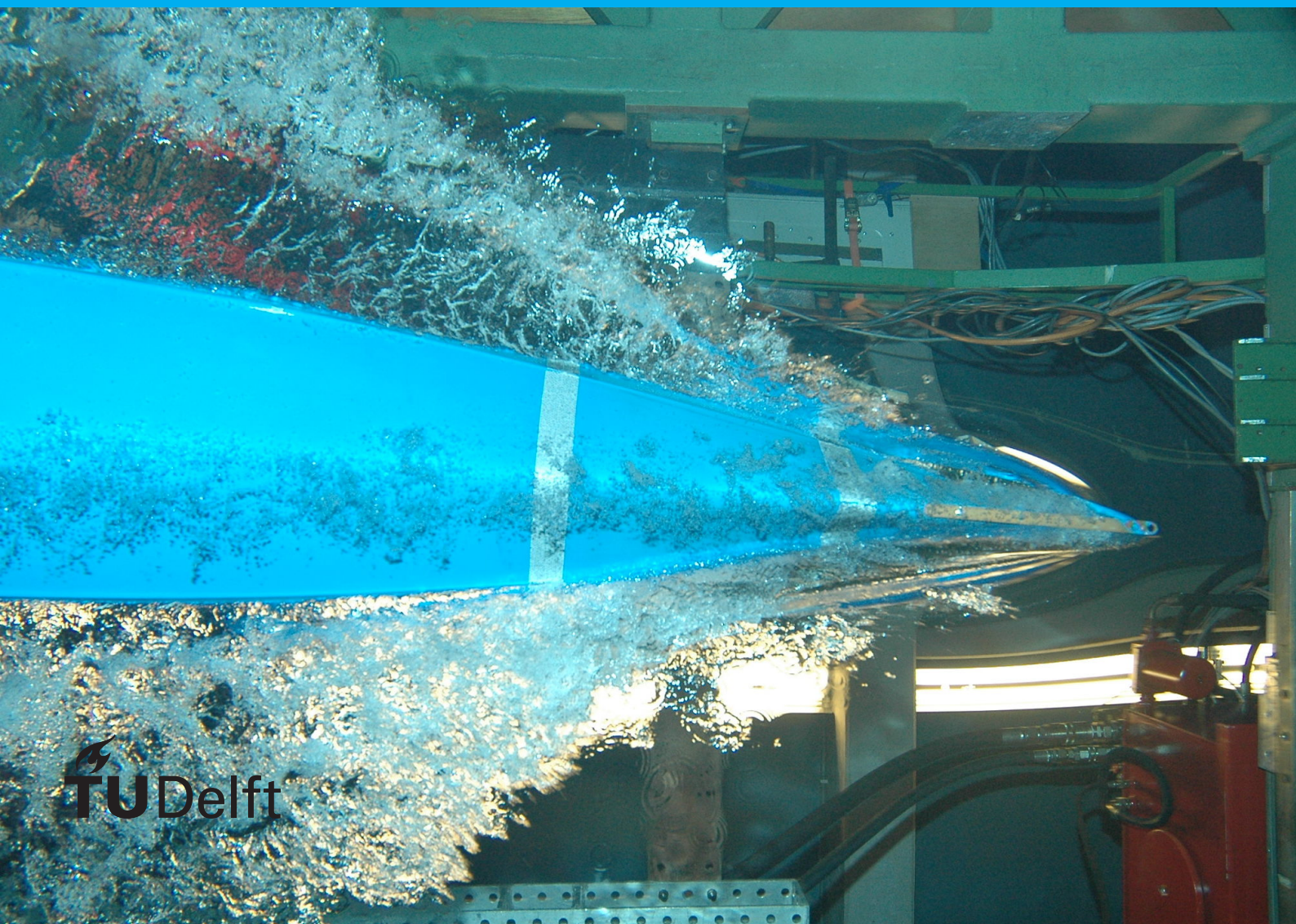
Monotonicity of Entropy

A Rigorous Proof of an Entropic Monotonicity Theorem

T.C.T. van Baar

Delft University of Technology



Monotonicity of Entropy

A Rigorous Proof of an Entropic Monotonicity Theorem

by

T.C.T. van Baar

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on Friday June 27, 2025 at 1:30 PM.

Student number:	5648319	
Project duration:	April 14, 2025 – June 27, 2025	
Thesis committee:	Prof. dr. M. P. T. Caspers,	TU Delft, supervisor
	dr. R. J. Fokkink,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Abstract

This thesis will be about explaining a proof of a theorem about entropy presented in a scientific article by Arstein, Barthe Ball and Naor [1] in detail. The original proof is complex, especially for bachelor-level students. The goal of this thesis is to break down that proof, add mathematics and make the ideas understandable for students at this level. The theorem is about mathematical entropy, a concept in probability theory, which is a measure for uncertainty or chaos in an event or outcome. For instance, consider a coin toss: the outcome is uncertain, and this uncertainty is measured by entropy. The greater the uncertainty, the higher the entropy. There is also a concept of entropy in the world of physics, where it is a measure for describing the uncertainty or randomness in which systems evolve. One of the goals of this theorem is its demonstration that these two types of entropy, though defined in different contexts, exhibit analogous behavior. In particular, it shows that the mathematical entropy behaves like the second law of thermodynamics, which states that entropy in an isolated system increases over time. As an illustrative example, when a glass of water is spilled on a table, the water gradually spreads out, increasing the disorder of the system. The theorem explains that if the entropy of a normalised sum of independent events is taken, then this entropy will increase with the amount of events that are summed. A normalized sum refers to the average obtained by summing independent events and dividing by their count. This seems logical at first, since the uncertainty of for example two separate events seems bigger than that of one event. However, the proof is complex and requires advanced analysis to prove.

The theorem is thus about the monotonicity of entropy of normalised sums. This thesis will connect the entropy to Fisher information, which enjoys nicer analytical properties to use. The concept is clear, entropy is a measure for uncertainty, while Fisher information quantifies the amount of information a random variable carries. This thesis will thus first connect the Fisher information and entropy and show that proving an increase of entropy can be done by proving a decrease in Fisher information. The proof of the decrease in Fisher information will need another theorem. This theorem is about connecting the Fisher information to the world of analysis. This requires some advanced analysis, like Green-Gauss on infinite surfaces and integrating the divergence of functions over hyperplanes. With this connection to the world of analysis eventually the decrease in Fisher information can be reached with some help from a lemma about commuting orthogonal projections in Hilbert spaces. All supporting theorems and lemmas will also be proved in detail in this thesis. In conclusion, this thesis will show a lot about the behaviour of entropy under normalised summation of independent events and show that these monotonically increase, which implies that the mathematical form of entropy behaves like the second law of thermodynamics.

Preface

This thesis is written as final requirement for the degree of Bachelor of Science in Applied Mathematics at Delft University of Technology. This thesis finishes four years of work to obtain this degree. I want to thank Martijn Caspers for the opportunity to do this project and the productive meetings. The approach was always direct and helpful to get me over the roadblocks presented by this project.

Before I started this project I had never even heard of the concept of entropy, but in researching this concept it has started to fascinate me more and more. I hope that the reader also finds this concept as interesting as I did when writing this.

Furthermore I want to thank the authors of the article of which this thesis will be about, Artstein, Ball, Barthe and Naor [1]. I want to emphasise that all the mathematical findigs in this thesis are done in this article. I merely explained it on a level for bachelor students and added mathematics to explain their findings. Therefore I want to thank them for adding these interesting findings to the scientific world and this opportunity to do this project about this article.

T.C.T. van Baar
Delft, June 2025

Contents

1 Introduction

1.1. Background of entropy in physics and information theory

Entropy is a concept that is widely recognized in physics, where it is often associated with chaos or uncertainty within a system. First introduced in the 19th century, entropy has played a central role in thermodynamics. However, this thesis will not focus on the physical interpretation of entropy. Instead, it will explore entropy from the perspective of information theory—a more mathematical formulation of the concept. It will actually be discussed later that the main point of the article is to link the result about mathematical entropy to a result from thermodynamics, namely the second law of thermodynamics.

The mathematical definition of entropy was introduced by Claude Shannon in the 1940s, in a groundbreaking work that laid the foundation for information theory [6]. Some of Shannon's results in this work will actually be used later in this thesis. A well-known anecdote illustrates the link between physical and mathematical entropy: when Shannon was trying to decide on a name for his invention, he asked the physicist John von Neumann. Von Neumann told him [9]: "You should call it entropy, for two reasons: first, because it's already used in statistical mechanics; and second, because nobody really knows what entropy is, so in a debate you'll always have the advantage." This already gives a connection between the physical form of entropy and the mathematical form, and maybe more importantly that entropy is a difficult concept to grasp. It is hoped that at the end of reading this thesis the reader has a clearer understanding of entropy and is at an advantage when the term entropy is used in a discussion.

1.2. Introduction to mathematical entropy

As said, the mathematical form of entropy is linked a lot with the physical form, which was about a state of chaos or uncertainty. This gives also the core idea of information theory, which looks at the uncertainty of a message. If a lot about a message is already known, then the informational value of this message is very low; we already pretty much knew what was going to be sent before it was sent. The uncertainty or surprisal of the message was low. On the other hand if on the other hand there is very little known about a message, the informational value of this message is very high. The uncertainty or surprisal of the message was high.

To measure this surprisal of an event E, a function is needed that increases if the probability $p(E)$ of an event decreases. When $p(E)$ is close to 1, the surprisal of the event is very low and vice versa. If $p(E)$ is close to 0, the surprisal of the event is very high. If the surprisal of E is defined by $S(E)$, then the relationship can actually only be described by the following function

$$S(E) = \log\left(\frac{1}{p(E)}\right) = -\log(p(E)).$$

Shannon defined the entropy Ent(X) of a discrete random variable X as the expected value of the surprisal of all possible events of X. This can be written as

$$\text{Ent}(X) = \mathbb{E}[S(X)] = \mathbb{E}[-\log(p(X)] = -\sum_{x\in\chi} p(x)\log(p(x)) \tag{1.1}$$

Here χ are all possible events which can happen in the random variable and $p(x) = \mathbb{P}(X = x)$. Thus if we know a lot about a random variable, then the entropy of a random variable is low and vice versa if we know very little about a random variable, then the entropy of a random variable is very high. It can be seen that entropy is thus a measure of uncertainty or chaos in the random variable.

This is introduced with a simple example, namely the coin flip. Take a coin flip with chance p for heads and q for tails. Then the entropy will be the biggest if the surprisal of the events is the highest. It can be deduced without calculations that this happens when $p = q = \frac{1}{2}$, since then the surprisal of the outcome is the biggest. If $p = 1$, then the entropy is the smallest, since the outcome is already known. The graph of the entropy for different p can be seen below.

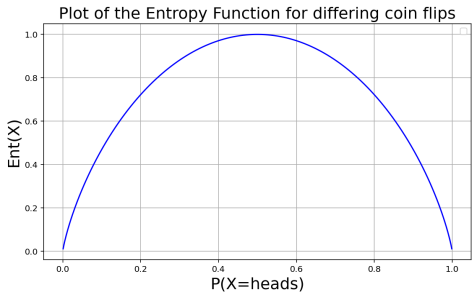


Figure 1.1: Binary entropy as a function of bias in a coin flip. From Entropy (information theory).

Figure 1.1 calculates the entropy with a base in the logarithm of 2. The base of the logarithm does not actually matter. All properties of entropy still hold for all bases. The choice of 2 comes from the fact that this concept was introduced in the information theory, which is interested in how many bits it for a computer takes to send a message. It is then logical to choose a base of 2. A bit can be seen as a yes-no question for a computer. If we look at the coin flip with $p = \frac{1}{2}$, it is clear that we need one yes-no question, so one bit. If $p = 1$ no yes-no questions are needed as can be also seen from Figure 1.1. This introduces the concept of which entropy was built on, the entropy of a random variable is equal to the least amount of bits it takes to determine the outcome of a random variable.

This is further explained with another example. Imagine a letter is sent from computer a to computer b and there are 4 possible letters, $\{A, B, C, D\}$, which can be sent. If all letters have the same probability to be sent, $p_i = \frac{1}{4}$, then it takes two yes-no questions for computer b to determine which letter was sent, as can be seen in the figure below. This is equal to 2 bits. Thus this can be seen as a random variable with entropy equal to 2.

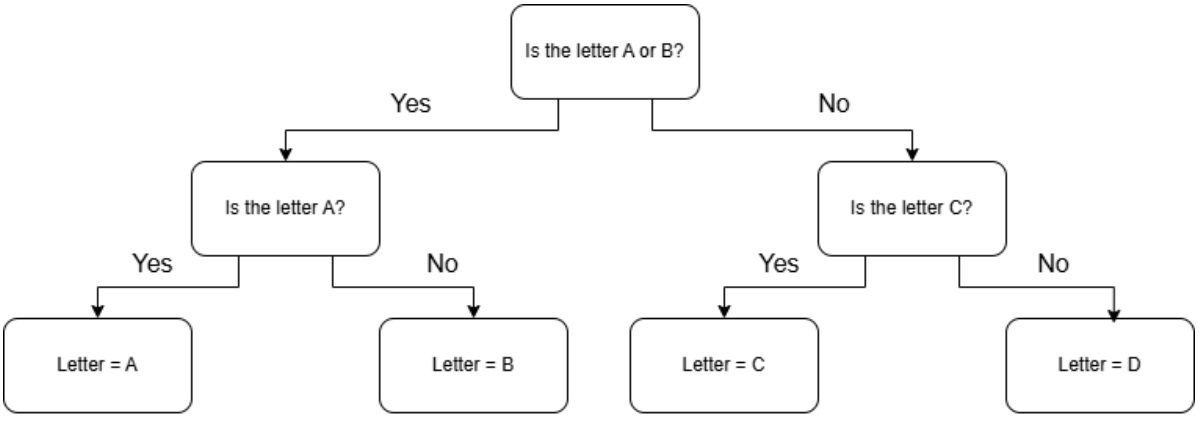


Figure 1.2: Yes-no flowchart in the case with equal probabilities).

Now take the same random variable, but now assume we know more about the probabilities. For example, $P(X = A) = \frac{1}{2}, P(X = B) = \frac{1}{4}, P(X = C) = P(X = D) = \frac{1}{8}$. Then it is logical to think the uncertainty of this random variable is less and thus the entropy will also be less. We show this in the next figure in which we

show that actually on average we only now need 1.75 yes-no questions, or equivalently 1.75 bits. The reader can easily calculate the entropy with the formula introduced in Equation (1.1) and see that the entropy is also equal to respectively 2 and 1.75.

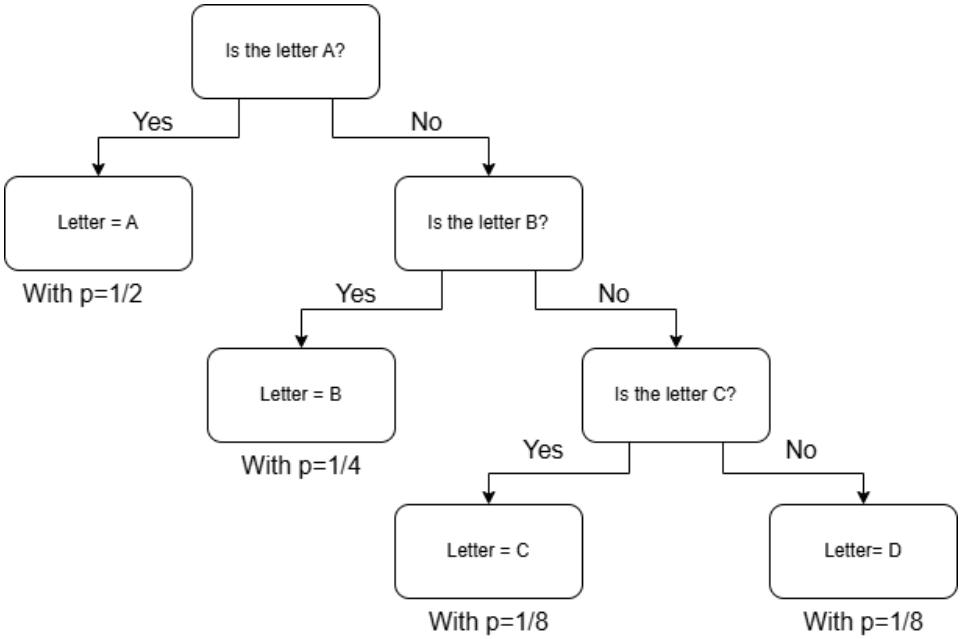


Figure 1.3: Yes-no flowchart in the case with different probabilities).

This is how the concept of entropy was introduced by Shannon, but the main point of thesis will be about entropy of continuous random variables. This will be explained in the next section, but these concepts about uncertainty and informational value will be the same. It is only not possible to then see it in bits, since it is not discrete anymore. It is not really important, but the base of the logarithm further used will be the normal e.

1.3. Entropy of a continuous random variable

Shannon did actually also introduce the form of entropy for continuous random variables in his first work. He did this by once again taking the expected value of the surprisal, but then for a continuous random variable X. Then the variable does not have a clear $p(X = x)$, but a density function $f : \mathbb{R} \to [0, \infty)$. He defined the entropy then as

$$\text{Ent}(X) = \mathbb{E}[-\log(f)] = -\int_{\mathbb{R}} f \log(f). \tag{1.2}$$

Shannon actually did not derive this but thought the continuous case was analogue to the discrete case. This is actually not the case, since taking the logarithm of a density is not clearly defined. A density can have a dimension and a logarithm is normally only defined for dimensionless functions. This is why this form of entropy is called differential entropy. The actual form for entropy for continuous random variables is the limiting density of discrete points. Differential entropy is a limiting case of the limiting density of discrete points and loses some of its association with the discrete entropy. For example, the differential entropy can be negative, which is impossible in the discrete case (you can never send a message with negative bits). Furthermore, since a logarithm is not well-defined for densities with a dimension, it is not invariant under scaling.

These properties mean that the differential entropy is not really useful in some cases. If the random variable has a dimensionless density for example, then the differential entropy is useful. It is not really important for the proof of the main theorem of the paper, but it is good to note that differential entropy only makes sense when the integral in Equation (1.2) is well-defined and makes sense. This is because the proof in this thesis will be about differential entropy.

1.4. Thesis outline

This bachelor thesis aims to explain the article by Artstein, Barthe, Ball, and Naor [1], which presents a proof of the monotonicity of entropy in the the theorems which will be introduced in Section 2.3. The article itself is quite dense and difficult for undergraduate students to follow. The goal of this thesis is to break down and clarify the proof, making it accessible to a bachelor-level mathematics student. The rest of this thesis will focus on carefully presenting and proving this theorem step by step.

2

Formulation of the Entropic Monotonicity Theorem

In this chapter we will introduce the main theorem to be proved in this bachelor thesis. These will be about the monotonicity of entropy in random variables. The rest of the chapters will be about developing a strategy to prove the theorems in this chapter.

2.1. Motivation behind the main theorem

Recall from Chapter 1 that we have defined the entropy of a random variable X with density $f : \mathbb{R} \to [0,\infty)$ as

$$\mathrm{Ent}(X) = -\int f \log(f) \tag{2.1}$$

provided that the integral makes sense. Now it is known that among random variables with variance 1 that the Standard Gaussian has the largest entropy. You can find a proof of that the Gaussian is actually the random variable with the largest entropy for a fixed variance in Soch [7], but we don't go into it any further here. Then, if X_i are independent copies of a random variable with expected value 0 and variance 1, then the normalized sums

$$Y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \tag{2.2}$$

approach the standard Gaussian as n tends to infinity. This is a result of the central limit theorem. Because as n increases, the normalised sum Y_n converges towards a Gaussian, which has the largest entropy, it makes sense to think that the entropy of the normalised sum also increases. The case where $n = 2$ was already proved by Shannon in the 1940's [6]. This result is enough to prove indeed that in big steps, in a sequence of powers of 2, the normalised sum's entropy keeps getting better, which we show by induction. We want to show that

$$\mathrm{Ent}(Y_{2^k}) \geq \mathrm{Ent}(Y_{2^{k-1}}). \tag{2.3}$$

For this we show that

$$Y_{2^k} = \frac{1}{\sqrt{2^k}} \sum_{i=1}^{2^k} X_i = \frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{2^{k-1}}} \sum_{i=1}^{2^{k-1}} X_i + \frac{1}{\sqrt{2^{k-1}}} \sum_{i=2^{k-1}+1}^{2^k} X_i \right) = \frac{1}{\sqrt{2}} \left(Y_{2^{k-1}} + Y_{2^{k-1}} \right). \tag{2.4}$$

We show in Equation (2.4) that Y_{2^k} is built of the normalised sum of 2 copies of $Y_{2^{k-1}}$. But since we can see $Y_{2^{k-1}}$ as a random variable with variance 1, we can apply the base case, which is $\mathrm{Ent}(Y_2) \geq \mathrm{Ent}(Y_1)$. This gives the desired result of Equation (2.3). Because of this it was naturally conjectured that the entire sequence Y_n was increasing with n. This problem actually remained open for a long time, until it was proven in the paper this thesis is about. It was even not proven that $\mathrm{Ent}(Y_3) \geq \mathrm{Ent}(Y_2)$ and this specifies the difficulty of this proof. There is no natural way to "build" the sum of three independent copies of X out of the sum of two.

The aim of this thesis is not only to prove this statement, but also generalise it even more for independent

random variables to a case where the variance need not be 1. This is done by generalising it even more by introducing a theorem for which the random variables are not restricted to be copies or better said, identically distributed.

2.2. Intuition behind the theorem

As said in the section before, it is quite difficult to even prove that the entropy of a normalised sum of 3 IID random variables is bigger than the entropy of a normalised sum of 2 IID random variables. To give some intuition that this is the case however, we give an example. If we take exponential distributions with parameter $\lambda = 1$ as IID random variables in the normalised sum, thus as X_i in (2.2), then we show numerically that the entropy of Y_n is increasing. The X_i were originally distributed with expected value 0 and variance 1, while the X_i chosen have expected value 1 and variance 1. This does not actually matter, since the entropy of a standard Gaussian is the same as the entropy of normally distributed random variable with expected value μ and variance 1. This is not explained further, but it is logical that the uncertainty does not depend on the expected value. What we want to show is that the entropy of Y_n, the normalised sum, goes towards the entropy of a standard Gaussian. Again Y_n goes towards a Gaussian distribution by the central limit theorem with variance 1. In the figure below the entropy of Y_n is plotted up until $n = 10$.

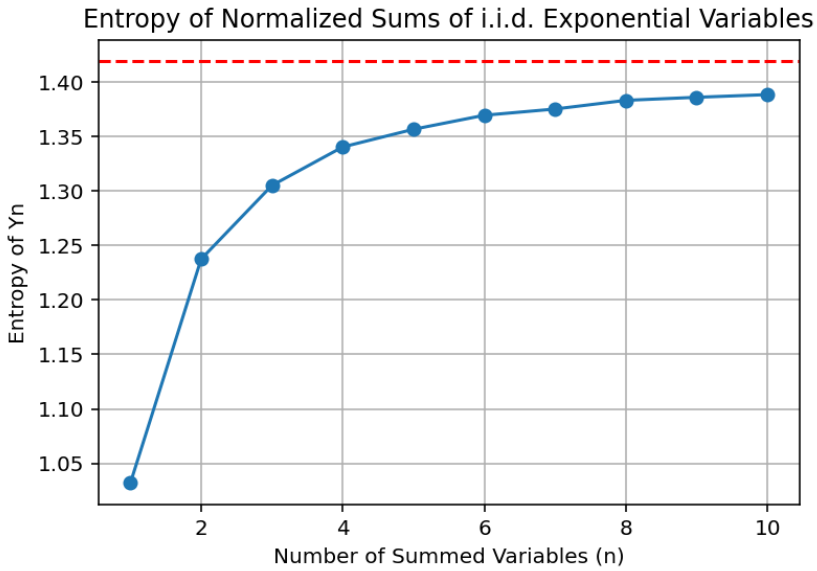


Figure 2.1: The entropy of Y_n plotted up until $n = 10$. The red-dashed line is the entropy of a standard Gaussian.

The entropy in figure 2.1 is numerically integrated by estimating the density with kernel density estimation and numerically integrating the integral in (2.1). The part about kernel density estimation is out of the scope of this thesis, but is only used to estimate the density in this case. The code with which the figure was plotted can be found in Appendix A. From the figure it is clear that the entropy is step-wise increasing towards the entropy of a standard Gaussian.

We could do this for different random variables with variance 1, but this does not actually prove anything. It does give some intuition however, that the theorems defined in the next section are correct. The theorems are even defined with less assumptions then we have defined up until now. It does not matter if the variance of the random variables are 1, the entropy will still be increasing. In that case not towards the entropy of a standard Gaussian, but towards the entropy of the Gaussian distribution with the same variance. There will even be a theorem defined in the case where the random variables are not identically distributed.

2.3. Entropic monotonicity theorem

First we introduce the theorem for independent random variables.

Theorem 1 (Entropy increases at every step)**.** *Let $X_1, X_2, \dots$ be independent and identically distributed square-integrable random variables. Then*

$$\mathrm{Ent}\left(\frac{X_1+\cdots+X_{n+1}}{\sqrt{n+1}}\right) \geq \mathrm{Ent}\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) \tag{2.5}$$

The main point of the result is that one can now see clearly that convergence in the central limit theorem is driven by an analogue of the second law of thermodynamics. The second law of thermodynamics states that in any energy transfer, the total entropy of an isolated system does not decrease. This can be compared to Theorem 1 by seeing that adding an extra random variable only increases the entropy, which can be seen as an energy transfer. Adding extra random variables thus only increases the entropy, which means that the normalised sum must converge to a Gaussian, since this has the largest entropy. This is why we can see that the central limit theorem is driven by an analogue of the second law of thermodynamics. Now we continue with a theorem for non-identically distributed random variables, since there are also versions of the central limit theorem for those.

Theorem 2. *Let $X_1, X_2, \dots, X_{n+1}$ be independent random variables and let $(a_1, \dots, a_{n+1}) \in S^n$ be a unit vector. Then*

$$\mathrm{Ent}\left(\sum_{i=1}^{n+1} a_i X_i\right) \geq \sum_{j=1}^{n+1} \frac{1-a_j^2}{n} \cdot \mathrm{Ent}\left(\frac{1}{\sqrt{1-a_j^2}} \cdot \sum_{i\neq j} a_i X_i\right). \tag{2.6}$$

In particular,

$$\mathrm{Ent}\left(\frac{X_1+\cdots+X_{n+1}}{\sqrt{n+1}}\right) \geq \frac{1}{n+1} \sum_{j=1}^{n+1} \mathrm{Ent}\left(\frac{1}{\sqrt{n}} \sum_{i\neq j} X_i\right). \tag{2.7}$$

First, we show that Equation (2.7) follows from Equation (2.6). We can choose each $a_i = \frac{1}{\sqrt{n+1}}$, since then

$$\sum_{i=1}^{n+1} a_i^2 = \sum_{i=1}^{n+1} (\frac{1}{\sqrt{n+1}})^2 = 1,$$

which implies $a \in S^n$. Then filling these a_i into Equation (2.6) gives

$$\mathrm{Ent}\left(\sum_{i=1}^{n+1} \frac{1}{\sqrt{n+1}} X_i\right) \geq \sum_{j=1}^{n+1} \frac{1-(\frac{1}{n+1})}{n} \cdot \mathrm{Ent}\left(\frac{1}{\sqrt{1-(\frac{1}{n+1})}} \cdot \sum_{i\neq j} \frac{1}{\sqrt{n+1}} X_i\right),$$

which gives

$$\mathrm{Ent}\left(\frac{1}{\sqrt{n+1}} \sum_{i=1}^{n+1} X_i\right) \geq \sum_{j=1}^{n+1} \frac{1}{n+1} \cdot \mathrm{Ent}\left(\frac{1}{\sqrt{\frac{n}{n+1}}} \cdot \sum_{i\neq j} \frac{1}{\sqrt{n+1}} X_i\right),$$

which can be seen to be exactly the same as Equation (2.7).

We are going to continue the rest of this thesis to prove Theorem 2 and show here that this theorem implies Theorem 1. If we asssume $X_1, \dots, X_{n+1}$ to be independent identically distributed square-integrable random variables, then we see by Theorem 2 that Equation (2.7) holds. If we compare this to Equation (2.5), we see that the left-hand sides are already equal. We are only left to prove that in this case

$$\mathrm{Ent}\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right) = \sum_{j=1}^{n+1} \frac{1}{n+1} \mathrm{Ent}\left(\frac{1}{\sqrt{n}} \sum_{i\neq j} X_i\right). \tag{2.8}$$

For this we need to see that with our assumptions it actually does not matter which j is not taken in the inner sum of the entropy on the right-side, since all X_i are identically distributed. Thus we can see that

$$\sum_{\substack{i=1,\\ i\neq j}}^{n+1} X_i = \sum_{i=1}^{n} X_i$$

for all j. Using this in the right-side of Equation (2.8) we see that

$$\sum_{j=1}^{n+1} \frac{1}{n+1} \operatorname{Ent}\left(\frac{1}{\sqrt{n}} \sum_{i \neq j} X_i\right) = \frac{1}{n+1} \sum_{j=1}^{n+1} \operatorname{Ent}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\right) = \frac{n+1}{n+1} \operatorname{Ent}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i\right).$$

This gives

$$\sum_{j=1}^{n+1} \frac{1}{n+1} \operatorname{Ent}\left(\frac{1}{\sqrt{n}} \sum_{i \neq j} X_i\right) = \operatorname{Ent}\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)$$

as desired, thus proving Theorem 2 is enough to prove Theorem 1.

3

Connection between the Fisher information and entropy

We start with proving Theorem 2 by transforming the problem of the form of entropy into another information-theoretic notion, the Fisher information of a random variable. The Fisher information is a way of measuring the amount of information a random variable X carries about an unknown parameter θ.

3.1. Background of the Fisher information

Let $f(x;\theta)$ be the probability density function for X conditioned on θ, which describes the probability that we observe a given outcome of X, given θ. If f changes a lot with respect to changes in θ, then it is easy to deduce the correct value of θ from the data or equivalently, the random variable provides a lot of information about the parameter. On the other hand, if f is flat and spread-out with respect to θ, then the random variable provides less information about the parameter. Here we introduce the score function, which measures the sensitivity of a change in θ and is

$$\frac{\partial}{\partial_\theta}\log(f(x;\theta)).$$

The Fisher information is the variance of the score function, since this measures how sensitive the likelihood function is to changes in the parameter. Now

$$\mathbb{V}ar(\frac{\partial}{\partial\theta}\log(f(x;\theta)) = \mathbb{E}((\frac{\partial}{\partial\theta}\log(f(x;\theta))^2) - \mathbb{E}(\frac{\partial}{\partial\theta}\log(f(x;\theta))^2 \tag{3.1}$$

We simplify this first by showing that if some regularity conditions are met (differentiability and integrability), then by using the chain rule

$$\mathbb{E}(\frac{\partial}{\partial\theta}\log(f(x;\theta)) = \int_{\mathbb{R}}\frac{\partial}{\partial\theta}\log(f(x;\theta))f(x;\theta)dx = \int_{\mathbb{R}}\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}(f(x;\theta))f(x;\theta)dx.$$

We can now see this term vanishes, because

$$\int_{\mathbb{R}}\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}(f(x;\theta))f(x;\theta)dx = \int_{\mathbb{R}}\frac{\partial}{\partial\theta}f(x;\theta)dx = \frac{\partial}{\partial\theta}\int_{\mathbb{R}}f(x;\theta)dx = \frac{\partial}{\partial\theta}1 = 0.$$

This means that

$$\mathbb{E}(\frac{\partial}{\partial\theta}\log(f(x;\theta))^2 = 0. \tag{3.2}$$

We continue with

$$\mathbb{E}((\frac{\partial}{\partial\theta}\log(f(x;\theta))^2) = \int_{\mathbb{R}}(\frac{\partial}{\partial\theta}\log(f(x;\theta))^2 f(x;\theta)dx = \int_{\mathbb{R}}\left(\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}(f(x;\theta))\right)^2 f(x;\theta)dx.$$

We can see this is equal to

$$\int_{\mathbb{R}}\frac{1}{f(x;\theta)}(\frac{\partial}{\partial\theta}(f(x;\theta))^2 dx. \tag{3.3}$$

Now filling Equations (3.2) and (3.3) into Equation (3.1) gives that the Fisher information of a random variable X with density f is

$$J(f) = \int_{\mathbb{R}} \frac{(f')^2}{f}$$

It is known that among all random variables with variance 1, the standard Gaussian has the smallest Fisher information, namely 1. If we remember that the more we knew about a random variable, the less the entropy got, we see that this is logical, since the standard Gausssian also had the biggest entropy.

3.2. Connecting Fisher information and entropy in a simple case

We now are going to try to connect the Fisher information and the entropy of a random variable. We start by introducing the adjoint Ornstein-Uhlenbeck semigroup, which are used in de Bruijn (see e.g. [8], Bakry and Emery [2] and Barron [3]. The operation of this semigroup is as follows

$$X \to X^{(t)},$$

where

$$X^{(t)} = \sqrt{e^{-2t}}X + \sqrt{1-e^{-2t}}G.$$

Here, $X^{(t)}$ is called the evolute at time t of the random variable X under the semigroup. We see that the evolute has the same distribution as an appropiately weighted sum of X with a standard Gaussian. The connection between Fisher information and entropy is made by Carlen and Soffer [4], which we will just state without proof. It says that entropy gap between a random Variable X and the standard Gaussian G can be written as an integral over all evolutes

$$\mathrm{Ent}(G) - \mathrm{Ent}(X) = \int_0^\infty \left(J(X^{(t)}) - 1\right) dt \tag{3.4}$$

Now we make the claim that proving the increase of entropy of IID random variables X, Theorem 1, can be deduced to proving a decrease of information, $J(S_{n+1}) \leq J(S_n)$. Here, S_n is the normalised sum of n IID copies of an evolute $X^{(t)}$. We start by showing that taking the evolute of Y_n, where Y_n is a normalised sum of X_i IID random variables, is the same as taking the normalised sum of n IID copies of an evolute $X^{(t)}$. We first show the evolute of Y_n,

$$Y_n^{(t)} = \left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)^{(t)} = \sqrt{e^{-2t}}\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) + \sqrt{1-e^{-2t}}G. \tag{3.5}$$

Next, we show that this is the same of taking the normalised sum of copies of an evolute $X^{(t)}$,

$$S_n = \frac{X_1^{(t)} + \ldots X_n^{(t)}}{\sqrt{n}} = \frac{\sqrt{e^{-2t}}X_1 + \sqrt{1-e^{-2t}}G_1 + \cdots + \sqrt{e^{-2t}}X_n + \sqrt{1-e^{-2t}}G_n}{\sqrt{n}}.$$

Here each G_i is just a standard Gaussian. This is equal to

$$\sqrt{e^{-2t}}\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) + \sqrt{1-e^{-2t}}\left(\frac{G_1 + \cdots + G_n}{\sqrt{n}}\right) = \sqrt{e^{-2t}}\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) + \sqrt{1-e^{-2t}}G. \tag{3.6}$$

In the last step we used the fact that taking the normalised sum of n standard Gaussians gives once again a random variable which has the distribution of a standard Gaussian. Furthermore, we used that taking IID copies of a random evolute $X^{(t)}$, which S_n does, is the same as taking the sum of the evolutes of all X_i, since the X_i are IID. Now we see that Equations (3.5) and (3.6) are the same and thus it does not matter if we first sum or first take the evolution. In other words the operation of the semi-group commutes with self-convolution. Now we show this proves that we only have to prove the decrease of information. If $J(S_{n+1}) \leq J(S_n)$, then since the Fisher information is non-negative

$$\int_0^\infty (J(S_{n+1}) - 1)\,dt \leq \int_0^\infty (J(S_n) - 1)\,dt,$$

which implies by what we proved earlier, $S_n = Y_n^{(t)}$, that

$$\int_0^\infty (J(Y_{n+1}^{(t)}) - 1)\,dt \leq \int_0^\infty (J(Y_n^{(t)}) - 1)\,dt \implies \mathrm{Ent}(G) - \mathrm{Ent}(Y_{n+1}) \leq \mathrm{Ent}(G) - \mathrm{Ent}(Y_n).$$

Here, we used Equation (3.4). But this shows exactly what we wanted to prove, $\mathrm{Ent}(Y_{n+1}) \geq \mathrm{Ent}(Y_n)$.

3.3. Connecting the fisher information and entropy in a more complex case

We now extend the proof of this claim to Theorem 2, where the X_i may not be identically distributed, but are independent. We summarise this claim in a theorem.

Theorem 3. *Let $X_1, X_2, \cdots, X_{n+1}$ be independent random variables and let $\hat{a} = (a_1, \ldots, a_{n+1}) \in S^n$ be a unit vector. If for every $b_1, \cdots, b_{n+1} \in \mathbb{R}$ satisfying*

$$\sum_{j=1}^{n+1} b_j \sqrt{1-a_j^2} = 1 \tag{3.7}$$

it holds that

$$J\left(\sum_{i=1}^{n+1} a_i X_i\right) \leq n \sum_{j=1}^{n+1} b_j^2 J\left(\frac{1}{\sqrt{1-a_j^2}} \sum_{i \neq j} a_i X_i\right), \tag{3.8}$$

then

$$\mathrm{Ent}\left(\sum_{i=1}^{n+1} a_i X_i\right) \geq \sum_{j=1}^{n+1} \frac{1-a_j^2}{n} \cdot \mathrm{Ent}\left(\frac{1}{\sqrt{1-a_j^2}} \cdot \sum_{i \neq j} a_i X_i\right).$$

Essentially this theorem is saying that if we can prove Inequality (3.8), then Theorem 2 is proved as well. This is exactly how we are going to prove Theorem 2, we first prove Theorem 3. Next, we prove that Inequality (3.8) actually holds for all b_j for which Inequality 3.7 holds. In conclusion, this proves Theorem 2, since we use the same assumptions to get the same conclusion.

Proof of Theorem 3:
Let $b_j = \frac{1}{n}\sqrt{1-a_j^2}$, then b_j fulfills Inequality 3.7, because

$$\sum_{j=1}^{n+1} \frac{1}{n}(1-a_j^2) = \frac{n+1}{n} - \frac{1}{n} = 1.$$

Here we used that $\hat{a} \in S^n$, which implies that $\sum_{j=1}^{n+1} a_j = 1$. Filling these b_j in Inequality (3.8) gives

$$J\left(\sum_{i=1}^{n+1} a_i X_i\right) \leq n \sum_{j=1}^{n+1} \frac{1}{n^2}(1-a_j^2) J\left(\frac{1}{\sqrt{1-a_j^2}} \sum_{i \neq j} a_i X_i\right). \tag{3.9}$$

Next, we remember that the only thing we assume about the X_i's are that they are independent. Now, since taking the evolute just adds a weighted Gaussian to the original random variable, the evolutes, $X_i^{(t)}$'s also satisfy Inequality (3.9). We first show just as in the earlier claim that it does not matter if we sum the evolutes or take the evolute of the sum of random variables. This follows the same arguments, thus

$$\sum_{i=1}^{n+1} a_i X_i^{(t)} = \sum_{i=1}^{n+1} \sqrt{e^{-2t}} a_i X_i + \sum_{i=1}^{n+1} \sqrt{1-e^{-2t}} a_i G_i$$

where each G_i is an independent standard Gaussian. We can now see that the summing these Gaussians gives once again a standard Gaussian, since

$$\mathbb{V}ar(\sum_{i=1}^{n+1} a_i G_i) = \sum_{i=1}^{n+1} a_i^2 \mathbb{V}ar(G_i) = \sum_{i=1}^{n+1} a_i^2 \cdot 1 = 1,$$

where in the last step we used that $\hat{a} \in S_n$. This gives

$$\sum_{i=1}^{n+1} a_i X_i^{(t)} = \sum_{i=1}^{n+1} \sqrt{e^{-2t}} a_i X_i + \sum_{i=1}^{n+1} \sqrt{1-e^{-2t}} G = \left(\sum_{i=1}^{n+1} a_i X_i\right)^{(t)}.$$

This gives exactly what we wanted to show. The arguments for $\frac{1}{\sqrt{1-a_j^2}}\sum_{i\neq j} a_i X_i$ are exactly the same, but the only difference is that we show

$$\mathbb{V}ar(\frac{1}{\sqrt{1-a_j^2}}\sum_{i\neq j} a_i G_i) = \frac{1}{1-a_j^2}\sum_{i\neq j} a_i^2 \mathbb{V}ar(G_i) = \frac{1}{1-a_j^2}\sum_{i\neq j} a_i^2 \cdot 1 = \frac{1}{1-a_j^2}(1-a_j^2) = 1,$$

The conclusion, summing the evolutes gives the same as taking the evolute of the sum is the same. Now with Inequality (3.9) and the previous results we can show that

$$\int_0^\infty (J\left(\sum_{i=1}^{n+1} a_i X_i^{(t)}\right) - 1)dt \leq n\sum_{j=1}^{n+1}\frac{1}{n^2}(1-a_j^2)\int_0^\infty (J\left(\frac{1}{\sqrt{1-a_j^2}}\sum_{i\neq j} a_i X_i^{(t)}\right) - 1)dt,$$

which is the same as

$$\int_0^\infty (J\left(\left(\sum_{i=1}^{n+1} a_i X_i\right)^{(t)}\right) - 1)dt \leq \sum_{j=1}^{n+1}\frac{1-a_j^2}{n}\int_0^\infty (J\left(\left(\frac{1}{\sqrt{1-a_j^2}}\sum_{i\neq j} a_i X_i\right)^{(t)}\right) - 1)dt.$$

But by Equation (3.4) this is equal to

$$\mathrm{Ent}(G) - \mathrm{Ent}\left(\sum_{i=1}^{n+1} a_i X_i\right) \leq \mathrm{Ent}(G) - \sum_{j=1}^{n+1}\frac{1-a_j^2}{n}\mathrm{Ent}\left(\frac{1}{\sqrt{1-a_j^2}}\sum_{i\neq j} a_i X_i\right),$$

which gives

$$\mathrm{Ent}\left(\sum_{i=1}^{n+1} a_i X_i\right) \geq \sum_{j=1}^{n+1}\frac{1-a_j^2}{n}\mathrm{Ent}\left(\frac{1}{\sqrt{1-a_j^2}}\sum_{i\neq j} a_i X_i\right). \qquad \square$$

Now to prove Inequality (3.8) holds for every b_j satisfying Inequality (3.7) we need a different theorem which we prove in Chapter 4, at last we prove Inequality (3.8) in Chapter 5, which finalises the proof of Theorem 2.

4

Variational characterisation of the information

In this chapter a theorem will be proved which is necessary for proving Inequality 3.8, which is needed to finalise the proof of Theorem (2). The theorem and proof will be split up into two parts, which will both be used in Chapter 5. For the first part we also introduce a lemma, needed to prove the first part.

4.1. Introducing the inequality

Theorem 4 (Variational characterisation of the information with inequality)**.** *Let $w : \mathbb{R}^n \to (0,\infty)$ be a continuously twice differentiable density on $\mathbb{R}^n$ with*

$$\int \frac{\|\nabla w\|^2}{w}, \quad \int \|\mathrm{Hess}(w)\| < \infty.$$

Let e be a unit vector and h the marginal density in direction e defined by

$$h(t) = \int_{te+e^\perp} w.$$

Then the Fisher information of the density h satisfies

$$J(h) \leq \int_{\mathbb{R}^n} \left(\frac{\mathrm{div}(pw)}{w}\right)^2 w,$$

for any continuously differentiable vector field $p : \mathbb{R}^n \to \mathbb{R}^n$ with the property that for every x,

$$\langle p(x), e\rangle = 1$$

(and, say, $\int \|p\| w < \infty$).

Remarks.

- The condition $\int \|p\| w < \infty$ is not important in applications but makes for a cleaner statement of the theorem. The authors of Artstein [1] thanked Giovanni Alberti for pointing out this simplified formulation of the result.
- Here the gradient of w is $\nabla w(x) = \begin{bmatrix} \frac{\partial w}{\partial x_1} \\ \frac{\partial w}{\partial x_2} \\ \vdots \\ \frac{\partial w}{\partial x_n} \end{bmatrix}$.

- The Hessian matrix of h is $\text{Hess}(h)(x) = \begin{bmatrix} \frac{\partial^2 h}{\partial x_1^2} & \frac{\partial^2 h}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 h}{\partial x_1 \partial x_n} \\ \frac{\partial^2 h}{\partial x_2 \partial x_1} & \frac{\partial^2 h}{\partial x_2^2} & \cdots & \frac{\partial^2 h}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 h}{\partial x_n \partial x_1} & \frac{\partial^2 h}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 h}{\partial x_n^2} \end{bmatrix}$.
- $\|\text{Hess}(h)(x)\| = \left(\sum_{i=1}^n \sum_{j=1}^n \left(\frac{\partial^2 h}{\partial x_i \partial x_j}\right)^2\right)^{1/2}$, so the norm of the Hessian matrix is like seeing the matrix as one big vector by putting the rows next to each other and taking the normal norm over that vector.
- The divergence of pw is $\text{div}(pw) = \sum_{i=1}^n \frac{\partial}{\partial x_i}\left(p_i(x)w(x)\right) = \nabla \cdot pw(x)$
- $h(t) = \int_{te+e^\perp} w$. means that h(t) is the integral over all vectors perpendicular to te. Thus, this is an integral over (n-1) dimensions. If we were to integrate $h(t)$ over t in the interval $(-\infty, \infty)$ we would integrate all of $\mathbb{R}^n$

4.1.1. Supporting lemma for the proof

First, a Lemma is introduced, which will be used in the proof of the first theorem.

Lemma 1. *If w and h satisfy the same assumptions of Theorem 4, then*

$$\int_{te+e^\perp} \text{div}(pw) = \int_{te+e^\perp} \partial_e w$$

Proof of Lemma 1

We prove this by decomposing the divergence into two parts. We do this by decomposing $pw(x)$ into $pw_e(x) = \langle pw(x), e\rangle e$, the projection of pw along e, and $pw_{e^\perp}(x) = pw(x) - pw_e(x)$. Then $pw(x) = pw_e(x) + pw_{e^\perp}(x)$. Then

$$\int_{te+e^\perp} \text{div}(pw(x)) = \int_{te+e^\perp} \text{div}_e(pw(x)) + \int_{te+e^\perp} \text{div}_{e^\perp}(pw(x)). \tag{4.1}$$

We first show that the most-right integral is zero by the Green-Gauss theorem. For this we use the ball with radius R, $B(R)$, on the hyperplane and take the limit of $R \to \infty$ to integrate over this hyperplane.

$$\left|\int_{te+e^\perp} \text{div}_{e^\perp}(pw(x))\right| = \left|\lim_{R\to\infty} \int_{B(R)} \text{div}_{e^\perp}(pw(x))\right|$$

Now we use Green-Gauss, where $S(R)$ is the boundary of the ball $B(R)$ and N(x) is the outward pointing normal vector to the ball:

$$\left|\lim_{R\to\infty} \int_{B(R)} \text{div}_{e^\perp}(pw(x))\right| = \left|\lim_{R\to\infty} \int_{S(R)} pw(x) \cdot N(x)\right| \leq \lim_{R\to\infty} \int_{S(R)} \|pw(x)\| = \lim_{R\to\infty} \int_{S(R)} \|p(x)\| w(x)$$

We assume that this last function is integrable over $\mathbb{R}^n$, thus by substitution (transformation formula)

$$\int_{\mathbb{R}^n} \|p(x)\| w(x) = \int_{\mathbb{R}_{\geq 0}} \int_{S(1)} \|p(rx)\| w(rx) r^{n-1} dx dr < \infty.$$

Since this integral is less than ∞, the inner integral decays to 0 as $r \to \infty$, thus

$$\lim_{r\to\infty} \int_{S(1)} \|p(rx)\| w(rx) r^{n-1} dx = \lim_{r\to\infty} \int_{S(r)} \|p(x)\| w(x) dx = 0.$$

This shows

$$\int_{te+e^\perp} \text{div}_{e^\perp}(pw(x)) = 0. \tag{4.2}$$

Now we continue with the other integral in Equation (4.1). Here we use the product rule for divergence with the constant vector field e and the scalar-valued function $\langle pw(x), e\rangle$

$$\int_{te+e^\perp} \text{div}_e(pw(x)) = \int_{te+e^\perp} \text{div}(\langle pw(x), e\rangle e) = \int_{te+e^\perp} \nabla(\langle pw(x), e\rangle) \cdot e + \langle pw(x), e\rangle \,\text{div}(e).$$

Now $\operatorname{div}(e) = 0$, since e is just a constant vector. This gives

$$\int_{te+e^\perp} \nabla(\langle pw(x), e\rangle) \cdot e = \int_{te+e^\perp} \partial_e \langle pw(x), e\rangle = \int_{te+e^\perp} \partial_e (w(x)\langle p(x), e\rangle) = \int_{te+e^\perp} \partial_e w(x)$$

where at the end we use that $\langle p(x), e\rangle = 1$. Now filling this all in Equation (4.1) gives

$$\int_{te+e^\perp} \operatorname{div}(pw(x)) = \int_{te+e^\perp} \partial_e w(x) + 0. \qquad \square$$

4.1.2. Proof of the inequality

Now we continue with the proof of Theorem 4.
Proof of Theorem 4:
We first derive a different version of the derivative of $h(t)$, which we use later

$$h'(t) = \frac{d}{dt} \int_{te+e^\perp} w(x).$$

We perform a change of variables to get t outside the integral boundary, since then we can apply the Leibniz rule. We use $x = te + y, \quad y \in e^\perp$, then

$$h'(t) = \frac{d}{dt} \int_{e^\perp} w(te+y).$$

This shows that the boundaries of the integral are now not dependent on t. Now using the change of differentiation and integration, which we can use since w is smooth and continuously differentiable twice and after applying the chain rule we get

$$h'(t) = \int_{e^\perp} \frac{d}{dt} w(te+y) = \int_{e^\perp} \nabla w(te+y) \cdot e = \int_{e^\perp} \partial_e w(te+y) = \int_{te+e^\perp} \partial_e w(x).$$

At the end we use that ∂_e of a function is the inproduct of the gradient of the function and e and we perform the original change of variables backwards. This gives thus

$$h'(t) = \int_{te+e^\perp} \partial_e w(x) \tag{4.3}$$

Now if

$$\int_{\mathbb{R}^n} \left(\frac{\operatorname{div}(pw)}{w}\right)^2 w$$

is finite, then $\operatorname{div}(pw)$ is integrable on $\mathbb{R}^n$. We show this by using a form of the Cauchy-Schwarz inequality, Hölder's inequality, with the following dot product. Note that if the integral is infinite, then the Fisher information is always less then this integral and the theorem follows trivially.

$$\langle f, g\rangle = \int fg \quad \text{and} \quad \|fg\|_1 \le \|f\|_2 \|g\|_2.$$

We use $f = \frac{\operatorname{div}(pw)}{\sqrt{w}}$ and $g = \sqrt{w}$ to get

$$\int_{\mathbb{R}^n} |\operatorname{div}(pw)| = \int_{\mathbb{R}^n} |\frac{\operatorname{div}(pw)}{\sqrt{w}} \sqrt{w}| \le \sqrt{\int_{\mathbb{R}^n} \frac{\operatorname{div}^2(pw)}{w}} \sqrt{\int_{\mathbb{R}^n} w} < \infty \quad . \tag{4.4}$$

The two integrals multiplied are less than infinity are because we have assumed the left integral is integrable and the right integral is integrable by definition, since it is a density. Thus $\operatorname{div}(pw)$ is integrable over $\mathbb{R}^n$. By the full Lebesgue version of Fubini's theorem $\operatorname{div}(pw)$ is integrable on almost every hyperplane perpendicular to e, since

$$\int_{\mathbb{R}^n} \operatorname{div}(pw) = \int_{\mathbb{R}} \int_{te+e^\perp} \operatorname{div}(pw).$$

We are going to use Lemma 1 to see that

$$\int_{te+e^\perp} \partial_e w = \int_{te+e^\perp} \operatorname{div}(pw) \tag{4.5}$$

We can now fill Equations (4.3) and (4.5) in the Fisher information of the marginal density h.

$$J(h) = \int_{\mathbb{R}} \frac{h'(t)^2}{h(t)} = \int_{\mathbb{R}} \frac{(\int_{te+e^\perp} \operatorname{div}(pw))^2}{\int_{te+e^\perp} w}.$$

Next, we use the normal version of the Cauchy-Schwarz inequality

$$|\langle f, g\rangle|^2 \leq \langle f, f\rangle\langle g, g\rangle.$$

We use the same functions as in Equation (4.4) to show

$$\left(\int_{te+e^\perp} \operatorname{div}(pw)\right)^2 = \left(\int_{te+e^\perp} \frac{\operatorname{div}(pw)}{\sqrt{w}}\sqrt{w}\right)^2 \leq \int_{te+e^\perp} \frac{\operatorname{div}(pw)^2}{w} \int_{te+e^\perp} w.$$

Now we fill this in to see that

$$\int_{\mathbb{R}} \frac{(\int_{te+e^\perp} \operatorname{div}(pw))^2}{\int_{te+e^\perp} w} \leq \int_{\mathbb{R}} \frac{\int_{te+e^\perp} \frac{\operatorname{div}(pw)^2}{w} \int_{te+e^\perp} w}{\int_{te+e^\perp} w} = \int_{\mathbb{R}} \int_{te+e^\perp} \frac{\operatorname{div}(pw)^2}{w}.$$

We now use Fubini's theorem again to complete the theorem. We use it in the way that if we integrate over all possible hyperplanes, we essentially integrate over $\mathbb{R}^n$.

$$J(h) \leq \int_{\mathbb{R}^n} \frac{\operatorname{div}(\mathrm{pw})^2}{w}. \qquad \square \tag{4.6}$$

4.2. Proof of equality in the theorem

We continue with showing the case where there is equality in the inequality of Theorem 4

Theorem 5 (Variational characterization of the information with equality)**.** *If w, h and p satisfy the same assumptions as in Theorem 4, then if w satisfies $\int \|x\|^2 w(x) < \infty$, then there is equality in the inequality*

$$J(h) \leq \int_{\mathbb{R}^n} \left(\frac{\operatorname{div}(pw)}{w}\right)^2 w,$$

for some suitable vector field p.

Proof of Theorem 5:
We are trying to find a suitable vector field p such that there is equality if $\int \|x\|^2 w(x) < \infty$. We start with showing with that there exists a p such that for all t

$$\operatorname{div}(pw) = \frac{h'(t)}{h(t)} w. \tag{4.7}$$

Since then

$$\int_{\mathbb{R}^n} \frac{\operatorname{div}(pw)^2}{w} = \int_{\mathbb{R}^n} \frac{h'(t)^2}{h(t)^2} w = \int_{\mathbb{R}} \frac{h'(t)^2}{h(t)^2} \int_{te+e^\perp} w.$$

But now we see that

$$\int_{\mathbb{R}} \frac{h'(t)^2}{h(t)^2} \int_{te+e^\perp} w = \int_{\mathbb{R}} \frac{h'(t)^2}{h(t)^2} h(t) = J(h).$$

Since we also need to ensure that $\langle p, e\rangle = 1$ identically, we construct p separately on each hyperplane perpendicular to e. This means we don't construct p on the normal x_i coordinates, but we fix an orthonormal basis of $e^\perp$ and index the coordinates of $y \in e^\perp$ with respect to this basis, $y_1, y_2, \cdots, y_{n-1}$. We then index $y_n = e$ to have an orthonormal basis of $\mathbb{R}$. We construct p separately on each hyperplane, which means that we construct it for each t. Remember that we split the space $\mathbb{R}^n$ in $te + y$. Regularity of p on the whole space can be ensured by using a "consistent" method for solving the equation on the separate hyperplanes. This means that we will show that p will not depend on t when solving for p. At last, we fix the n^{th} component of p, p_n, to be 1, so that $\langle p, e\rangle = 1$.

Then the last thing p needs to satisfy is

$$\operatorname{div}_{e^\perp}(pw) = \frac{h'(t)}{h(t)} w - \partial_e w \tag{4.8}$$

since then by splitting the divergence in the same way as Equation (4.1), using the fact that $\operatorname{div}_e(pw) = \partial_e(w\langle p,e\rangle)$ and filling in Equation (4.8) we see that then

$$\operatorname{div}(pw) = \frac{h'(t)}{h(t)}w - \partial_e w + \partial_e(w\langle p,e\rangle) = \frac{h'(t)}{h(t)}w,$$

where in the last step we used that we construct p such that $\langle p,e\rangle = 1$. Thus we see if we construct p such that Equation (4.8) holds, then Equation (4.7) holds as desired.

We can see that $\operatorname{div}_{e^\perp}(pw)$ is integrable on $te+e^\perp$, since $\frac{h'(t)}{h(t)}$ is just a constant and w and $\partial_e w$ are integrable by the assumptions of w. We furthermore see by Equation (4.3) that $\int_{te+e^\perp}\partial_e w = h'(t)$ and that thus

$$\int_{te+e^\perp}\frac{h'(t)}{h(t)}w - \int_{te+e^\perp}\partial_e w = \frac{h'(t)}{h(t)}\int_{te+e^\perp} w - h'(t) = 0 \tag{4.9}$$

where in the penultimate step we used that $\int_{te+e^\perp} w = h(t)$. This ensures that the integral of the divergence on the hyperplane is 0 when we define the divergence as in Equation (4.8). This corresponds with what we showed earlier, that the divergence on the hyperplane needs to be 0 by Green-Gauss, as shown in Equation (4.2).

The hypothesis $\int \|x\|^2 w(x)\,dx < \infty$ is needed only if we wish to construct a p for which $\int \|p(x)\|^2 w(x)\,dx < \infty$. For each real t and each $y \in e^\perp$, set

$$F(t,y) = \frac{h'(t)}{h(t)}w(te+y) - \partial_e w(te+y).$$

We are first going to show that the assumption $\int_{\mathbb{R}^n}\frac{\|\nabla w\|^2}{w} < \infty$, implies $\int_{\mathbb{R}^n}\frac{F^2}{w} < \infty$.

$$F^2 = \frac{h'^2}{h^2}w^2 - 2\frac{h'}{h}w\partial_e w + \partial_e w^2 \implies \frac{F^2}{w} = \frac{h'^2}{h^2}w - 2\frac{h'}{h}\partial_e w + \frac{\partial_e w^2}{w}$$

Now we show that all three of the terms on the right-side are integrable over $\mathbb{R}^n$ and thus showing that the left-side is integrable. We start with $\frac{h'^2(t)}{h^2(t)}w(te+y)$, we show using Fubini's theorem again that

$$\int_{\mathbb{R}^n}\frac{h'^2(t)}{h^2(t)}w(te+y) = \int_{\mathbb{R}}\frac{h'^2(t)}{h^2(t)}\int_{te+y} w(te+y) = \int_{\mathbb{R}}\frac{h'^2(t)}{h^2(t)}h(t) = J(h).$$

This shows it is integrable, since we assume the Fisher information is finite, otherwise any vector field which gives an indefinite integral gives equality in (4.6) . Following the same arguments we show that the second term is integrable

$$\int_{\mathbb{R}^n}\frac{h(t)}{h(t)}w(te+y) = \int_{\mathbb{R}}\frac{h(t)}{h(t)}\int_{te+y} w(te+y) = \int_{\mathbb{R}}\frac{h'(t)}{h(t)}h(t) = J(h).$$

Now to show that the third therm, $\frac{|(\partial_e w)^2|}{w}$ is integrable , we use that

$$\frac{(\partial_e w)^2}{w} \le \frac{\|\nabla w\|^2}{w}$$

and since the right-side is integrable so is the left side.
Next we use Hölder's inequality again, just as in (4.4), to prove that $|F|\|x\|$ is integrable.

$$\int_{\mathbb{R}^n}|F|\|x\| = \int_{\mathbb{R}^n}\frac{|F|}{\sqrt{w}}\|x\|\sqrt{w} \le \sqrt{\int_{\mathbb{R}^n}\frac{F^2}{w}}\sqrt{\int_{\mathbb{R}^n}\|x\|^2 w}.$$

We have earlier proved that $\frac{F^2}{w}$ is integrable and we have earlier assumed that $\|x\|^2 w$ is integrable and thus we have proven that $\int |F|\|x\| < \infty$. This integrability can be used to find solutions to our equation

$$\operatorname{div}_{e^\perp}(pw) = F(t,y)$$

satisfying $\int_{\mathbb{R}^n} \|pw\| < \infty$, for example in the following way. We build the remaining $n-1$ components of p and show that the divergence equation (4.8) holds as desired. Remember that in this process t is fixed and we do this process for every t, thus every hyperplane. In the next part the p will thus be independent of t. We use the same orthonormal basis of $e^\perp$ as earlier. Let $K(y_2,\dots,y_{n-1}) = \int F(t,y_1,y_2,\dots,y_{n-1})dy_1$ and choose a rapidly decreasing density g on the line in the y_1 direction. Then the function

$$F(t,y) - g(y_1)K(y_2,\dots,y_{n-1})$$

integrates to zero on every line in the y_1 direction, since

$$\int F(t,y) - g(y_1)K(y_2,\dots,y_{n-1})dy_1 = \int F(t,y)dy_1 - K(y_2,\dots,y_{n-1})\int g(y_1)dy_1$$

and since g is a density, this is equal to

$$\int F(t,y)dy_1 - K(y_2,\dots,y_{n-1})1 = \int F(t,y)dy_1 - \int F(t,y)dy_1 = 0.$$

This means that we can find $p_1(y)$ such that

$$\frac{\partial}{\partial y_1}(p_1(y)w(te+y)) = F(t,y) - g(y_1)K(y_2,\dots,y_{n-1}) \tag{4.10}$$

and we set $p_1 w$ by

$$p_1(y)w(te+y) = \int_{-\infty}^{y_1} F(t,s,y_2,\dots,y_{n-1}) - g(s)K(y_2,\dots,y_{n-1})ds. \tag{4.11}$$

First, we note that $p_1 w$ goes to 0 at infinity, since g is rapidly decreasing and that $|F|\|x\|$ decays in all directions, thus so does F. This is also the reason the entire constructed p will be integrable. Next we continue to do the same steps as before, but with a new function F_2. We make this the residual of Equation (4.10). Take $F_2(y) = g(y_1)K(y_2,\dots,y_{n-1})$ and take

$$K_2(y) = \int F_2(y_1,y_2,\dots,y_{n-1})dy_2 = g(y_1)\int K(y_2,\cdots,y_{n-1})dy_2 = g(y_1)\int\int F(t,y)dy_1dy_2$$

and choose a rapidly decreasing density g_2 on this line. Then following the same arguments the function

$$F_2(y) - g_2(y_2)K_2(y)$$

integrates to zero on every line in the y_2 direction. Thus we can find $p_2(y)$, such that

$$\frac{\partial}{\partial y_2}(p_2(y)w(te+y)) = F_2(y) - g_2(y_2)K_2(y)$$

which gives

$$\frac{\partial}{\partial y_2}(p_2(y)w(te+y)) = F_2(y) - g_2(y_2)g(y_1)\int\int F(t,y)dy_1dy_2. \tag{4.12}$$

We continue this process $n-1$ times and then fill Equations (4.10) and (4.12) and the rest of the partial derivatives into the divergence and show that Equation (4.8) holds as desired and we found our p.

$$\mathrm{div}_{e^\perp}(p(y)w(te+y)) = \sum_{i=1}^{n-1}\frac{\partial}{\partial y_i}(p(y)w(te+y)) =$$

$$F(t,y) - g(y_1)K(y) + g(y_1)K(y) - g_2(y_2)K_2(y) + \cdots + g_{n-2}(y_{n-2})K_{n-1}(y) - 0,$$

which shows

$$\mathrm{div}_{e^\perp}(pw) = F(t,y).$$

For this to hold, we only need to show that $\frac{\partial}{\partial y_{n-1}}(p_{n-1}(y)w(te+y)) = g(y_{n-2})K_{n-2}(y) - 0$. If we follow the same steps, we see that

$$\frac{\partial}{\partial y_{n-1}}(p_{n-1}(y)w(te+y)) = g_{n-2}(y_{n-2})K_{n-2}(y) - g_{n-1}(y_{n-1})K_{n-1}(y). \tag{4.13}$$

Here, K_{n-1} is just the integral over the entire hyperplane, this integral thus does not depend on any variables. Now looking at Equation (4.12), we can see that

$$g_{n-1}(y_{n-1})K_{n-1}(y) = g_{n-1}(y_{n-1})\cdots g_1(y_1)\int_{te+e^{\perp}} F(t,y) = 0$$

since each g is a rapidly decreasing density and we can see the integral is 0 from Equation (4.9). This proves Equation (4.13) and therefore the entire theorem. At last we show the constructed p as follows. We note that p was only independent of t, because t was fixed. In reality this t does influence p, but the p is constructed in the exact same way as described in the part before. Therefore, since p is a function from $\mathbb{R}^n \longrightarrow \mathbb{R}^n$, thus

$$p(t,y) = \begin{bmatrix} p_1(t,y) \\ p_2(t,y) \\ \vdots \\ p_{n-1}(t,y) \\ 1 \end{bmatrix}. \tag{4.14}$$

Remember that $p_n(y) = 1$, so that $\langle p, e\rangle = 1$. This is done regardless of t. From Equations (4.11),(4.12) and (4.13) we can see that respectively

$$p_1(t,y) = \frac{\int_{-\infty}^{y_1} F(t,s,y_2,\ldots,y_{n-1}) - g(s)K(y_2,\ldots,y_{n-1})ds}{w(te+y)}$$

,

$$p_2(t,y) = \frac{\int_{-\infty}^{y_2} F_2(y_1,s,y_3,\ldots,y_{n-1}) - g_2(s)g_1(y_1)(\int\int F(t,y)dy_2dy_1)ds}{w(te+y)}$$

and

$$p_{n-1}(t,y) = \frac{\int_{-\infty}^{y_{n-1}} F_{n-1}(y_1,\ldots,y_{n-2},s)ds}{w(te+y)}.$$

Now filling these three equations and all other p_i's constructed with the same steps into Equation (4.14) gives the constructed p.□

5

Finalising the proof of the Entropic Monotonicity Theorem

In this chapter, we complete the remaining steps from Chapter 3 to finalize the proof of Theorem 2, and consequently, Theorem 1. Specifically, we establish that Inequality (3.8) holds for all b_j as stated in Theorem 3. The proof of this relies on two supporting lemmas, which are also proven in this chapter.

5.1. The assumptions of the proof of the Entropic Monotonicity Theorem

In the proof of the Theorem in Section 5.2, let f_i be the density of independent random variables X_i and consider the product density

$$w(x_1, \cdots, x_{n+1}) = f_1(x_1) \cdots f_{n+1}(x_{n+1}).$$

The density of $\sum_{i=1}^{n+1} a_i X_i$ is the marginal of w in the direction of $(a_1, \cdots, a_n) \in S_n$. We shall show in the next section that if w satisfies the conditions of Theorems 4 and 5, then Theorem 6 holds. Remember that the assumptions for Theorem 4 are that w is twice continuously differentiable and that

$$\int \frac{\|\nabla w\|^2}{w}, \int \|\operatorname{Hess}(w)\| < \infty$$

and that the assumption for Theorem 5 is

$$\int \|x\|^2 w < \infty.$$

We only show that the assumption $\int \frac{\|\nabla w\|^2}{w} < \infty$ holds for non-trivial cases. Non-trivial means that the Fisher information is finite. Otherwise all the theorems hold trivially. In the case that the Fisher information is finite, we have that

$$\int \frac{f_i'^2}{f_i} < \infty \tag{5.1}$$

for all i. We summarise the proof of the assumptions in two different lemmas, one for each theorem's assumptions.

Lemma 2 (Assumptions for Theorem 4)**.** *Let, $X_1, \cdots, X_{n+1}$ be independent random variables with density functions $f_1, \cdots f_{n+1}$ and finite Fisher information, then the product density w is twice continuously differentiable and*

$$\int \frac{\|\nabla w\|^2}{w} < \infty.$$

Remark: The assumption that $\|\operatorname{Hes}(w)\|$ is also integrable is thought to follow from this lemma, but is not discussed in this thesis and thought of to be true.

Proof of Lemma 2:
Since the Fisher information is finite for each random variable, we can see from (5.1) that each density is twice

continuously differentiable. From this we can deduce that the product density is also twice continuously differentiable. Next, we see that

$$\|\nabla w\|^2 = (f_1'(x_1)\cdots f_{n+1}(x_{n+1}))^2 + \cdots + (f_1(x_1)\cdots f_{n+1}'(x_{n+1}))^2$$

which gives

$$\frac{\|\nabla w\|^2}{w} = \frac{(f_1'(x_1)\cdots f_{n+1}(x_{n+1}))^2}{f_1(x_1)\cdots f_{n+1}(x_{n+1}} + \cdots + \frac{(f_1(x_1)\cdots f_{n+1}'(x_{n+1}))^2}{f_1(x_1)\cdots f_{n+1}(x_{n+1}}$$

which is equal to

$$f_2(x_2)\cdots f_{n+1}(x_{n+1})\frac{f_1'(x_1)^2}{f_1(x_1)} + f_1(x_1)\cdots f_n(x_n)\frac{f_{n+1}'(x_{n+1})^2}{f_{n+1}(x_{n+1})}.$$

This results in

$$\int \frac{\|\nabla w\|^2}{w} = \int f_2(x_2)\cdots f_{n+1}(x_{n+1})\left(\frac{f_1'(x_1)^2}{f_1(x_1)}\right) + \cdots + \int f_1(x_1)\cdots f_n(x_n)\left(\frac{f_{n+1}'(x_{n+1})^2}{f_{n+1}(x_{n+1})}\right) < \infty.$$

We can make the conclusion that this integral is less than infinity from the fact that all Fisher informations are finite, (5.1), and that each integral is multiplied with densities, which are also finite. This means that each integral is less than infinity and so is the sum of all these integrals. □

Next, we continue with the assumption for Theorem 5. In that case, we don't need that the Fisher information is finite, but that the variance of each X_i is finite. If the variance is infinite, we are once again in a trivial case and the subsequent theorems follow trivially.

Lemma 3 (Assumptions for Theorem 5)**.** *Let, $X_1, \cdots, X_{n+1}$ be independent random variables with density functions $f_1, \cdots f_{n+1}$ and finite variance, then*

$$\int_{\mathbb{R}^{n+1}} \|x\|^2 w < \infty,$$

where w is the product density.

Proof of Lemma 3
We first show that

$$\int_{\mathbb{R}^{n+1}} \|x\|^2 w = \int_{\mathbb{R}^{n+1}} \sum_{i=1}^{n+1} (x_i^2) w = \sum_{i=1}^{n+1} \int_{\mathbb{R}^{n+1}} x_i^2 w. \tag{5.2}$$

Next we fill in the product density and split up the integrals to show

$$\sum_{i=1}^{n+1} \int_{\mathbb{R}^{n+1}} x_i^2 f_1(x_1)\cdots f_{n+1}(x_{n+1}) dx_{n+1}\cdots dx_1 = \sum_{i=1}^{n+1} \int_{\mathbb{R}} x_i^2 f_i(x_i) dx_i \prod_{j=1, j\neq i}^{n} \int_{\mathbb{R}} f_j(x_j) dx_j.$$

We simply integrate all the densities out of the equation. We can do this because all the densities which are not dependent on x_i have integrals which are equal to 1 when integrating over their entire domain. This gives

$$\sum_{i=1}^{n+1} \int_{\mathbb{R}} x_i^2 f_i(x_i) dx_i \prod_{j=1, j\neq i}^{n} \int_{\mathbb{R}} f_j(x_j) dx_j / = \sum_{i=1}^{n+1} \int_{\mathbb{R}} x_i^2 f_i(x_i) dx_i \cdot 1 \cdots \cdot 1.$$

But now remember the definition of an expected value to see that

$$\sum_{i=1}^{n+1} \int_{\mathbb{R}} x_i^2 f_i(x_i) dx_i = \sum_{i=1}^{n+1} \mathbb{E}(X_i^2). \tag{5.3}$$

Combining Equations (5.2) and (5.3) gives

$$\int_{\mathbb{R}^{n+1}} \|x\|^2 w = \sum_{i=1}^{n+1} \mathbb{E}(X_i^2).$$

But we can see that saying that this integral is less than infinity is equal to saying that each X_i has finite variance, since the variance is defined as

$$\mathbb{V}ar(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2.$$

Remember that the variance of each X_i was assumed to be finite. This proves the lemma. □

5.2. The final piece of the proof of the Entropic Monotonicity Theorem

Theorem 6. *Let $X_1, X_2 \dots X_{n+1}$ be independent random variables, let $\hat{a} = (a_1, \dots, a_{n+1}) \in S_n$ be a unit vector and $b_1, \dots, b_{n+1} \in \mathbb{R}$, then*

$$J\left(\sum_{i=1}^{n+1} a_i X_i\right) \leq n \sum_{j=1}^{n+1} b_j^2 J\left(\frac{1}{\sqrt{1-a_j^2}} \sum_{i \neq j} a_i X_i\right).$$

Proof of Theorem 6
For every j, denote a new vector

$$\hat{a}^j = \frac{1}{\sqrt{1-a_j^2}}(a_1, \cdots, a_{j-1}, 0, a_{j+1}, \cdots, a_{n+1})$$

which is also a unit vector, since

$$\sum_{i=1}^{n+1} (\hat{a}_i^j)^2 = \frac{1}{1-a_j^2} \sum_{i \neq j} a_i^2 = \frac{1}{1-a_j^2} \cdot (1-a_j^2) = 1.$$

Now let $p^j : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ be a vector field which realises the information of the marginal of w in direction $\hat{a}^j$ as in Theorem 5. This gives

$$J\left(\frac{1}{\sqrt{1-a_j^2}} \sum_{i \neq j} a_i X_i\right) = \int_{\mathbb{R}^{n+1}} \left(\frac{\operatorname{div}(w p^j)}{w}\right)^2 w. \tag{5.4}$$

Furthermore, we want to show that we can assume that p^j does not depend on x_j and its j^{th} coordinate is 0. We do this by applying Theorem 5 again, but then in n dimensions. We can do this by dropping the j^{th} coordinate, since $\hat{a}_j^j = 0$ and because of that it follows that X_j will not add anything to the Fisher information. This gives a vector field $p_2^j : \mathbb{R}^n \to \mathbb{R}^n$ which realises the information of the marginal of w_2 in the direction $\hat{a}^j$, where

$$w_2(x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_{n+1}) = f_1(x_1) \cdots f_{j-1}(x_{j-1}) f_{j+1}(x_{j+1}) \cdots f_{n+1}(x_{n+1})$$

such that

$$J\left(\frac{1}{\sqrt{1-a_j^2}} \sum_{i \neq j} a_i X_i\right) = \int_{\mathbb{R}^n} \left(\frac{\operatorname{div}(w_2 p_2^j)}{w_2}\right)^2 w_2. \tag{5.5}$$

Essentially what we want to show is that we can construct p^j by taking it to be exactly the same as p_2^j, but adding an extra j^{th} coordinate which is 0. This would also mean p^j would not depend on x_j, since that coordinate did not even exist when p_2^j was made. We start by showing that

$$\operatorname{div}(w p^j) = \sum_{k=1, k \neq j}^{n+1} \left(\frac{\partial}{\partial x_k}(w p_2^j)_k\right) + \frac{\partial}{\partial x_j}(w p^j)_j = f_j(x_j) \sum_{k=1, k \neq j}^{n+1} \left(\frac{\partial}{\partial x_k}(w_2 p_2^j)_k\right) = f_j(x_j) \operatorname{div}(w_2 p_2^j). \tag{5.6}$$

In the second step we used that $(w p^j)_j = 0$ and every other function of $p^j, (p^j)_k = (p_2^j)_k$. We can see from Equations (5.4) and (5.5) that for this construction to work we need to be able to fill the constructed p_j into Equation (5.4) and get the same result as Equation (5.5). We fill the constructed p^j into (5.4) and use (5.6) to see that

$$\int_{\mathbb{R}^{n+1}} \left(\frac{\operatorname{div}(w p^j)}{w}\right)^2 w = \int_{\mathbb{R}^{n+1}} \left(\frac{f_j(x_j) \operatorname{div}(w_2 p_2^j)}{f_j(x_j) w_2}\right)^2 f_j(x_j) w_2 = \int_{\mathbb{R}^{n+1}} \left(\frac{\operatorname{div}(w_2 p_2^j)}{w_2}\right)^2 f_j(x_j) w_2.$$

Now we split the integrals over the domain of $\mathbb{R}^n$ and x_j and use the fact that $f_j(x_j)$ is a density to see that

$$\int_{\mathbb{R}^{n+1}} \left(\frac{\operatorname{div}(w p^j)}{w}\right)^2 w = \int_{\mathbb{R}^n} \left(\frac{\operatorname{div}(w_2 p_2^j)}{w_2}\right)^2 w_2 \int_{\mathbb{R}} f_j(x_j) dx_j = \int_{\mathbb{R}^n} \left(\frac{\operatorname{div}(w_2 p_2^j)}{w_2}\right)^2 w_2 = J\left(\frac{1}{\sqrt{1-a_j^2}} \sum_{i \neq j} a_i X_i\right).$$

This has shown that in (5.4) we can use the fact that p_j is independent of the x_j coordinate and its j^{th} coordinate is 0.
Next, we consider the vector field $p:\mathbb{R}^{n+1}\to\mathbb{R}^{n+1}$ given by $p=\sum_{j=1}^{n+1} b_j p^j$. To make sure we can use Theorem 4 we have to show that $\langle p,\hat{a}\rangle = 1$ and $\int \|p\| w < \infty$. The second part is clearly satisfied, since each p_j was constructed to have $\int \|p_j\| w < \infty$. Next, we show that $\langle p,\hat{a}\rangle = 1$.

$$\langle p,\hat{a}\rangle = \sum_{i=1}^{n+1} a_i p_i = \sum_{i=1}^{n+1} a_i \sum_{j=1}^{n+1} b_j p_i^j = \sum_{j=1}^{n+1} b_j \sum_{i=1}^{n+1} a_i p_i^j, \tag{5.7}$$

by changing the sums. Next, since $\langle \hat{a}_j, p_j\rangle = 1$,

$$\langle \hat{a}_j, p_j\rangle = \frac{1}{\sqrt{1-a_j^2}} \sum_{i\neq j} p_i^j a_i = 1 \implies \sum_{i\neq j} p_i^j a_i = \sqrt{1-a_j^2}.$$

Filling this in Equation (5.7) and using that $\sum_{j=1}^{n+1} b_j\sqrt{1-a_j^2} = 1$, we get that

$$\langle p,\hat{a}\rangle = \sum_{j=1}^{n+1} b_j \sum_{i=1}^{n+1} a_i p_i^j = \sum_{j=1}^{n+1} b_j\sqrt{1-a_j^2} = 1.$$

This makes sure we can use Theorem 4 to see that

$$J\left(\sum_{i=1}^{n+1} a_i X_i\right) \le \int_{\mathbb{R}^{n+1}} \left(\frac{\operatorname{div}(wp)}{w}\right)^2 w = \int_{\mathbb{R}^{n+1}} \left(\sum_{j=1}^{n+1} b_j \frac{\operatorname{div}(wp^j)}{w}\right)^2 w. \tag{5.8}$$

Let y_j denote $b_j\frac{\operatorname{div}(wp^j)}{w}$. Our aim is to show that in $L_2(w)$ (the Hilbert space of absolutely square integrable functions with weight w):

$$\|y_1+\cdots+y_{n+1}\|^2 \le n(\|y_1\|^2+\cdots+\|y_{n+1}\|^2). \tag{5.9}$$

This would prove the Theorem, since then

$$\int_{\mathbb{R}^n} \left(\sum_{j=1}^{n+1} b_j \frac{\operatorname{div}(wp^j)}{w}\right)^2 w = \|\sum_{j=1}^{n+1} y_j\|^2 \le n(\sum_{j=1}^{n+1}\|y_j\|^2) = n\sum_{j=1}^{n+1} b_j^2 \int_{\mathbb{R}} \left(\frac{\operatorname{div}(wp^j)}{w}\right)^2 w.$$

Now, using Equations (5.4) and (5.8) we can see that then

$$J\left(\sum_{i=1}^{n+1} a_i X_i\right) \le n\sum_{j=1}^{n+1} b_j^2 J\left(\frac{1}{\sqrt{1-a_j^2}}\sum_{i\neq j} a_i X_i\right).$$

To prove this we use Lemma 5. We do this by introducing $n+1$ commuting orthogonal projections such that for every $1\le j\le n+1$, $T_1\cdots T_{n+1} y_j = 0$. At last we need to show $T_i y_i = y_i$ for each i. Since then by the lemma

$$\|T_1 y_1+\cdots+T_{n+1} y_{n+1}\|^2 \le n\left(\|y_1\|^2+\cdots+\|y_{n+1}\|^2\right).$$

and because of the last step this is equal to

$$\|y_1+\cdots+y_{n+1}\|^2 \le n\left(\|y_1\|^2+\cdots+\|y_{n+1}\|^2\right),$$

and we have shown Equation (5.9). For this we define $T_i : L_2(w)\to L_2(w)$ by

$$(T_i\phi)(x) = \int_{\mathbb{R}} \phi(x_1,\cdots,u_i,\cdots,x_{n+1}) f_i(u_i)\,du_i.$$

Essentially integrating out the i^{th} coordinate against f_i. By Lemma 6 these are $n+1$ commuting orthogonal projections. Next, we show that for every $1\le j\le n+1$, $T_1\cdots T_{n+1} y_j = 0$. We can see that if we take all T_i over an arbitrary function ϕ that then $T_1\cdots T_{n+1}\phi =$

$$\int_{\mathbb{R}^{n+1}} \phi(u_1,\cdots,u_{n+1}) f_1(u_1)\cdots f_{n+1}(u_{n+1})\,du_1\ldots du_{n+1} = \int_{\mathbb{R}^{n+1}} \phi(u_1,\cdots,u_{n+1}) w\,du_1\ldots du_{n+1}. \tag{5.10}$$

We are integrating over the entire domain, so we can leave out the the variables. Now in our case we use y_i as ϕ, so filling this in Equation (5.10) gives

$$T_1 \cdots T_{n+1} y_i = \int_{\mathbb{R}^{n+1}} y_i \, w = \int_{\mathbb{R}^{n+1}} b_j \frac{\operatorname{div}(w p^j)}{w} w = b_j \int_{\mathbb{R}^{n+1}} \operatorname{div}(w p^j) \tag{5.11}$$

We are now going to show this integral goes to 0 by using Green-Gauss. This follows the same strategy as used earlier in this thesis when Equation (4.2) was proved. Therefore the usage here will not be described in detail. Let $B(r)$ be a ball in $n+1$ dimensions with radius r, $S(r)$ be the boundary of this ball and $N(x)$ the outward pointing normal vector, then

$$\int_{\mathbb{R}^{n+1}} \operatorname{div}(w p^j) = \lim_{r\to\infty} \int_{B(r)} \operatorname{div}(w p^j) = \lim_{r\to\infty} \int_{S(r)} w p^j \cdot N(x) \leq \lim_{r\to\infty} \int_{S(r)} \| w p^j \| = 0. \tag{5.12}$$

In the last step we used that by definition $\int_{\mathbb{R}^{n+1}} w \| p^j \| < \infty$ and that thus the integral becomes 0 if we take the limit to infinity. Combining Equations (5.11) and (5.12) gives what we want, that $T_1 \cdots T_{n+1} y_i = 0$ for each j. To complete the proof we need to show that $T_i y_i = y_i$.

$$T_i y_i = \int_{\mathbb{R}} y_i(x_1, \cdots, u_i, \cdots, x_{n+1}) f_i(u_i) \, du_i = y_i \int_{\mathbb{R}} f_i(u_i) \, du_i = y_i.$$

Here, we used that each y_i is specificaly constructed to be independent of the i^{th} coordinate, so we can just take it out of the integral and use that f_i is a density. This proves Equation (5.9) and therefore the entire Theorem. □

5.3. First supporting lemma

To prove the first supporting lemma, we actually need another general result about Hilbert spaces. This is summarised in the lemma below.

Lemma 4. *Let P_1 and P_2 be commuting orthogonal projections in a Hilbert space H. Then an orthogonal decomposition of H can be made with*

$$H = \bigoplus_{\varepsilon \in \{0,1\}^2} H_\varepsilon, \quad H_\varepsilon = \{x : P_1 x = \varepsilon_1 x, P_2 x = \varepsilon_2 x\}.$$

Proof of Lemma 4
We start with the use of Equation 2.7 from Conway [5] to see that an orthogonal projection, call it P_1, projects vectors in the hilbert space to a linear closed subspace, call it M_1, such that $\operatorname{Ran}(P_1) = M_1$. Furthermore, from Conway we see that $\operatorname{Ker}(P_1) = M_1^\perp = \operatorname{Ran}(1 - P_1)$, where $(1 - P_1)$ is also an orthogonal projection. We see that we can write H as a direct sum of M_1 and $M_1^\perp$. This is because for every $x \in H$ we can write it as $x = P_1 x + (1 - P_1)x$. Here, by definition $P_1 x \in M_1$, since it is in the range of the orthogonal projection P_1 and $(1 - P_1)x \in M_1^\perp$, since it is in the range of the orthogonal projection $(1 - P_1)$.

But we can write it even cleaner, as the direct sum of H_ε with $\varepsilon \in \{0, 1\}$ and $H_\varepsilon = \{x : P_1 x = \varepsilon x\}$. This can be done because $H_1 = M_1$ and $H_0 = M_1^\perp$. Since H_0 is exactly the definition of the kernel of P_1 and for every $x \in M_1$, $P_1 x = x$, since P_1 sends the x to M_1, but it is already in M_1. Now we continue with showing it for 2 orthogonal projections.

We can write every $x \in H$ as $x = P_1 P_2 x + P_1(1 - P_2)x + (1 - P_1)P_2 x + (1 - P_1)(1 - P_2)x$, thus following the same arguments we can see that this H is a direct sum like

$$H = \bigoplus_{\varepsilon \in \{0,1\}^2} H_\varepsilon, \quad H_\varepsilon = \{x : P_1 x = \varepsilon_1 x, P_2 x = \varepsilon_2 x\}$$

This decomposition can only be performed if the projections P_1 and P_2 are commuting orthogonal projections. For example, the term $P_1 P_2 x$ is intended to correspond to the subspace $H_{(1,1)}$, which equals $\operatorname{Ran}(P_1) \cap \operatorname{Ran}(P_2)$. However, for $P_1 P_2 x$ to actually lie in both $\operatorname{Ran}(P_1)$ and $\operatorname{Ran}(P_2)$, we require that

$$P_1(P_2 x) = P_2(P_1 x),$$

i.e., the projections must commute. Without commutativity, it is possible that $P_2 x \in \mathrm{Ran}(P_2)$, but applying P_1 may move it out of $\mathrm{Ran}(P_2)$. In this case, $P_1 P_2 x \notin \mathrm{Ran}(P_2)$, so it does not lie in the intersection $\mathrm{Ran}(P_1) \cap \mathrm{Ran}(P_2)$.
Now we need to show each subspace is orthogonal. Take ε and ε', such that $\varepsilon \neq \varepsilon'$, then take $x \in H_\varepsilon$ and $y \in H_{\varepsilon'}$, then we need to show that $\langle x, y\rangle = 0$. By definition, there exists some j such that $\varepsilon_j \neq \varepsilon'_j$ and the fact that each P_i is an orthogonal projection (such that $\langle P_i x, y\rangle = \langle x, P_i y\rangle$), we see that

$$\langle x, y\rangle = \langle P_j x, y\rangle = \langle \varepsilon_j x, y\rangle = \varepsilon_j \langle x, y\rangle \tag{5.13}$$

and that

$$\langle x, y\rangle = \langle x, P_j y\rangle = \langle x, \varepsilon'_j y\rangle = \varepsilon'_j \langle x, y\rangle. \tag{5.14}$$

In the last two equations we have used that $\varepsilon_j x = P_j x$ and $\varepsilon'_j y = P_j y$, which follows from the definitions that $x \in H_\varepsilon$ and $y \in H_{\varepsilon'}$. Combining Equations (5.13) and (5.14) we see that $\langle x, y\rangle = 0$, since $\varepsilon_j \neq \varepsilon'_j$. This shows that all H_ε are orthogonal subspaces. □

Now we have the result we needed to prove the first supporting lemma, which we do in the proof below.

Lemma 5. *Let $T_1, \ldots, T_m$ be m commuting orthogonal projections in a Hilbert space H. Assume that we have m vectors $y_1, \ldots, y_m$ such that for every $1 \leq j \leq m$, $T_1 \cdots T_m y_j = 0$. Then*

$$\|T_1 y_1 + \cdots + T_m y_m\|^2 \leq (m-1)\left(\|y_1\|^2 + \cdots + \|y_m\|^2\right).$$

Proof of Lemma 5
We first show that we can decompose the Hilbert space with the commuting projections into an orthogonal decomposition. For this we use Lemma 4 and extend the argument for m commuting orthogonal projections instead of 2. It can be seen that then we have an orthogonally decomposition of H with

$$H = \bigoplus_{\varepsilon \in \{0,1\}^m} H_\varepsilon, \quad H_\varepsilon = \{x : T_i x = \varepsilon_i x, 1 \leq i \leq m\}$$

Since this is an orthogonal decomposition, we can write each $\phi \in H$ as $\sum_{\varepsilon \in \{0,1\}^m} \phi_\varepsilon$, where each $\phi_\varepsilon \in H_\varepsilon$. We use orthogonality to show

$$\|\phi\|^2 = \|\sum_{\varepsilon \in \{0,1\}^m} \phi_\varepsilon\|^2 = \langle \sum_{\varepsilon \in \{0,1\}^m} \phi_\varepsilon, \sum_{\varepsilon \in \{0,1\}^m} \phi_\varepsilon\rangle = \sum_{\varepsilon \in \{0,1\}^m} \|\phi_\varepsilon\|^2.$$

Here, we used that every ϕ_ε is orthogonal to each other and thus that for $\phi_{\varepsilon 1} \neq \phi_{\varepsilon 2}$, $\langle \phi_{\varepsilon 1}, \phi_{\varepsilon 2}\rangle = 0$. We do this for each y_i and write it as $y_i = \sum_{\varepsilon \in \{0,1\}^m} y_\varepsilon^i$. Then we can write

$$T_1 y_1 + \cdots + T_m y = \sum_{i=1}^m \sum_{\varepsilon \in \{0,1\}^m} T_i y_\varepsilon^i = \sum_{\varepsilon \in \{0,1\}^m} \sum_{i=1}^m T_i y_\varepsilon^i.$$

In the last step we just switched the sums. Now the second sum can be rewritten, $\sum_{i=1}^m T_i y_\varepsilon^i$. Since $T_i y_\varepsilon^i = \varepsilon_i y_\varepsilon^i$ and remember that $\varepsilon_i = 1$ or $\varepsilon_i = 0$, we see that $T_i y_\varepsilon^i$ only adds something to the sum if $\varepsilon_i = 1$. Thus,

$$\sum_{\varepsilon \in \{0,1\}^m} \sum_{i=1}^m T_i y_\varepsilon^i = \sum_{\varepsilon \in \{0,1\}^m} \sum_{\substack{i=1,\\ \varepsilon_i = 1}}^m T_i y_\varepsilon^i = \sum_{\varepsilon \in \{0,1\}^m} \sum_{\substack{i=1,\\ \varepsilon_i = 1}}^m y_\varepsilon^i.$$

We use this and $\|\phi\|^2 = \sum_{\varepsilon \in \{0,1\}^m} \|\phi_\varepsilon\|^2$ to show that

$$\|T_1 y_1 + \cdots + T_m y_m\|^2 = \sum_{\varepsilon \in \{0,1\}^m} \|\sum_{\substack{i=1,\\ \varepsilon_i = 1}}^m y_\varepsilon^i\|^2 \tag{5.15}$$

Now we use that in a Hilbert space

$$\|\sum_{i=1}^m u_i\|^2 \leq m \sum_{i=1}^m \|u_i\|^2, \tag{5.16}$$

which we will prove in a bit. Then by assumption in the Lemma $T_1, \cdots T_m y_j = 0$. This gives that if all $\varepsilon_i = 1$, then it does not add anything to the sum. This can be seen from the construction of each H_ε. Remember that

in the construction y_ε^i with $\varepsilon = 1$ is the part with $T_1 \cdots T_m y_i$, which was 0 by the assumption. So we can see that every vector on the right-hand side of Equation (5.15) is a sum of at most $m-1$ summands and using Equation (5.16) we see that

$$\|T_1 y_1 + \cdots + T_m y_m\|^2 \leq \sum_{\varepsilon \in \{0,1\}^m} (m-1) \sum_{\substack{i=1,\\ \varepsilon_i = 1}}^{m} \|y_\varepsilon^i\|^2 = (m-1) \sum_{i=1}^{m} \|y_i\|^2.$$

In the last step we switched the sums again and transformed back to the original vectors. This proves the Lemma, so we only need to prove Equation (5.16).

$$\|\sum_{i=1}^{m} u_i\|^2 = \langle \sum_{i=1}^{m} u_i, \sum_{j=1}^{m} u_j \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} \langle u_i, u_j \rangle = \sum_{i=1}^{m} \|u_i\|^2 + \sum_{i \neq j} \langle u_i, u_j \rangle \leq \sum_{i=1}^{m} \|u_i\|^2 + \sum_{i \neq j} \|u_i\| \|u_j\|$$

We next use the AM-GM inequality to show

$$\sum_{i=1}^{m} \|u_i\|^2 + \sum_{i \neq j} \|u_i\| \|u_j\| \leq \sum_{i=1}^{m} \|u_i\|^2 + \frac{1}{2} \sum_{i \neq j} (\|u_i\|^2 + \|u_j\|^2)$$

The AM-GM inequality is

$$\frac{x_1 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 \cdots x_n}$$

and we use it with $n = 2$ and $x_1 = \|u_i\|$ and $x_2 = \|u_j\|$. Next, we count how many times each u_i appears in the second sum. It appears $m-1$ times for each $i \neq j$, but since we also need to count $j \neq i$, it is counted $2(m-1)$ times. This gives

$$\|\sum_{i=1}^{m} u_i\|^2 \leq \sum_{i=1}^{m} \|u_i\|^2 + \frac{1}{2} \sum_{i \neq j} (\|u_i\|^2 + \|u_j\|^2) = \sum_{i=1}^{m} \|u_i\|^2 + \frac{2(m-1)}{2} \sum_{i=1}^{m} \|u_i\|^2 = m \sum_{i=1}^{m} \|u_i\|^2$$

which proves the lemma. □

5.4. Second supporting lemma

Lemma 6. *Let $X_1, \cdots, X_{n+1}$ be independent random variables with corresponding densities $f_1(x_1), \cdots, f_{n+1}(x_{n+1})$ with w the product density, then $T_i : L_2(w) \to L_2(w)$ defined by*

$$(T_i \phi)(x) = \int_{\mathbb{R}} \phi(x_1, \cdots, u_i, \cdots, x_{n+1}) f_i(u_i) du_i$$

are commuting orthogonal projections.

Proof of Lemma 6:
To show each T_i is orthogonal, we have to show that $T_i^2 \phi = T_i \phi$ and $\langle T_i \phi_1, \phi_2 \rangle = \langle \phi_1, T_i \phi_2 \rangle \ \forall \phi, \phi_1, \phi_2 \in L_2(w)$. We start with $T_i^2 \phi = T_i \phi$:

$$T_i^2 \phi = T_i \int_{\mathbb{R}} \phi(x_1, \cdots, u_i, \cdots, x_{n+1}) f_i(u_i) du_i = \int_{\mathbb{R}} \int_{\mathbb{R}} \phi(x_1, \cdots, u_i, \cdots, x_{n+1}) f_i(u_i) du_i f_i(v_i) dv_i,$$

but the function over which we take T_i the second time, the inner integral, was already independent of the i^{th} coordinate, so we can take it out of the outer integral and using the fact that f_i is a density we get

$$T_i^2 \phi = \int_{\mathbb{R}} \phi(x_1, \cdots, u_i, \cdots, x_{n+1}) f_i(u_i) du_i \int_{\mathbb{R}} f_i(u_i) du_i = \int_{\mathbb{R}} \phi(x_1, \cdots, u_i, \cdots, x_{n+1}) f_i(u_i) du_i \cdot 1 = T_i \phi.$$

Next, we show that $\langle T_i \phi_1, \phi_2 \rangle = \langle \phi_1, T_i \phi_2 \rangle$, we use here that $d\mathbf{x}_{n+1} = dx_1 \ldots x_{n+1}$ for notation purposes.

$$\langle T_i \phi_1, \phi_2 \rangle = \int_{\mathbb{R}^{n+1}} (T_i \phi_1) \cdot \overline{\phi_2} \cdot w \, d\mathbf{x_{n+1}} = \int_{\mathbb{R}^{n+1}} \int_{\mathbb{R}} \phi_1(x_1, \cdots, u_i, \cdots, x_{n+1}) f_i(u_i) du_i \overline{\phi_2(x_1, \cdots, x_{n+1})} w d\mathbf{x_{n+1}}.$$

By using the fact that w is a product density and decomposing the domain of the integral over $\mathbb{R}^{n+1}$ into $\mathbb{R}^n$ and $\mathbb{R}$, where we integrate over the variable x_i we see that this is equal to

$$\langle T_i \phi_1, \phi_2 \rangle = \int_{\mathbb{R}^n} \prod_{j \neq i} f_j(x_j) \left(\int_{\mathbb{R}} \phi_1(x_1, \cdots, u_i, \cdots, x_{n+1}) f_i(u_i) du_i \int_{\mathbb{R}} \overline{\phi_2(x_1, \cdots, x_i \cdots, x_{n+1})} f_i(x_i) dx_i \right) d\mathbf{x_n}. \quad (5.17)$$

Now we show that we can decompose $\langle \phi_1, T_i \phi_2 \rangle$ in the same way:

$$\langle \phi_1, T_i \phi_2 \rangle = \int_{\mathbb{R}^{n+1}} \phi_1 \cdot (\overline{T_i \phi_2}) \cdot w \, d\mathbf{x_{n+1}} = \int_{\mathbb{R}^{n+1}} \phi_1(x_1, \cdots, x_{n+1}) \int_{\mathbb{R}} \overline{\phi_2(x_1, \cdots, u_i, \cdots, x_{n+1})} f_i(u_i) du_i dw \mathbf{x_{n+1}}.$$

Following the same arguments as before, we show that this is equal to

$$\langle \phi_1, T_i \phi_2 \rangle \int_{\mathbb{R}^n} \prod_{j \neq i} f_j(x_j) \left(\int_{\mathbb{R}} \phi_1(x_1, \cdots, x_i, \cdots, x_{n+1}) f_i(x_i) dx_i \int_{\mathbb{R}} \overline{\phi_2(x_1, \cdots, u_i, \cdots, x_{n+1})} f_i(u_i) du_i \right) d\mathbf{x_n}. \quad (5.18)$$

Combing Equations (5.17) and (5.18) shows that $\langle T_i \phi_1, \phi_2 \rangle = \langle \phi_1, T_i \phi_2 \rangle$. This is because in both equations x_i and u_i are just dummy variables who are integrated out of the function. At last we need to show that they commute, for this we show that $T_i T_j \phi = T_j T_i \phi$. W.l.o.g. we assume $i \leq j$,then

$$T_i T_j \phi = \int_{\mathbb{R}} \int_{\mathbb{R}} \phi(x_1, \cdots, u_i, \cdots, u_j, \cdots, x_{n+1}) f_j(u_j) f_i(u_i) du_j du_i.$$

Following the same arguments we can show that $T_j T_i \phi =$

$$T_j T_i \phi = \int_{\mathbb{R}} \int_{\mathbb{R}} \phi(x_1, \cdots, u_i, \cdots, u_j, \cdots, x_{n+1}) f_i(u_i) f_j(u_j) du_i du_j.$$

But these last two expressions are just equal, since we can switch the integrals. This shows $T_i T_j \phi = T_j T_i \phi$ which was necessary to show they commute. □

6

Conclusion

The conclusion of this thesis will be mostly about pointing out where compared to the original article [1] new mathematical findings were added. Furthermore it will explain the few parts that are still missing and could be improved.

First of all, the structure of the proof in this thesis differs from the article. The part about the connection between the Fisher information and the entropy in Chapter 3 is structured a lot more clearly to show what is left to prove, while the article keeps this more condensed and expects the reader to understand it. In the actual proof, the theorems are split up and more lemmas are introduced to keep the structure nice to read, while the original article only uses one supporting lemma and only one theorem. These are not actual mathematical findings, but these do add to the flow for the reader and keeping it accessible for students who are not specialised in this area.

In Chapter 4 a lot more mathematical details and findings are added. For example Lemma 1 is described in one paragraph in the original article, while the rigorous proof to actually understand it takes one page. This can also be seen in the proof of Theorem 5. The eventual p which is constructed in this theorem is only described vaguely in one paragraph, but it is difficult to understand what each component of the function it actually is. Constructing p rigorously takes more than an entire page in the proof in this thesis.

In Chapter 5 the same happens, a lot more mathematical details and findings are added in the proofs. In the proof of Theorem 6 the article is especially quick with its calculations. Taking the integral in Equation (5.4) actually has the wrong domain where the integral is taken over, namely $\mathbb{R}^n$ instead of $\mathbb{R}^{n+1}$. This is probably, because the authors think this is easy to see, but explaining that we can reduce one dimension actually takes another page, while the article does it in a few sentences. Especially when we take into account the part where the article says it can be assumed that p^j does not depend on the j^{th} coordinate and its own j^{th} coordinate is 0. The last part where a lot of new mathematical findings were added is in Lemma 5, but this makes more sense. The authors probably presume that the reader has a lot of experience in Hilbert spaces, while a bachelor student has little to no experience with it. This is the reason this needs a lot of added explanation in its proof.

There are also still some small gaps in this thesis. If we go to the proof of Theorem 5 the article mentions that each component of pw tends to 0 at infinity. This is vaguely proved in this thesis, but a more rigorous proof would be even better. A relative bigger gap in this thesis is that in the proof of Theorem 6 it is only proved for random variables for which the product density w satisfies the assumptions of Theorem 4. The assumption that $\|\operatorname{Hess}(w)\|$ is integrable is not proved. It is kind of logical and is not used in the proof, thus the reason why it was seen to be reasonable to omit this assumption. The proof of the other assumptions is not mentioned at all in the article however, so these parts can be seen as entirely new mathematical findings.

Bibliography

[1] Shiri Artstein, Keith M. Ball, Franck Barthe, and Assaf Naor. Solution of shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4):975–982, 2004. doi: 10.1090/S0894-0347-04-00459-X.

[2] D. Bakry and M. Emery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX*, volume 1123 of *Lecture Notes in Mathematics*, pages 179–206. Springer, 1985.

[3] A. R. Barron. Entropy and the central limit theorem. *Ann. Probab.*, 14:336–342, 1986.

[4] E. A. Carlen and A. Soffer. Entropy production by block variable summation and central limit theorems. *Commun. Math. Phys.*, 140(2):339–371, 1991.

[5] John B. Conway. *A Course in Functional Analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer, New York, 2nd edition, 1990. ISBN 978-0-387-97245-5.

[6] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL, 1949. MR11:258e. See the section on "Entropy of a Sum of Two Ensembles" for the case where the entropy of the normalized sum of two random variables exceeds that of a single variable.

[7] Joram Soch. Normal distribution maximises entropy, 2020. URL `https://statproofbook.github.io/P/norm-maxent.html`. Accessed: 2025-06-02.

[8] A. J. Stam. Some inequalities satisfied by the quantities of information of fisher and shannon. *Info. Control*, 2:101–112, 1959.

[9] Eric T. H. Wang. Neumann–shannon anecdote. `https://www.eoht.info/page/Neumann-Shannon%20anecdote`, n.d. URL `https://www.eoht.info/page/Neumann-Shannon%20anecdote`. Accessed: 2025-06-19.

A

Increasing entropy plot

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import gaussian_kde,norm
from scipy.integrate import simpson
# Settings
np.random.seed(0)
n_samples = 100_000
max_n = 10

entropies = []

# Loop over 1 to 10 summed variables
for n in range(1, max_n + 1):
    # Generate n independent Exponential(1) random variables
    Xs = np.random.exponential(scale=1.0, size=(n, n_samples))    # Normalised sum
    Y_n = np.sum(Xs, axis=0) / np.sqrt(n)

    # Estimate density using KDE
    kde = gaussian_kde(Y_n)
    z_vals = np.linspace(min(Y_n), max(Y_n), 1000)
    f_z_vals = kde(z_vals)

    # Compute entropy
    integrand = -f_z_vals * np.log(f_z_vals)
    #numerically approximate the integral
    entropy = simpson(integrand, z_vals)
    entropies.append(entropy)

z_vals_gauss = np.linspace(-6, 6, 1000)
f_gauss = norm.pdf(z_vals_gauss)
integrand_gauss = -f_gauss * np.log(f_gauss)
gaussian_entropy = simpson(integrand_gauss, z_vals_gauss)

# Plot entropy vs number of summed variables
plt.plot(range(1, max_n + 1), entropies, marker='o')
plt.axhline(y=gaussian_entropy, color='red', linestyle='--', label="Entropy of Standard Gaussian (numer
plt.title("Entropy of Normalized Sums of i.i.d. Exponential Variables")
plt.xlabel("Number of Summed Variables (n)")
plt.ylabel("Entropy of Yn")
plt.grid(True)
```

```
plt.show()
```