - 4h(ikl) - h(jkl) \geqslant 0 \end{aligned} \tag{3.8}$$

where $i, j, k, l \in N$ are different elements. Note that there are six instances of this inequality for $n = 4$, because i index can be swapped with j, and k index can be swapped with l reciprocally.

The resulting linear program is always feasible and its optimal value provides an upper bound on the optimal value of (3.6).

Chapter 4

Implementation

In the previous chapters, we could see that the connected information can be easily calculated given the maximum entropies of the corresponding orders.

In this chapter, we will go through part of the implementation of this project, as well as tests, analyze the time complexity of the resulting program and evaluate performance on real data originating from “Network inference and maximum entropy estimation on information diagrams” [9] paper.

For the sake of brevity, some parts of the code will be omitted. Full implementation including tests, examples and documentation can be found in the GitHub repository [6].

4.1 Code and implementation details

This project is implemented in Julia 1.7. For solving the linear programming problems, JuMP package for Julia and MOSEK solver [2] were used. Those tools provide algorithms for solving LP problems, thus we only need to correctly set objective function, variables and constraints.

We will take a closer look at the function computing maximum of entropy.

```
function estimate_max_entropy(
    k::Int64,
    distr_cards::Vector{Int64},
    entropy_constraints::Dict{Vector{Int64}, Float64};
    lower_bound = false)::Float64
```

In this function definition, the arguments are:

- `k` – maximal possible cardinality of marginals of entropy from constraints,
- `distr_cards` – probability distributions cardinalities of corresponding random variables,
- `entropy_constraints` – entropy values,
- `lower_bound` – optional argument, if true, computes the lower bound of approximation, otherwise calculates the upper bound.

Returned float value is the estimated entropy for given inputs.

Following the approach from the previous chapter (3.3) we use entropy vectors as variables. Those will be refed as *generators* later.

First, we initialize the JuMP model that uses the MOSEK solver.

```
model = Model(
    optimizer_with_attributes(Mosek.Optimizer, "QUIET" => true))
```

Then, we define $2^n + 1$ variables for optimization, the same as the number of generators for n random variables. Based on the entropy definition, a lower bound of their values is set to 0.

```
# non-negativity constraints
# $h(A) \geqslant 0, \forall A \in \mathcal{P}(N)$
@variable(model, h[1:(2^distributions_n + 1)] >= 0)
```

After, we gradually introduce constraints described in the approximation section (3.3).

```
# monotonicity (part of $h \in \Gamma_n$)
# $h(N) \geqslant h(N \backslash i), \forall i \in N$
@constraint(model, h[subset_A] >= h[subset_B])
...
# submodularity (part of $h \in \Gamma_n$)
# $h(A) \leqslant h(B) + h(A \backslash B), \forall B \subseteq A$
@constraint(
    model,
    h[subset_intersect] + h[subset_union] <= h[subset_C] + h[subset_D])
...
# given entropy constraints
# $h(A) = h_p(A), \forall A \in \mathcal{P}_k(N)$
@constraint(model, h[subset_A] == entropy_constraints[subset_A])
...
# cardinality constraints
# $h(A) \leqslant \log_2 |\mathcal{X}_a|, \forall A \in \mathcal{P}(N) \backslash \mathcal{P}_k(N)$
@constraint(model, h[subset_A] <= log(2, cardinality))
...
if lower_bound
    # Ingleton inequality
    @constraint(
        model,
        h[it] + h[jt] + h[il] + h[tl] - h[ij] - h[t] - h[l]
        - h[itl] - h[jtl] >= 0)
else
    # Zhang-Yeung inequalities
    @constraint(
        model,
        3(h[it] + h[il] + h[tl]) + h[jt] + h[jl] - h[i]
        - 2(h[t] + h[l]) - h[ij] - 4h[itl] - h[jtl] >= 0)
end
```

Finally, we define the objective function for the model and solve the problem

```
# Maximize $H(X_1, X_2, \ldots, X_n)$
@objective(model, Max, h[subset_to_index[collect(1:distributions_n)]])
optimize!(model);
return objective_value(model)
```

The returned value is the lower or upper bound of entropy, depending on the passed `lower_bound` parameter.

The mentioned module additionally implements several other functions, however, most of them are based on the above-described function.

4.2 Time complexity

From the implementation point of view, the discussed function can be divided into two parts. The first one is setting the constraints for the model for the solver, the second is finding the solution for the model itself.

During the function execution constraints are separately set for each variable. There are the same amount of variables as the number of subsets we can obtain from generators. It equals 2^n, where n is the number of generators, or, similarly, the number of random variables. Setting constraints takes $O(2^(n-1))$, thus, overal, setting constraints takes $O(2^{n(n-1)})$.

As for the MOSEK solver, even though it is not possible to say the exact time complexity, it is, in practice, guaranteed [1] to be less than $200 * O(n^3)$.

Although time measurements on a random machine are not representative and time results shouldn't be used as an algorithm efficiency reference, nevertheless, we examine time on one particular machine to have a rough idea of calculation duration. Those experiments can be found and reproduced in the GitHub repository [6] `examples/time_performance.ipynb` Jupiter Notebook.

While the function time complexity itself depends only on the number of random variables, precomputation of data also rely on $|\mathcal{X}|$ and dimensions of a probability distribution, because we need to calculate entropic constraints for each subset of random variables. Precomputation time complexity in our implementation is $O(2^n * \prod_{i=1}^{n} |\mathcal{X}_i|)$.

4.3 Testing

The implementation approach's correctness was tested on smaller examples located in the same repository [6] as the main module in the `test` directory.

The first test is based on XOR example 2.6 described through the theory part of the thesis. The second example represents the randomly generated probability distribution for 3 random variables with cardinalities 2, 3 and 4 respectively.

These examples as well can be found in `exapmles/example.ipynb` Jupiter Notebook.

4.4 Quality of approximation on the real data

In order to evaluate the quality of the approximation, we used the same data as in “Network inference and maximum entropy estimation on information diagrams” [9] paper to compare the resulting approximations of connected information. The measured neuroimaging data represents resting-state human brain networks.

Since the provided data measurements are continuous, we first discretized them to 2, 3 and 4 levels (bins) the same way as it was done in the mentioned paper [9], that is, using equiquantal (equiprobable) binning. After that, it was needed to calculate all the possible entropies for discretized data to utilize as constraints for the model in the module. Finally, we run the approximation function with acquired data.

In the paper [9] the *total correlation* $I_N = \sum_i H(X_i) - H(X)$ was defined together with the ratios I^k/I_N, measuring this way the percentage of information contained in k-dimensional measurements.

Following their approximation approach it could be seen, that as optimization variables were used so-called *atoms* and subsequent optimization was conducted under non-negativity constraints for them. However, some information measures, specifically mutual information, represented by atoms can be negative. Therefore, the feasible set of this problem does not cover all possible entropies. The resulting entropy approximation is, thus,
the lower bound of the actual optimal value of maximum entropy and subsequent connected information has to be the upper bound of the optimal connected information value.

As opposed to the approach from paper [9], we used *generators* as variables for the optimization problem. Because generators represent entropy, which is always non-negative, the feasible set covers all possible entropies consistent with constraints. However, because we are modeling only properties of entropic vectors, there is no guarantee that probability distribution with such entropy would exist. That means that we are approximating the upper bound of entropy and, consequently, the lower bound of connected information. Therefore, our resulting numbers should have a lower value comparing to the approximations from the paper [9].

Discretization level	Our results			Observations from the paper
	I^2	I_N	I^2/I_N	I^2/I_N
2-level	1.3721	2.1006	0.6532	0.93
3-level	2.1332	3.7126	0.5745	1.00

Table 4.1: Comparison of approximations.

The results of our approximation turned out to have a smaller value than approximations from the paper [9], meaning that results are consistent.

The calculations of experiments for 2, 3 and 4-level discretized data, both regular and its surrogate, up to 10th order of connected information approximation, could be found in module repository [6] in `examples/DMN_data_results.ipynb` Jupiter Notebook.

Chapter 5

Conclusion

We approximated the stated problem of maximizing entropy under the entropic constraints using the information theory technics described in the theory chapter. The obtained formulation has entropic variables and linear constraints based on the entropy vector's properties.

As an implementation output of this thesis, we got an LP-based solver in Julia language. After implementing it for smaller instances of the problem, we managed to scale it, making the solver generic in terms of the input size of parameters. Although, as it was assessed later, the time complexity of the solver heavily depends on the number of distributions and precomputation time complexity, on top of it, also depends on the size of the sample space of given distributions. Therefore it is still recommended to choose a number of random variables and their cardinality wisely, not to drastically increase the solver runtime. Besides, in the same module were also implemented helper functions, such as the computation of all entropies for constraints or discretizing given data to the required level.

Further, evaluating the same data, we compared the results of our approximation and once presented in the mentioned paper [9]. Results turned out to be consistent because our resulting lower bound of connected information was indeed less than approximated upper bound from the paper [9].

Bibliography

[1] E. D. Andersen. Complexity of solving conic quadratic problems, 2013.

[2] E. D. Andersen and K. D. Andersen. The mosek interior point optimizer for linear programming: an implementation of the homogeneous algorithm. *High performance optimization*, pages 197–232, 2000.

[3] J. R. Banavar, A. Maritan, and I. Volkov. Applications of the principle of maximum entropy: from physics to ecology. *Journal of Physics: Condensed Matter*, 22(6):063101, 2010.

[4] S.-C. Fang, J. R. Rajasekera, and H.-S. J. Tsao. *Entropy optimization and mathematical programming*, volume 8. Springer Science & Business Media, 1997.

[5] S. F. Gull and J. Skilling. Maximum entropy method in image processing. In *Iee proceedings f (communications, radar and signal processing)*, volume 131, pages 646–659. IET, 1984.

[6] A. Ibatullina. Connectedinformation.jl repository. https://github.com/ann-ib/ConnectedInformation.jl, 2023.

[7] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[8] S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 832–838. IEEE, 2005.

[9] E. A. Martin, J. Hlinka, A. Meinke, F. Děchtěrenko, J. Tintěra, I. Oliver, and J. Davidsen. Network inference and maximum entropy estimation on information diagrams. *Scientific reports*, 7(1):1–15, 2017.

[10] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[11] J. Solomon. *Numerical algorithms: methods for computer vision, machine learning, and graphics.* CRC press, 2015.

[12] R. W. Yeung. *Information theory and network coding.* Springer Science & Business Media, 2008.

[13] Z. Zhang and R. W. Yeung. On characterization of entropy function via information inequalities. *IEEE Transactions on Information Theory*, 44(4):1440–1452, 1998.

Appendix A

Attachments

- ConnectedInformation.zip ...... Archive of the module repository[1]
- text .... Source code of the thesis
  - Thesis_Text.pdf .... Thesis in PDF format
  - Thesis_Assignment.pdf ...... Thesis assignment in PDF format

[1] https://github.com/ann-ib/ConnectedInformation.jl