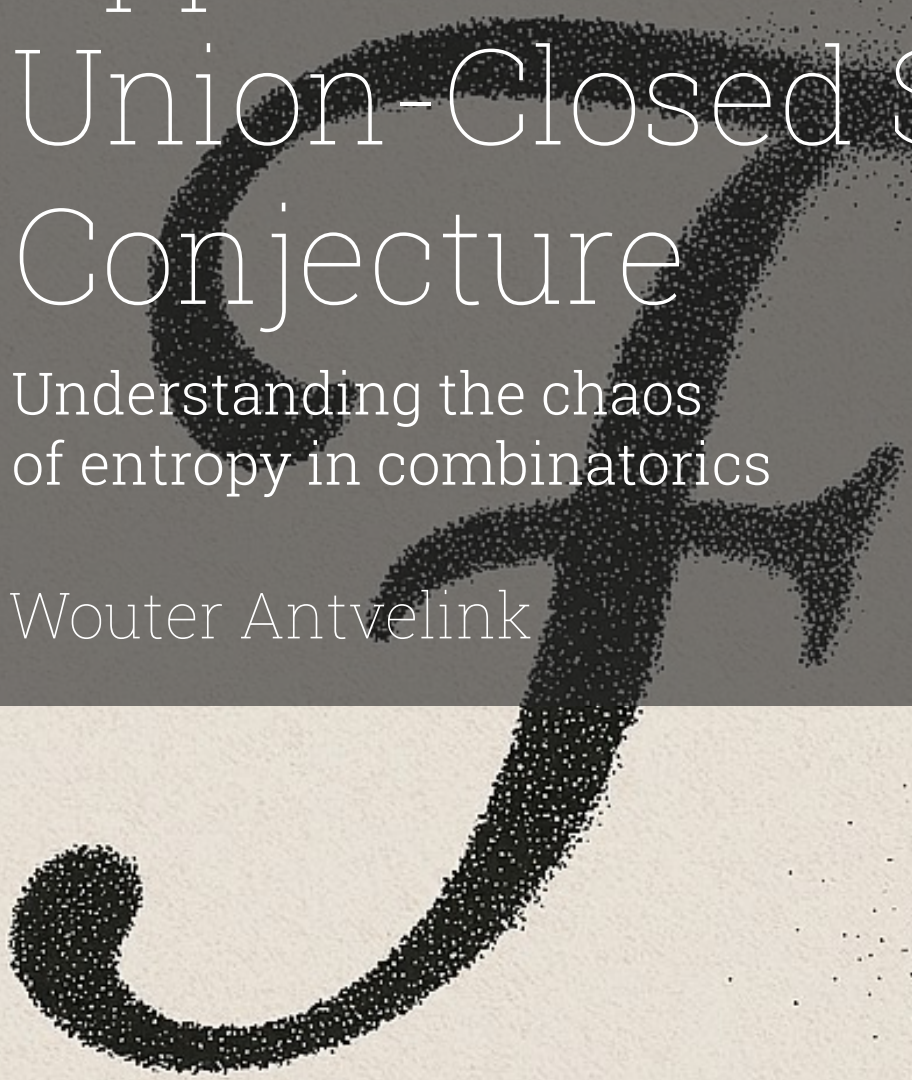
An Entropy-Based Approach to the Union-Closed Sets Conjecture

Understanding the chaos of entropy in combinatorics

Wouter Antvelink



Delft University of Technology

TU Delft

Understanding the chaos of entropy in combinatorics

by

Wouter Antvelink

Supervisors:	Anurag Bishnoi, Barbara Terhal
Second Supervisors:	Rik Versendaal, Yaroslav Blanter
Project Duration:	March 2025–July 2025
Faculty:	Faculty of Mathematics and Faculty of Physics, Delft

Cover:	Visualization of entropy-based dispersion of the mathematical symbol of a family of sets, generated with OpenAI's ChatGPT on July 1, 2025
Style:	TU Delft Report Style, with modifications by Daan Zwaneveld
URL:	`http://repository.tudelft.nl/`.

TU Delft

Preface

This thesis was written as part of my undergraduate studies in mathematics and physics, with the aim of exploring a modern, information-theoretic approach to one of combinatorics' most enduring open problems: the Union-Closed Sets Conjecture. It is intended for undergraduate students from both mathematics and physics and for those curious about how tools from information theory, such as entropy, can be fruitfully applied outside their traditional domains.

My initial interest in this topic arose from a fascination with how a concept like entropy, which I originally encountered in my physics studies, could yield concrete results for a long-standing mathematical problem. As a student of both physics and mathematics, I have often seen mathematical ideas applied to physics, but it is far less common to see concepts from physics provide new insights into mathematics. This reversal made the topic all the more compelling to explore.

Throughout the writing process, I learned not only about entropy and Union-Closed families, but also about how modern mathematical research progresses and what it takes to structure and carry out such a complex project.

I would like to thank Anurag Bishnoi for his guidance, insightful discussions, and for encouraging me to pursue this topic in the first place. I am also grateful to Barbara Terhal for her helpful suggestions on how to make this thesis more approachable for students with a background in physics.

Wouter Antvelink
Delft, July 2025

Abstract

The Union-Closed Sets Conjecture (UCSC), posed by Peter Frankl in 1979, asserts that every finite union-closed family of sets contains an element that appears in at least half of its member sets. Despite this seemingly simple formulation, the conjecture has remained unresolved for decades. Recently in November 2022 Gilmer developed a novel approach based on the information theoretic concept: *entropy*, which has offered fresh insights and promising partial results. In this thesis, we provide a self-contained introduction to entropy and explore its utility in combinatorics, specifically its application to the UCSC. We thoroughly examine the groundbreaking entropy-based proof that establishes a constant lower bound of $\frac{3-\sqrt{5}}{2} \approx 0.382$. Additionally, we shortly discuss the small further improvements made to this bound. We then consider a related conjecture proposed by Nagel, which states: *the kth most frequent element in a union-closed family appears in at least a fraction $\frac{1}{2^{k-1}+1}$ of the sets.* We shall discuss how the new entropy approach could also be used there to improve the bounds on the sizes of the family. All these explorations have been focused on understanding recent developments and to create a unified narrative. Due to time constraints, the thesis did not pursue new results. However, the comprehensive understanding gained has given some promising directions for future research.

Contents

Preface i

Abstract ii

1 Introduction 1

2 Preliminaries 3
- 2.1 Elementary definitions . . . 3
  - 2.1.1 Union-closed families of sets . . . 3
  - 2.1.2 Characteristic vector of a set . . . 4
  - 2.1.3 Random variable . . . 4
- 2.2 Notation . . . 5

3 Entropy in Combinatorics 7
- 3.1 What is Entropy? . . . 7
- 3.2 The intuition . . . 8
  - 3.2.1 The expected Surprise . . . 8
  - 3.2.2 Information transfer and the surprisal . . . 9
  - 3.2.3 Entropy examples . . . 11
  - 3.2.4 Conditional and joint entropy . . . 12
- 3.3 Properties . . . 14
- 3.4 Entropy in Combinatorics . . . 18
  - 3.4.1 Binomial sum bound . . . 19
  - 3.4.2 The Bregman-Minc Inequality . . . 20
  - 3.4.3 A connecting theme? . . . 23
- 3.5 Why entropy? . . . 23
  - 3.5.1 Connection between entropy and counting . . . 23
  - 3.5.2 Other useful properties . . . 24
  - 3.5.3 Is entropy unique? . . . 24
  - 3.5.4 Conclusion . . . 25

4 Probabilities and the Union-Closed Set Conjecture 26
- 4.1 the Conjecture . . . 26
  - 4.1.1 Examples from the conjecture . . . 26
  - 4.1.2 Finite families . . . 28
- 4.2 Entropy and sets . . . 28

5 Proving A constant bound of for the Union-Closed Sets Conjecture 32
- 5.1 The structure of the proof . . . 32
- 5.2 Proof of the main theorem . . . 33
- 5.3 The proof of Lemma 5.3 . . . 35
  - 5.3.1 Lemma from Chase and Lovett . . . 36

5.3.2 The lemma from Boppana . . . 37

6 Further improvements on the lower bound and Sawin 42
6.1 Further improvements on the lower bound . . . 42
6.2 Sawin . . . 42
6.2.1 Sawin's idea for further improvement . . . 42
6.3 Notation and preliminaries . . . 43
6.4 lemma 3 from Sawin . . . 44
6.4.1 Overview and idea . . . 44
6.4.2 The start of the proof . . . 45
6.4.3 If *mu* minimizes *C(mu)* it also minimizes *D(mu)* . . . 45
6.4.4 Analysing *Fmu(q)* . . . 46
6.4.5 The strictly concave section *(a,1)* . . . 48
6.4.6 The convex section [0,a] . . . 49
6.4.7 Combining the sections . . . 49
6.4.8 Recap and back to equation . . . 49
6.4.9 Analysing *w* and *v*. . . . 50

7 The kth most frequent element 52
7.1 Notation . . . 52
7.2 The *k*th most frequent conjecture . . . 52
7.2.1 A tight bound . . . 53
7.2.2 The Union-Closed Sets Conjecture implies the kth most frequent conjecture . . . 53
7.3 How Das and Wu use the entropy method for the kth most frequent conjecture . . . 54

8 Conclusion and Further Research 57
8.1 Further Research . . . 57

References 59

A Additional proofs 61
A.1 Limit of *xlog(x)* . . . 61
A.2 Entropy of fully dependent random variables . . . 61
A.3 proof of the concavity of entropy . . . 62
A.4 Jensen's inequality . . . 63
A.5 The limit of *f(x,y)* along the corner points of the boundary . . . 64
A.6 *G(x)* is strictly increasing . . . 65
A.7 the reverse direction Rolle's theorem . . . 65
A.8 The roots of *v* . . . 66
A.9 Probability distribution on a strictly convex section *[0,a]* . . . 67

B The proof that the logarithmic form of entropy is characterized by four basic properties 68

1

Introduction

Combinatorics, at its core, is the study of discrete structures and their enumeration, and that might feel far away from probabilities. However, over the past few decades, a remarkable tool from information theory, entropy, has become a powerful method for proving combinatorial results. It has been used to simplify existing arguments and proof new results for longstanding conjectures. This thesis is motivated by the relatively surprising effectiveness of entropy-based techniques in problems in combinatorics, with a particular focus on extremal set theory.

One such longstanding conjecture in extremal set theory is the **Union-Closed Sets Conjecture** (UCSC), posed by Peter Frankl in 1979:

> ***If $\mathcal{F}$ is a finite union-closed family of sets, and it isn't the family containing only the empty set, then there exists an element that belongs to at least half the sets in $\mathcal{F}$.***

Despite its seemingly simple formulation, the conjecture has resisted proof for over four decades and remains one of the most famous open problems in the field.

Traditional approaches to the conjecture rely on complicated combinatorial arguments, but those have never led to any constant bound (i.e. the most frequent element is in the fraction α of all the sets in the family). For an excellent overview of many of those approaches, see [7]. However, recently, entropy has offered a fresh perspective. The work of Gilmer (2022) [16], transformed the problem into one of bounding the entropy of certain random variables associated to sets in the union-closed family. This shift has led to many new partial results.[2, 19, 8, 12, 10].

The goal of this thesis is to present a self-contained introduction to entropy and its combinatorial applications for relative newcomers. And to explain in detail the new entropy-approach to the Union-Closed Sets Conjecture. Such that the reader can understand the proof for a new bound on said conjecture. Another initial goal was to improve on Das and Wu [12] by applying structures to increase the bound on the Union-Closed Sets Conjecture, such as done by [8, 19, 31] to the related kth most frequent conjecture. However, this goal was never reached due to the time constrains of this project. Nonetheless, the notes from that research are included in this thesis.

Structure

The thesis is organized as follows: First we discuss some notation and elementary definitions in chapter 2, then we introduce the concept of entropy and discuss some properties followed by a short survey of some results in combinatorics that have been proven using entropy, such as the Bregman-Minc inequality. In chapter 4, we present the Union-Closed Sets Conjecture and the entropy formulation of it. After which we prove the $\frac{3-\sqrt{5}}{2}$ bound for the UCSC, following a line of reasoning similar to [10] and [1]. In the last part of the main body do we discuss our further exploration and notes on the proof by Das and Wu [12] and Sawin [28] in our failed attempt to apply the framework of Cambie [8] to the paper from Das and Wu. And then we conclude with a summary of this thesis and further research recommendations.

2

Preliminaries

This thesis is meant for a wider audience and assumes only a basic understanding of set theory. We will discuss most other concepts and their properties in detail. Therefore, we encourage more advanced readers to skip over parts that are already familiar to them. For the start of the discussion about entropy, we refer to Chapter 3.

2.1. Elementary definitions

In the formulation of the Union-Closed Sets Conjecture, we encounter the notion "union-closed family of sets". It is fitting therefore to first discuss these concepts. Afterwards, we will shortly discuss the characteristic vector of a set and the definition of a random variable.

2.1.1. Union-closed families of sets

Definition 2.1. Family of sets
A **Family** is a collection or set of sets that includes the empty set, typically denoted by $\mathcal{F}$, where each element of $\mathcal{F}$ is called a **member-set**. Formally, $\mathcal{F} \subseteq \mathcal{P}(U(\mathcal{F}))$, where $U(\mathcal{F}) = \cup_{A \in F}(A)$ is the *ground-set* or *universe* of $\mathcal{F}$ and $\mathcal{P}(U(\mathcal{F}))$ is its power set.

Example 2.2. If we have the three sets: $A = \{1, 2, 3\}$, $B = \{1, rood\}$ and $C = \{223, \square\}$ then a family $\mathcal{B}$ could be defined as: $\mathcal{B} = \{A, B, C\} = \{\emptyset, \{1, 2, 3\}, \{1, rood\}, \{223, \square\}\}$. And $U(\mathcal{B}) = \{1, 2, 3, rood, 223, \square\}$

Definition 2.3. A family $\mathcal{F}$ is **union-closed** if:
$\forall A, B \in \mathcal{F} \implies A \cup B \in \mathcal{F}$.

Example 2.4. The family $\mathcal{Z} = \Big\{\emptyset, \{1\}, \{2\}, \{1, 2\}, \{1, 2, 3, 4\}\Big\}$ is **union-closed** because for every two sets $A, B \in \mathcal{Z}$, their union ($A \cup B$) is an element of $\mathcal{Z}$: $\{1\} \cup \{2\} = \{1, 2\} \in \mathcal{Z}$ and $\{1, 2\} \cup \{1, 2, 3, 4\} = \{1, 2, 3, 4\} \in \mathcal{Z}$ and $\emptyset \cup \{2\} = \{2\} \in \mathcal{Z}$ etc.

Example 2.5. Is the family $\mathcal{X} = \Big\{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 2, 3\}\Big\}$ union-closed? (the answer is on the next page).

$\mathcal{X}$ is **not union-closed**, because $\{2\}, \{3\} \in \mathcal{X}$, however $\{2\} \cup \{3\} = \{2, 3\} \notin \mathcal{X}$.

2.1.2. Characteristic vector of a set

Definition 2.6. Let $U = \{u_1, u_2, \ldots, u_n\}$ be the finite ground set of a family with a fixed ordering of elements. For any subset $S \subseteq U$, the *characteristic vector* (or *indicator vector*) $\vec{x}_S \in \{0, 1\}^n$ is defined as follows:

$$(\vec{x}_S)_i \equiv \begin{cases} 1 & \text{if } u_i \in S, \\ 0 & \text{if } u_i \notin S. \end{cases} \tag{2.1}$$

Example 2.7. If we have the family $\mathcal{F}$:

$$\mathcal{F} = \{\emptyset, \{1\}, \{1, 2\}\} = \{\emptyset, A, B\},$$

then the ground-set of this family is $\{1, 2\}$. And the characteristic vectors are as follows:

$$\vec{x}_\emptyset = (0, 0)^T,$$
$$\vec{x}_A = (1, 0)^T,$$
$$\vec{x}_B = (1, 1)^T.$$

2.1.3. Random variable

Definition 2.8. A random variable X is defined as a measurable function $X : \Omega \to \mathbf{E}$, where Ω is part of the probability space $(\Omega, \mathcal{A}, P)$ and $\mathbf{E}$ is part of the measurable space $(\mathbf{E}, \mathcal{B})$ (see [13] for more information on these definitions). Random variables are often denoted as capital letters, such as X and Y.

For the purposes of this thesis, we can be a little less rigorous and narrow the concept of a random variable more down to: *A **random variable** X is a function that has an outcome space $\mathbf{E}$ where each element $x \in \mathbf{E}$ has a probability to be the outcome of X.*

The outcome space $\mathbf{E}$ is most often numerical, i.e. $\mathbf{E} = \{0, 1, 2, 3, ..., n\}$, but it can also be a binary vector space: $\mathbf{E} = \{0, 1\}^n$ or anything else. Random variables that take only two values will be referred to as **binary random variables**.

The probability that X takes on a specific value in this outcome space ($x \in E$) is written as:

$$p_x = p(x) = \Pr(X = x) = \Pr(\{\omega \in \Omega \mid X(\omega) \in S\}). \tag{2.2}$$

We note that we will use p_x and $p(x)$ interchangeably. And we will use μ or ν to refer to the general probability distribution over the outcome space.

When discussing the **conditional probability** of two random variables, we denote the probability that X takes the value x given that Y is known to be y as:

$$\Pr\big(X = x \mid Y = y\big) = p_{x|y}. \tag{2.3}$$

And when referring to **joint probability** we will write that as:

$$\Pr(X = x, Y = y) = p_{x,y}, \tag{2.4}$$

Which denotes the probability that X is x and Y is y simultaneously.

Example 2.9. Consider a fair coin toss, where the probabilities of obtaining heads or tails are both 50%. We assign the value 0 to the outcome "heads" and the value 1 to the outcome "tails". We can represent this situation with a random variable Y defined as a function from the sample space $\{0, 1\}$ to the outcomes 0 or 1, where 0 corresponds to heads and 1 to tails. Each outcome has equal probability.

A random variable with outcome space $\{0, 1\}$ is called a **Bernoulli random variable**. We denote this by writing

$$X \sim \text{Ber}(p(1)),$$

where $p(1) = \Pr(X = 1)$ is the probability of obtaining the outcome 1. Furthermore, we note that this definition is very similar to the notion of a *binary random variable*, and these two concepts can therefore be used interchangeably.

So in this example: $Y \sim \text{Ber}(0.5)$.

Example 2.10. We can also imagine that we sit on a sunny day in front of a café. And we are looking at the door of the café, curious who is going to come out of it next. Then our random variable could be defined as a function from the sample space: { *all the people in the bar at that time* } to the outcome: *any one person*. The probabilities could be different for all the people in the café. Maybe $Alex$ has just arrived, so he has a low probability to come out of the door next. While $Brenda$ has already finished her coffee, so she has a higher probability of coming out.

Example 2.11. *In this last example, we will hint to how random variables can be connected to the notion of a family of a set.* If we have the family $\mathcal{F}$, such as in Example 2.7:

$$\mathcal{F} = \{\emptyset, \{1\}, \{1, 2\}\}.$$

We now sample a set C uniformly at random from the family $\mathcal{F}$ and consider its characteristic vector $\vec{x}_C \in 0, 1^2$. This characteristic vector can be interpreted as a pair of Bernoulli random variables, (X_1, X_2), where $X_1 \sim \text{Ber}\left(\frac{2}{3}\right)$ and $X_2 \sim \text{Ber}\left(\frac{1}{3}\right)$. We also note that the distributions of X_i are not independent:

$$\Pr(X_2 = 1 | X_1 = 1) = \frac{1}{2} \neq \frac{1}{3} = \Pr(X_2 = 1).$$

Now that we have settled some definitions, we will specify certain notations:

2.2. Notation

Size of a Family
The number of sets in a family of sets $\mathcal{F}$ is denoted by $|\mathcal{F}|$. And we call a Family finite when $|\mathcal{F}| < \infty$.

Logarithm Notation
The notation $\log(x)$ refers to the base-2 logarithm, i.e., $\log_2(x)$.

Union of Random Variables

Throughout this thesis, when X and Y are random variables taking values in $0, 1^n$, we use the notation $X \cup Y$ to denote the componentwise maximum, i.e.,

$$(X \cup Y)_i = \max(X_i, Y_i). \tag{2.5}$$

This notation reflects the behaviour of a union between sets, when X and Y are interpreted as characteristic vectors of sets. We will discuss this notation more in Chapter 5.

Size of sets

We denote the number of elements in a set as $|S|$. A set is finite when $|S| < \infty$.

$[\boldsymbol{n}]$

The notation $[n]$ represents the set of the first n natural numbers:

$$[n] = \{1, 2, 3, \ldots, n\}.$$

List of random variables

When talking about the random variable representation of the characteristic vector of a set, we often denote the list of binary random variables as: $X = (X_1, X_2, \ldots, X_n)$. And let $X_{<i}$ denote $(X_1, X_2, ..., X_{i-1})$.

Entropy

With the word 'entropy' we will mean Shannon entropy, unless stated otherwise, such as 'physical entropy'.

3

Entropy in Combinatorics

Entropy is a fascinating concept with roots in many disciplines and applications in even more. In recent years, it has also gained a foothold in combinatorics, particularly in relation to the Union-Closed Sets Conjecture. Understanding entropy is therefore interesting not only for this specific conjecture, but also for exploring its potential applications in other areas of research.

In the following, we first define entropy and develop an intuition for it, after which we demonstrate its connection to set systems. We conclude this chapter with a discussion of the underlying theme in entropy-based bounds and why entropy fits naturally within combinatorics.

Since entropy relies on fundamental statistical concepts such as random variables, we refer the reader to the preliminaries (Chapter 2) and to [13] for further background. This chapter is written for students at the bachelor level; for more advanced readers or ambitious bachelor students, we recommend [15, 27].

"Nobody knows what entropy really is, so in a debate you will always have the advantage."

– John von Neumann (via Claude Shannon)

3.1. What is Entropy?

Shannon entropy originated in information theory, where it was introduced by Claude Shannon in his groundbreaking 1948 paper *A Mathematical Theory of Communication [30]*. During this time, Shannon was motivated by a central question: *how efficiently can information be encoded and transmitted over a communication channel?* He sought a mathematical measure of information from a source, and he realized quickly that 'surprise' and 'information' are integrally linked; when a message is more surprising it delivers more information and if the message is totally expected (0 surprise) then there is no information. He then realized that if a source emits symbols with known probabilities, then the 'surprise' and

thus 'information' of each symbol should depend on how unlikely it is. This led him to define entropy as the expected value of the information content (or 'surprisal') of outcomes of a random variable. And when he had to name it, he talked to John von Neumann, who encouraged him to draw inspiration from thermodynamics, where entropy was already widely associated with chaos and surprise.

Definition 3.1. For a discrete random variable X taking values in a finite set: $x \in \mathbf{E}$ with probability distribution $\Pr(X = x) = p_x$, the *Shannon* **entropy** of X is defined as:

$$\mathrm{H}(X) \equiv -\sum_{x \in \mathbf{E}} p_x \log p_x, \tag{3.1}$$

$$\equiv \mathbb{E}_{x \in E}[-\log(p_x)], \tag{3.2}$$

where the base-2 logarithm reflects a binary encoding and gives the unit *bits*. Some authors also use the natural log version of entropy, but this doesn't change the theorems as much as $\log(x) = \ln(x)/\ln(2)$. We will refer to entropy both in terms of 'bits' and as a unit-less quantity, depending on the context.

Shannon's entropy is one of the fundamental tools in information theory, and it is used in areas such as machine learning (e.g., decision trees and regularization), cryptography (e.g., randomness extractors), data compression (e.g., Huffman coding), and theoretical computer science (e.g., complexity theory and randomness).

But as this thesis shows: the relevance of entropy has far outgrown its original setting; it has found fertile ground in combinatorics. Over the past few decades, mathematicians have used entropy in a plurality of proofs as a powerful probabilistic technique for proving bounds on discrete structures. This approach often involves associating a random variable with a combinatorial object and then applying entropy inequalities to extract a bound.

In this chapter, we work to get an intuition for entropy and review its key properties. Then we will apply our intuition and knowledge to applications of entropy in combinatorics. This strong and intuitive foundation will then support our exploration of the new entropy based approach on the Union-Closed Sets Conjecture and its partial results (Chapter 4, Chapter 5).

3.2. The intuition

In the physical sciences, entropy is often described as a measure of disorder or uncertainty. Shannon entropy, while distinct, can be interpreted in a similar spirit: as a measure of the uncertainty inherent in a probability distribution. To illustrate this, I will explore entropy from two perspectives: first, the more intuitive notion of *surprise*, and then from an *information transfer* perspective, which helps to reveal the entropy expression.

3.2.1. The expected Surprise

One way to interpret Shannon entropy is as the expected value of "surprise", where the surprisal function $\log(1/p_x)$ represents the surprise associated with observing the outcome x. Notice that this surprisal increases as the probability $p_x \leq 1$ decreases, mirroring the intuition that rarer outcomes are more surprising.

Example 3.2. let's say we have a random variable X with two possible outcomes: x, y where $p_x = 0.9$ (and $p_y = 0.1$) the surprise when you would get x is low: $\log(1/0.9) \approx 0.15$. The surprise of getting y is

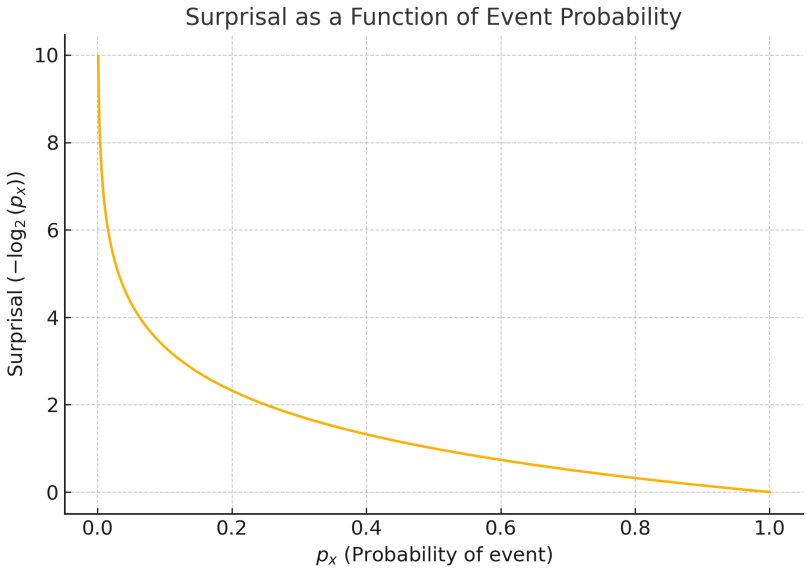


Figure 3.1: Surprisal as a function of event probability. We note that the "surprisal" of a certain event is 0, while the surprisal increases as the probability of the event decreases.

high: $\log(1/0.1) \approx 3.32$. The Entropy and the expectation of the surprise is thus:

$$\mathrm{H}(X) = \mathbb{E}[\log(1/p_x)] = (0.9) \cdot \log(1/0.9) \; + \; (0.1) \cdot \log(1/0.1) \approx 0.47.$$

3.2.2. Information transfer and the surprisal

Having understood entropy as the expected value of *surprisal*, we are led to ask: *what justifies the form of the surprisal function?* This is what we will now examine.

When Shannon thought of the entropy function, he was working to quantify *information* from a process. And when we follow that thought we will uncover the form of the surprisal. Let us start by considering a random variable X, and suppose we wish to communicate the outcome of X using bits.

Example 3.3. If X has 2^2 equally likely outcomes, such as *A, B, C* and *D*, we need only two bits to encode the outcome: *A* = '00', *B* = '10', *C* = '01' and *D* = '11'.

Remark 3.4. *if X has 2^m equally likely outcomes, then we need exactly m bits to represent any outcome, thus we will always send m bits on average.*

$$\mathrm{H}(X_{uniform}) = \mathbb{E}[\#\textbf{\textit{bits}}] = \log(2^m) = m. \tag{3.3}$$

However, Equation 3.3 only takes the amount of outcomes into account, not the probabilities of every outcome. To see why that is not enough, we discuss the next example and broaden our scope to include non-uniform distributions.

Example 3.5. Consider a random variable Z representing the outcome of an egg-and-spoon race between four siblings: the oldest, Hanna; the middle child, Clara; and the youngest twins, Tim and Oliver. The probabilities of each sibling winning the race are as follows: $p_{Hanna} = 0.6$, $p_{Clara} = 0.2$, and $p_{Tim} = p_{Oliver} = 0.1$.

The siblings' dad wants to send the result of the race to their mom over the internet. Being a bit of a geek, he wants to encode the outcome of Z as efficiently as possible, minimizing the number of bits he needs to transmit. He devises the following encoding scheme:

- If Hanna wins, send '0' (1 bit),
- If Clara wins, send '10' (2 bits),
- If Tim wins, send '110' (3 bits),
- If Oliver wins, send '111' (3 bits).

The expected number of bits required to transmit the outcome using this scheme is:

$$\begin{aligned}\mathbb{E}[\text{bits}] &= 0.6 \cdot 1 + 0.2 \cdot 2 + 0.1 \cdot 3 + 0.1 \cdot 3, \\ &= 0.6 + 0.4 + 0.3 + 0.3 = 1.6 \text{ bits}.\end{aligned}$$

We observe that, using this strategy, the dad only needs to send 1.6 bits on average. This is fewer than the 2 bits required in the case of a uniform distribution over four outcomes (Example 3.3).

Now we can calculate the entropy of this random variable:

$$\mathrm{H}(Z) = -(0.6)\log(0.6) - (0.2)\log(0.2) - 2 \cdot (0.1)\log(0.1) = 1.57 < 1.60 \text{ bits}.$$

Why is our calculated average of bits still higher than the entropy? That is what we will discuss next.

So the last example illustrated how the expected number of bits required depends on the probabilities of the outcomes, and it suggests that when the distribution is non-uniform, less information may be needed to convey the outcome of an event on average. However, we saw that our simple encoding method does not yet reach the same level of accuracy as the entropy function. The discrepancy between the entropy and the average bit rate per outcome in the example above (3.5) is because encoding schemes per random variable are non-optimal. We will illustrate this in the next example.

Example 3.6. Take a binary random variable X with an outcome space of only $\{a, b\}$ and $p(a) = 0.8$. Subsequently, $\mathrm{H}(X) \approx 0.722$. But if we want to encode one outcome of X, we still need 1 bit ("1" = a, "0" = b). Let us now look at 3 instances of this random variable: X_1, X_2, X_3, then there are 8 outcomes. Then we can derive a more optimal coding scheme (using the Huffman coding algorithm):

Sequence	bit string
aaa	1
aab	001
abb	00001
aba	010
bbb	00000
bba	00011
bab	00010
baa	011

Table 3.1: Example Huffman encoding for three instances of a binary random variable $\Pr(X = a) = 0.8$.

So the average bits per three instances of the random variable are:

$$1 \cdot (0.8^3) + 3 \cdot \big(3 \cdot (0.8^2 \cdot 0.2)\big) + 3 \cdot \big(5 \cdot (0.8 \cdot 0.2^2)\big) + 5 \cdot (0.2^3) \approx 3 \cdot 0.728.$$

We observe that, on average, only 0.728 bits are needed per instance of the random variable. This is already much closer to 0.722 than 1.

In fact, if we keep going with the scheme from Example 3.6, then we arrive in the limit at the entropy. In information theory, Shannon's source coding theorem informally states that:

Theorem 3.7 (Source Coding Theorem)**.** *[30, 20, 11] Let X be a discrete random variable with a finite outcome space. Then, for a sequence of n independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots, X_n$ drawn from X, the following holds:*

- *Any lossless source code must use at least $\mathrm{H}(X)$ bits per symbol on average.*
- *For any $\varepsilon > 0$, there exists a source coding scheme that compresses the source to $\mathrm{H}(X) + \varepsilon$ bits per symbol with vanishing probability of error as $n \to \infty$.*

In other words, for large n, $n\,\mathrm{H}(X)$ bits are both necessary and sufficient to encode the sequence $X_1, \ldots, X_n$.

Furthermore, we want to note that Shannon entropy can be characterized via multiple assumptions, and we have included in the appendix a proof of the logarithmic form of Shannon entropy, that follows from four basic properties (see the Appendix B).

3.2.3. Entropy examples

Now that we have gotten an understanding of Shannon entropy, it is important to discuss some examples to nurture our intuition.

Example 3.8 (Binary random variables)**.** A binary random variable, has two possible outcomes, i.e. $|\mathbf{E}| = 2$, and the outcomes occur with probabilities p and $1 - p$. The entropy can be calculated using the definition 3.1 and the entropy of a binary random variable with probability p is denoted by $h_2(p)$ and is called the *binary entropy*:

$$h_2(p) = p \log(1/p) + (1 - p) \log(1/(1 - p)). \tag{3.4}$$

We have plotted the binary entropy as a function of p down below. Furthermore, do we note that the entropy of a binary random variable and the entropy of a Bernoulli random variable are the same.

Example 3.9 (Certain Outcome (Zero Entropy))**.** Suppose the coin is rigged so that it always lands heads. Then there is no uncertainty and intuitively: there is no surprise. Thus, the entropy is zero:

$$\mathrm{H}(X) = \lim_{p \to 1} -[p \log(p) + (1 - p) \log(1 - p)] = 0,$$

where we can use l'Hopital's rule to show that $\lim_{x \to 0} x \log(x) = 0$ (see remark A.1 in Appendix A).

Example 3.10 (Rain and umbrellas)**.** As a bored mathematics student, you find yourself staring blankly at your computer screen day after day. Eventually, your gaze drifts to the window. Outside, the weather is predictably dismal. You see small figures hustling through the rain, umbrellas overhead, rushing to get indoors.

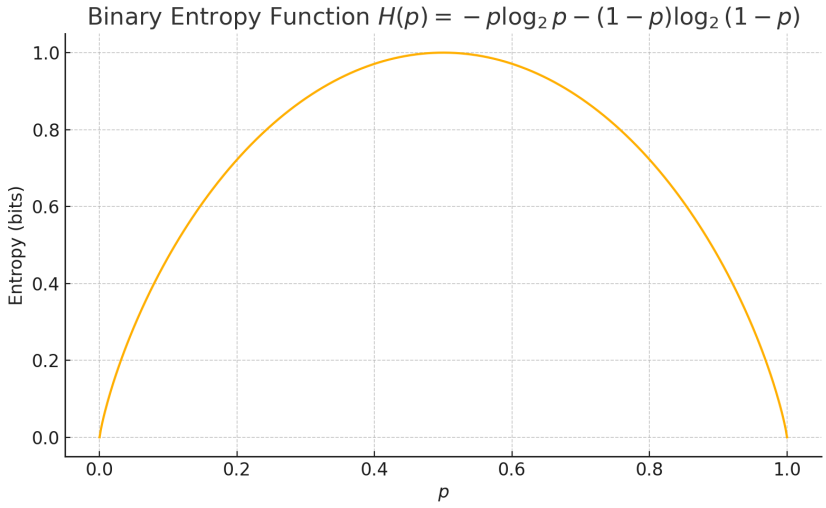


Figure 3.2: Binary entropy $h_2(p)$ as a function of p. We note that this function is clearly concave and that it achieves it's maximum at $p = 0.5$.

But then, like a true mathematician, a thought strikes you: What is the conditional entropy of "Rain" and "a person carrying an umbrella"? Naturally, you decide to find out.

You define a random variable X with outcome space $\mathbf{E}_x = \{1, 0\}$ representing respectively if any given person out of your window is carrying an umbrella or not. Then you define another binary random variable Y with outcome space $\mathbf{E}_y = \{1, 0\}$, representing rain '1' or no rain '0'. Committed, you make daily observations about the weather and how many people carry an umbrella, and you derive the (conditional) probabilities in the table below.

	Umbrella / $\Pr(X = 1)$	**No Umbrella** / $\Pr(X = 0)$	**Total**
Rain/ $\Pr(Y = 1)$	0.18	0.02	0.20
No Rain/ $\Pr(Y = 0)$	0.08	0.72	0.80
Total	0.26	0.74	1.00

Table 3.2: Joint probability table of Rain and Umbrella

We can now calculate the entropy of these random variables:

$$\mathrm{H}(X) = 0.827,$$
$$\mathrm{H}(Y) = 0.722.$$

We note that the entropy of "umbrellas" is bigger than that of "Rain".

3.2.4. Conditional and joint entropy

Now that we have established a clear intuition for entropy, we will turn to its properties. To do so, we first need to define *conditional entropy* and *joint entropy*.

Definition 3.11 (Joint entropy)**.** For two random variables X, Y we write the *joint entropy* as:

$$\mathrm{H}(X, Y) \equiv -\sum_{x,y} \Pr(X = x, Y = y) \log\left(\Pr(X = x, Y = y)\right).$$

This construction naturally extends to any finite number of random variables. Given a list of random variables $X = (X_1, X_2, \ldots, X_n)$, the outcome space of the joint variable is

$$\mathbf{E}_{X_1} \times \mathbf{E}_{X_2} \times \cdots \times \mathbf{E}_{X_n} = \mathbf{E}_X.$$

A specific outcome is denoted by $\vec{x} \in \mathbf{E}_X$ or informally $x \in \mathbf{E}$. And the joint entropy of X is given by:

$$\begin{aligned}
&\mathrm{H}(X) = \mathrm{H}(X_1, X_2, \ldots, X_n), \\
&= - \sum_{\vec{x} \in \mathbf{E_{X_1}} \times \cdots \times \mathbf{E_{X_n}}} \Pr(X = \vec{x}) \log \left(\Pr(X = \vec{x}) \right), \\
&= - \sum_{\vec{x} \in \mathbf{E_{X_1}} \times \cdots \times \mathbf{E_{X_n}}} \Pr\left[(X_1, \ldots, X_n) = (x_1, \ldots, x_n) \right] \log \left(\Pr[(X_1, \ldots, X_n) = (x_1, \ldots, x_n)] \right).
\end{aligned}$$

Intuitively, joint entropy measures the total uncertainty (or "surprise") associated with two or more random variables considered together. If the variables are *independent*, the joint entropy equals the sum of their individual entropies: $\mathrm{H}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y)$ (we will prove this later). On the other hand, if two variables are *fully dependent* on each other, the joint entropy is equal to the entropy of just one of them: $\mathrm{H}(X, Y) = \mathrm{H}(X)$. (As a side note, completely dependent random variables share the same entropy, see remark A.2 in Appendix A)

Example 3.12. Returning to Example 3.10 on rain and umbrellas, we can compute the joint entropy of umbrellas (X) and rain (Y).

$$\begin{aligned}
\mathrm{H}(X, Y) &= - \sum_{x \in \{0,1\}} \sum_{y \in \{0,1\}} \Pr(X = x, Y = y) \, \log \Pr(X = x, Y = y), \\
&= -\Big[0.18 \log(0.18) + 0.02 \log(0.02) + 0.08 \log(0.08) + 0.72 \log(0.72) \Big], \\
&\approx -\big(-0.4453 - 0.1129 - 0.2915 - 0.3413 \big), \\
&\approx 1.19 \text{ bits}.
\end{aligned}$$

We note that this value is quite close to the values of the individual random variables. That means that the added information/surprise of having both variables instead of only one is not very big. And thus we can conclude that the two random variables are correlated.

Definition 3.13 (Conditional entropy)**.** Given two random variables X and Y, we define the *conditional entropy* as:

$$\begin{aligned}
\mathbf{H}(X|Y) &\equiv \mathbb{E}_Y [\mathrm{H}(X|Y = y)] \\
&= \sum_y \Pr(Y = y) \, \mathrm{H}(X|Y = y) \\
&= \sum_y \Pr(Y = y) \sum_x - \Pr(X = x|Y = y) \log \Pr(X = x|Y = y)
\end{aligned}$$

Intuitively, is conditional entropy $\mathrm{H}(X|Y)$ the measure for the amount of additional information coming from X when we already know Y. Some important special cases for conditional entropy are:

- If X is independent of Y, then $\mathrm{H}(X \mid Y) = \mathrm{H}(X)$. (This is easily proven when one realizes that for two independent random variables X, Y: $\Pr(X = x|Y = y) = \Pr(X = x)$)
- If $X = Y$, then $\mathrm{H}(X \mid Y) = 0$.

Example 3.14. Again we go back to Example 3.10, and then we can calculate the entropy of umbrellas given that it is raining ($X|Y$). To do that we use the conditional probability definition, which is defined for $\Pr(Y = y) > 0$ as:

$$\Pr(X = x \mid Y = y) \equiv \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)}. \tag{3.5}$$

We then apply this to the example:

$$\begin{aligned}
\Pr(X = 1 \mid Y = 1) &= \tfrac{0.18}{0.20} = 0.9, \quad \Pr(X = 0 \mid Y = 1) = \tfrac{0.02}{0.20} = 0.1, \\
\Pr(X = 1 \mid Y = 0) &= \tfrac{0.08}{0.80} = 0.1, \quad \Pr(X = 0 \mid Y = 0) = \tfrac{0.72}{0.80} = 0.9,
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}(X \mid Y = 1) &= -\big(0.9 \log(0.9) + 0.1 \log(0.1)\big) \approx 0.469 \text{ bits}, \\
\mathrm{H}(X \mid Y = 0) &= -\big(0.1 \log(0.1) + 0.9 \log(0.9)\big) \approx 0.469 \text{ bits},
\end{aligned}$$

$$\begin{aligned}
\mathrm{H}(X \mid Y) &= \Pr(Y = 1)\ \mathrm{H}(X \mid Y = 1) + \Pr(Y = 0)\ \mathrm{H}(X \mid Y = 0), \\
&= 0.20 \cdot 0.469 + 0.80 \cdot 0.469 = 0.469 \text{ bits}.
\end{aligned}$$

We note that the entropy is low, this is because the probability distributions of umbrella and rain are strongly linked. So the extra information from having an umbrella, is low when we already know if it is raining or not.

3.3. Properties

We will now develop the basic properties of Shannon entropy such as concavity, subadditivity, the chain rule and more.

Remark 3.15 (Concavity of entropy)**.** *Let X_1 and X_2 be discrete random variables with respective probabilities distributions μ_{X_1} and μ_{X_2}, and let $\lambda \in [0, 1]$. Define a new random variable X with exactly $\mu_X = \lambda\mu_{X_1} + (1 - \lambda)\mu_{X_2}$ Then the entropy of X satisfies:*

$$\mathrm{H}(X) \geq \lambda\, \mathrm{H}(X_1) + (1 - \lambda)\, \mathrm{H}(X_2),$$

with equality if and only if X_1 and X_2 have the same distribution.

We will give here a short idea of the proof, we refer to Remark A.3 in the appendix for the full proof.

Proof. The entropy of a random variable X with finite outcome space $\mathbf{E}$ is defined as:

$$\begin{aligned}
\mathrm{H}(X) &= -\sum_{x \in \mathbf{E}} p_x \log p_x, \\
&= \sum_{x \in \mathbf{E}} f(p_x),
\end{aligned}$$

where $f(p_x) = -p_x \log(p_x)$. We differentiate $f(p_x)$ twice with respect to p_x to obtain on the interval $p_x \in (0, 1]$:

$$\frac{d^2}{d(p_x)^2} f(p_x) = -\frac{1}{p_x} < 0.$$

We note that the second derivative is strictly negative on the interval $(0, 1]$ and therefore f is a concave function. The entropy function is the sum over the function f and the sum of concave functions is concave. Therefore, $\mathrm{H}(X)$ is concave. (for a thorough proof, see Remark A.3). □

Theorem 3.16 (Jensen's inequality)**.** *Let f be a concave function and let X be a real-valued random variable with finite expectation, then:*

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)], \tag{3.6}$$

where the inequality flips for convex functions. And this inequality becomes strict if $f(X)$ is strictly concave and $X \neq \mathbb{E}[X]$ i.e. the random variable X takes more than one value.

This is technically a small extension on Jensen's inequality and the proof of this theorem is in Appendix A at Remark A.4. To get a better intuition of Jensen's inequality we will shortly discuss it visually. We illustrate Jensen's inequality for a concave function using the expectation form. Let $f(x) = \sqrt{x}$, which is concave on the interval $[0, 1]$. Consider two values $x_1 = 0.2$ and $x_2 = 0.8$, with associated probabilities $p_1 = 0.4$ and $p_2 = 0.6$. Define a discrete random variable X such that

$$\Pr(X = x_1) = p_1, \qquad \Pr(X = x_2) = p_2.$$

Then:

$$\mathbb{E}[X] = p_1 x_1 + p_2 x_2, \quad \text{and} \quad \mathbb{E}[f(X)] = p_1 f(x_1) + p_2 f(x_2).$$

Since f is concave, Jensen's inequality tells us that:

$$f\big(\mathbb{E}[X]\big) \geq \mathbb{E}[f(X)].$$

Geometrically, this means the point $f(\mathbb{E}[X])$ on the curve lies above the chord connecting $(x_1, f(x_1))$ and $(x_2, f(x_2))$, and in particular above the point $\mathbb{E}[f(X)]$, which lies on that chord (see Figure 3.3).

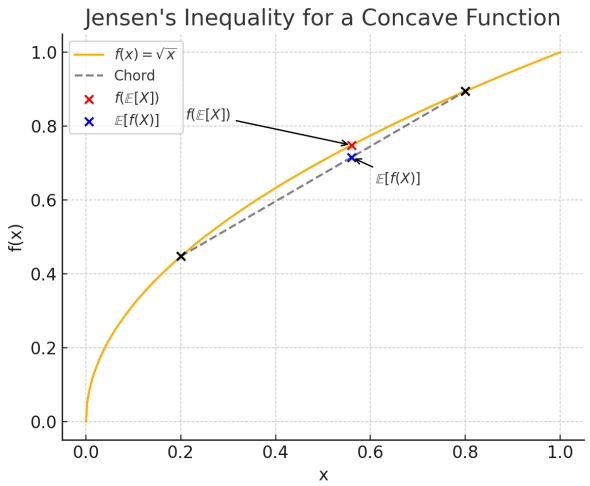


Figure 3.3: Visual illustration of Jensen's inequality for the concave function $f(x) = \sqrt{x}$.

Remark 3.17 (The entropy of a uniform random variable)**.** *let X be a uniformly distributed random variable with an outcome space $\mathbf{E}$ and $|\mathbf{E}| = m$, then $\forall x \in \mathbf{E}$, $p_x = 1/m$.*

$$\mathrm{H}(X) = -m \cdot \left(\frac{1}{m}\right) \log\left(\frac{1}{m}\right) = \log(m).$$

Now we can use Theorem 3.16 and Remark 3.15 to prove our first big property of entropy.

Theorem 3.18 (Uniform bound)**.** *Let X be a random variable with outcome space $\mathbf{E}$ such it has m outcomes with non-zero probability ($m = |\operatorname{support}(X)|$). Then the entropy of this random variable is bounded:*

$$\mathrm{H}(X) = -\sum_{x \in \mathbf{E}} p_x \log p_x \leq \log(m),$$

with equality if and only if $p_x = \frac{1}{m}$ for all $x \in \mathbf{E}$ (i.e. only when X is uniformly distributed).

Proof. We will prove this with the use of Jensen's inequality and the fact that the entropy function is concave.

$$\begin{aligned}\mathrm{H}(X) &= \sum_{s \in \mathbf{E}} -p_s \log(p_s) = \sum_{s \in \mathbf{E}} f(p_s), \\ &= \mathbb{E}[f(p_s)] \cdot m < m \cdot f\left(\mathbb{E}[p_s]\right) = m \cdot f\left(\frac{1}{m} \sum_{s \in \mathbf{E}} p_s\right) = -m \frac{1}{m} \log\left(\frac{1}{m}\right) = \log(m).\end{aligned}$$

With only equality if p_s is constant (i.e. with a uniform distribution). □

This uniform bound is intuitively reasonable, since the uniform distribution maximizes unpredictability and, consequently, surprise and entropy.

Lemma 3.19 (Subadditivity)**.** *For all discrete random variables X, Y with finite outcome space:*

$$\mathrm{H}(X, Y) \leq \mathrm{H}(X) + \mathrm{H}(Y).$$

Or when we repeat this for $X_1, X_2, \ldots, X_n$:

$$\mathrm{H}(X_1, X_2, \ldots, X_n) \leq \mathrm{H}(X_1) + \mathrm{H}(X_2) + \cdots + \mathrm{H}(X_n).$$

Proof.

$$\begin{aligned}
\mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) &= \sum_x -p_x \log p_x - \sum_y p_y \log p_y + \sum_{x,y} p_{x,y} \log p_{x,y}, \\
&= \sum_{x,y} \left(-p_{x,y} \log p_x - p_{x,y} \log p_y + p_{x,y} \log p_{x,y}\right), \\
&= \sum_{x,y} p_{x,y} \log \frac{p_{x,y}}{p_x p_y} = \sum_{x,y} -p_{x,y} \log \frac{p_x p_y}{p_{x,y}}, \\
&= \sum_{x,y} p_{x,y} g\left(\frac{p_x p_y}{p_{x,y}}\right), \\
&\geq g\left(\sum_{x,y} p_{x,y} \frac{p_x p_y}{p_{x,y}}\right) = f(1) = 0,
\end{aligned}$$

where we use in the third line that: $\sum_y p_x, y = p_x$ and in the last line we use that $-\log(x)$ is a convex function, and then we use Jensen's inequality. □

This lemma is again intuitive as the surprise of random variables together can't be bigger than the sum of the surprise of the single random variables.

Example 3.20. When we look back at Example 3.12 then we can see that: $\mathrm{H}(X,Y) = 1.19 < 0.827 + 0.722 = \mathrm{H}(X) + \mathrm{H}(Y)$. We see that in this case, the joint entropy is strictly smaller. This is because X and Y are not independent.

Theorem 3.21 (The Chain Rule of Shannon Entropy)**.** *For two discrete random variables X, Y, the entropy satisfies the following chain rule:*

$$\mathrm{H}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y|X). \tag{3.7}$$

Or with more than two variables and recalling the notation that we specified earlier (see 2.2):

$$\mathrm{H}(X_1, X_2, \ldots, X_i) = \sum_{i-1}^{n} \mathrm{H}(X_i | X_{<i}). \tag{3.8}$$

Proof. Recall the definitions of entropy and conditional entropy 3.1 & 3.13:

$$\mathrm{H}(X,Y) = -\sum_{x,y} p_{x,y} \log(p_{x,y}), \tag{3.9}$$

$$\mathrm{H}(Y|X) = -\sum_{x,y} p_{x,y} \log(p_{y|x}). \tag{3.10}$$

Starting from 3.9, and using the definition of conditional probabilities:

$$\begin{aligned}
\mathrm{H}(X,Y) &= -\sum_{x,y} p_{x,y} \log\left[p_{y|x} \cdot p_x\right], \\
&= -\sum_{x,y} p_{x,y} \left[\log p_x + \log p_{y|x}\right], \\
&= -\sum_{x,y} p_{x,y} \log p_x - \sum_{x,y} p_{x,y} \log p_{y|x}.
\end{aligned}$$

Notice the first term does not depend on y, thus we can simplify it by summing over y:

$$\mathrm{H}(X,Y) = -\sum_x \log p_x \sum_y p_{x,y} - \sum_{x,y} p_{x,y} \log p_{y|x}.$$

And since $\sum_y p_{x,y} = p_x$, we have:

$$\begin{aligned}\mathrm{H}(X,Y) &= -\sum_x p_x \log p_x - \sum_{x,y} p_{x,y} \log p_{y|x}, \\ &= \mathrm{H}(X) + \mathrm{H}(Y|X).\end{aligned}$$

□

Remark 3.22. *If X and Y are independent random variables, then:*

$$\mathrm{H}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y)$$

Proof. from the chain rule:

$$\begin{aligned}\mathrm{H}(X,Y) &= \mathrm{H}(X) + \mathrm{H}(Y|X), \\ &= \mathrm{H}(X) + \mathrm{H}(Y),\end{aligned}$$

where we used that $\mathrm{H}(Y|X) = \mathrm{H}(Y)$ for independent variables. □

We have now covered nearly all the fundamental properties of entropy. One final theorem from information theory remains to be discussed: the data processing inequality. We will list here the entropy variant of this inequality, and we will give an intuitive proof. For more information on the data processing inequality, Markov chains and the information formula, we refer to [13].

Theorem 3.23 (The data processing inequality for entropy)**.**

$$H\big(X \mid f(Y)\big) \;\geq\; H\big(X \mid Y\big),$$

for any (possibly random) function f.

An intuitive justification of the data processing inequality is that a function $f(Y)$ can never contain more information than the original variable Y, since it is merely a processed version of it. When we condition on $f(Y)$ instead of Y, we are given less information, which leads to greater uncertainty and therefore, greater entropy.

3.4. Entropy in Combinatorics

Having developed an intuition for entropy and explored its key properties, we now turn to its applications in combinatorics. In this section, we present some classic entropy-based proofs. And understanding these proofs will help our intuition for the potential applications of entropy.

We will begin with perhaps the most well-known example among mathematicians unfamiliar with the rest of entropy.

3.4.1. Binomial sum bound

Theorem 3.24 (Entropy bound on binomial sums)**.** *let k be such that $0 < k \leq n/2$, then:*

$$\sum_{i=0}^{k} \binom{n}{i} \leq 2^{n \cdot h_2\left(k/n\right)} = \left(\frac{n}{k}\right)^{k} \left(\frac{n}{n-k}\right)^{n-k},$$

where h_2 is the Bernoulli entropy formula (see Example 3.4).

At first glance, it is not at all obvious that entropy can assist in proving this inequality. However, as we work through the proof, we will see how entropy can be applied here.

Proof. The proof will follow approximately these steps: 1. To bound the sum of binomial coefficients up to k, we first define a random variable of which (by design) its entropy is equal to the logarithm of this binomial sum. 2. Now we can bound the entropy of this random variable with the properties discussed in the last section. And thus, put a bound on the logarithm of the binomial sum.

Let k be such that $0 < k \leq n/2$. Then we define the ordered list of n binary random variables: $X = (X_1, X_2, \ldots, X_n)$ such that $X_1 + X_2 + \cdots + X_n \leq k$. And X is uniformly distributed over all these possible lists. The number of different ordered lists $(x_1, x_2, \ldots x_n)$ that exist, such that $x_1 + x_2 + \cdots + x_n = p$ is equal to $\binom{n}{p}$. So all possible lists such that $x_1 + x_2 + \cdots + x_n \leq p$ is:

$$\sum_{i=0}^{k} \binom{n}{i}.$$

X is uniformly distributed over all those possible lists so: (remember the uniform bound of entropy)

$$\mathrm{H}(X) = \mathrm{H}(X_1, X_2, \ldots, X_n) = \log\Big(\sum_{i=0}^{k} \binom{n}{i}\Big).$$

We also know from sub-additivity (Theorem 3.19) that:

$$\log\Big(\sum_{i=0}^{k} \binom{n}{i}\Big) = \mathrm{H}(X_1, X_2, \ldots, X_n) \leq \mathrm{H}(X_1) + \mathrm{H}(X_2) + \cdots + \mathrm{H}(X_n) = \sum_{i=1}^{n} h_2\left(\Pr(X_i = 1)\right).$$

Each X_i is a Bernoulli random variable (it has two outcomes) with $\Pr(X_i = 1 | X_1 + X_2 + \cdots + X_n = m) = m/n$. This probability is because X is uniformly distributed over all lists, so the 1's and 0's are equally distributed over the string. So:

$$\Pr(X_i = 1 \mid X_1 + \cdots + X_n = m) = \frac{m}{n} \leq \frac{k}{n} \leq \frac{1}{2}.$$

We can also compute the precise probability $\Pr(X_i = 1)$ by averaging over all the possible values of the total number of ones:

$$\Pr(X_i = 1) = \sum_{m=0}^{k} \Pr(X_i = 1 \mid \sum_{j=1}^{n} X_j = m) \cdot \Pr\left(\sum_{j=1}^{n} X_j = m\right) = \sum_{m=0}^{k} \frac{m}{n} \cdot \frac{\binom{n}{m}}{\sum_{i=0}^{k}\binom{n}{i}}.$$

However, for our purpose of bounding the binomial sum, this is too unwieldy. So we instead bound the probability that X_i is one by the simpler term: k/n. This is possible because:

$$\Pr\left(X_i = 1 | \sum_{j=1}^{n} X_j = m\right) \leq k/n$$

for all m between 0 and k. And thus the average over these probabilities will also be smaller than k/n. Next:

$$\Pr(X_i = 1) \leq \frac{k}{n} \Rightarrow \quad h_2(X_i) \leq h_2\left(\frac{k}{n}\right).$$

The implication is due to the fact that the Bernoulli entropy function is increasing on the interval $(0, 0.5)$ (see figure 3.2). And $k/n \leq 1/2$ thus the entropy of these Bernoulli random variables is smaller than $h_2(k/n)$.

Putting this all together, we obtain:

$$\begin{aligned}\log(\sum_{i=0}^{k}\binom{n}{i}) &= \mathrm{H}(X),\\ &\leq \mathrm{H}(X_1) + \mathrm{H}(X_2) + \cdots + \mathrm{H}(X_n),\\ &\leq n h_2\left(\frac{k}{n}\right),\end{aligned}$$

as required. □

3.4.2. The Bregman-Minc Inequality

Another famous theorem that can be proved using entropy is the Bregman-Minc inequality. It gives an upper bound on the permanent of a $\{0, 1\}$ valued matrix A in terms of its row sums. And these kinds of matrices can be seen as an adjacency matrix of a bipartite graph. So then a bound on the permanent gives a bound on the number of perfect matchings in the graph.

Even though this inequality was originally proven by Bregman in 1973 without entropy [6]. In 1979 Radhakrishnan proved it with the use of entropy [27] and it is that proof that we will discuss here.

Theorem 3.25. *Let A be a $n \times n$ matrix with entries in $\{0, 1\}$, and let d_i denote the sum of the i-th row, then Bregman's theorem [6] states:*

$$\mathrm{per}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i,\sigma(i)} \leq \prod_{i=1}^{n} (d_i!)^{1/d_i}, \tag{3.11}$$

where S_n is the set of all permutations of length n or equivalently all the ordered lists of n elements.

The idea of the proof from Radhakrishnan is to define a random variable such that the logarithm of the permanent is equal to the entropy of this random variable just like that the logarithm of the binomial sum was equal to the entropy of our random variable in the proof of binomial sum entropy bound (Theorem 3.24). And then we can bound the entropy of our random variable and thus the logarithm of the permanent with use of the properties of entropy.

Proof. Let $A = (a_{ij}) \in \{0,1\}^{n\times n}$ be a $\{0,1\}$-valued matrix, where the sum of the i-th row is d_i. The permanent of this matrix is:

$$\operatorname{per}(A) = \sum_{\sigma\in S_n} \prod_{i=1}^{n} a_{i,\sigma(i)}.$$

We notice that $\prod_{i=1}^{n} a_{i,\sigma(i)} \neq 0$ only for permutations in the set: $\Omega \subseteq S_n, \Omega = \{\sigma \in S_n : \forall i \in [n], a_{i,\sigma(i)} = 1\}$. Such that for example $\sigma = (2,4,5,7,3,6,1)$, then $\sigma_1 = 2$, $\sigma_2 = 4$ etc. And with k we denote the size $|\Omega|$.

For example, we take the matrix $\hat{A}$ as shown below:

$$\hat{A} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix},$$

then the only permutations such that the product is not zero are: $(1,2,3)$ and $(3,1,2)$. So $\Omega = \{(1,2,3),(3,1,2)\}$.

Next, we define a random variable $X = (X_1, X_2, \ldots, X_n)$, where X is uniformly distributed over all permutations in Ω and $X_1, X_2, \ldots, X_n$ represent $\sigma_1, \sigma_2, \ldots, \sigma_n$.

It follows that $\mathrm{H}(X) = \log(k)$, because X is uniformly distributed over k possibilities, see Remark 3.17. We also note that the permanent is the sum of 1 over all permutations in Ω and thus $\operatorname{per}(A) = k$. Thus:

$$\log\big(\operatorname{per}(A)\big) = \log(k) = \mathrm{H}(X).$$

Subsequently, we apply the chain rule to $\mathrm{H}(X)$:

$$\mathrm{H}(X) = \sum_{i=1}^{n} \mathrm{H}(X_i | X_{<i}).$$

However, we have to be careful because the reveal order matters for the uncertainty. So the order in which we get to know X_i's matters. To explain what we mean, we will go back to the example matrix from earlier.

We will take the permutation: $\hat{\sigma} = (1,2,3)$ and subsequently we will look at the number of possibilities for $\hat{\sigma}_i$ given that we know what $\hat{\sigma}_{<i}$ is.

We proceed by the reveal order $(1,2,3)$ such that we will look first at the first row, then at the second row, and lastly at the third row.

we recall the matrix $\hat{A}$:

$$\hat{A} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}.$$

(1) We first look at the first row: there are two 1's, so there are two possibilities for $\hat{\sigma}_1$, namely 1 and 3. We then "reveal" that $\hat{\sigma}_1 = 1$. (2) For the second row there are two 1's, but we already revealed that, $\hat{\sigma}_1 = 1$ so we can't take the first column again, so 2 is the only possibility for $\hat{\sigma}_2$. (3) And for, $\hat{\sigma}_3$ there is only one choice: 3. Thus, the number of possibilities per "reveal" is: 2,1,1 (we will later define this number of possibilities as $N_i(X, Y)$).

Now if we chose the different reveal order: $(3, 2, 1)$ then the number of available possibilities per reveal are: 2,2,1.

We thus see that the reveal order matters, consequently we need to take the expectation over all reveal orders!

So we go back to the proof, and we define another random variable $Y = (Y_1, Y_2, \ldots, Y_n)$, which is uniformly distributed over all, $\tau \in S_n$ and this random variable represents the possible *reveal orders*.

Following this, we apply the chain rule:

$$\mathrm{H}(X) = \mathbb{E}_Y\big[\sum_{i=1}^{n} \mathrm{H}\left(X_i | X_j : Y_j < Y_i\right)\big].$$

Thereafter, we define the number of possibilities given $X_j : Y_j < Y_i$ as $N_i(X, Y)$. That is, N_i counts how many 1's in row i correspond to columns that have not already been selected by earlier rows. Furthermore, $(X_i | X_j : Y_j < Y_i)$ is a random variable over $N_i(X, Y)$ possibilities. And then the uniform bound implies:

$$\mathbb{E}_Y\big[\mathrm{H}\left(X_i | X_j : Y_j < Y_i\right)\big] \leq \mathbb{E}_{X,Y}\big[\log N_i(X, Y)\big].$$

Because Y is uniformly distributed over all permutations, $N_i(X, Y)$ is uniformly distributed over $[d_i]$. To understand why, we will have a closer look at row i.

Row i has d_i ones, so there are d_i columns for which $a_{ij} = 1$. We denote the set of these columns C. Each of these columns is used by exactly one row (possibly row i itself). Let us consider all the rows that 'choose' these columns: $\{r_1, r_2, \ldots, r_{d_i}\}$ such that $Y_{r_k} \in C$. These 'compete' with row i for the columns it can potentially choose.

Since we reveal the rows in a random order according to the uniform random variable Y, the position of row i among these d_i competitor rows is uniformly random. Thus, row i is equally likely to be first among its competitors or last, etc.

The order of i within its competitor rows determines N_i. And thus N_i varies uniformly over $[d_i]$ (it has probability $1/d_i$ to be any value in $\{1, 2, 3, \ldots, d_i\}$, which implies that the expectation or average of N_i is:

$$\mathbb{E}_{X,Y}[\log N_i] = \frac{1}{d_i}\sum_{k=1}^{d_i}\log k = \frac{\log(d_i!)}{d_i}.$$

Putting it all together, we obtain:

$$\log \operatorname{per} A = \mathrm{H}(X) \leq \sum_{i=1}^{n}\mathbb{E}_{X,Y}[\log N_i] = \sum_{i=1}^{n}\frac{\log(d_i!)}{d_i}.$$

Taking the exponent of both sides, we obtain the desired inequality. □

We have examined two compelling proofs that make use of entropy, but there exists a wide range of other applications. For further reading, we refer the reader to [15, 21, 23]. Having understood these examples, we are naturally led to ask: why does entropy work so well in these contexts, and is there a unifying theme underlying the proofs that employ it?

3.4.3. A connecting theme?

Given our limited exposure to the full range of entropy's applications in combinatorics, it is difficult to draw broad conclusions about a general entropy-based approach. However, two patterns stand out. First, all the entropy-based proofs we encountered are used to establish bounds or inequalities, not equalities. Second, these proofs typically follow a common structure: (1) Define random variables whose entropy is similar to the expression we want to bound. (2) Use entropy's properties to bound the entropy of these variables and thereby bounding the original expression.

We will see this pattern come back in our proof of the bound on the Union-Closed Sets Conjecture.

3.5. Why entropy?

Another question one might ask would be: Why does entropy work so well? And: Might there be another function that would work even better? Here we will discuss that shortly.

3.5.1. Connection between entropy and counting

Entropy has proven to be an extremely effective tool in combinatorics. This can be partly explained by two properties, that provide a numerical handle on set sizes and set structures. (1) The maximum entropy for any distribution over a set S, such that $|S| = m$ is $\log m$, achieved when all the outcomes are equally likely. We therefore can translate any argument about entropy into a statement about the size $|S|$. (2) Moreover, most structures within a family of sets $\mathcal{F}$ can be translated in to non-uniform distributions over these sets. And because these structure induced distributions are non-uniform they lower the entropy. Directly **linking entropy and structure**. For example:

If many sets in $\mathcal{F}$ share common elements, then the coordinate variables X_i have low entropy $\mathrm{H}(X_i)$, reducing the total entropy $\mathrm{H}(X)$. On the other hand, if sets are disjoint or highly diverse, then the entropy increases, as each coordinate carries more uncertainty. (We will see this in action in our proof of the bound on the Union-Closed Sets Conjecture).

3.5.2. Other useful properties

We have thus seen the connection between entropy and counting. However, these properties aren't the only useful properties of entropy and we will here give a short and non-extensive list:

- **Non-negativity** $\mathrm{H}(X) \geq 0$, with equality if and only if X is deterministic.
- **Maximum at uniformity** The entropy function of a random variable with a certain number of outcomes is maximized when the random variable is uniformly distributed over all outcomes.
- **Chain rule** $\mathrm{H}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y \mid X) \leq \mathrm{H}(X) + \mathrm{H}(Y)$ and for independent variables: $\mathrm{H}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y)$.
- **Concavity** Entropy is a concave function of the probability distribution and therefore we can use Jensen's inequality
- **Invariance under relabelling** Entropy depends only on the distribution of X, not on the labels of its outcomes.
- **Monotonicity under conditioning** $\mathrm{H}(X \mid Y) \leq \mathrm{H}(X)$, with equality if and only if X and Y are independent.

We will see that many of these properties will be used in the next chapters.

3.5.3. Is entropy unique?

When we understand the usefulness of entropy we might get greedy and wonder if there is any other function that is even better than entropy. However, the Russian mathematician Khinchin showed already in 1957 that entropy can be characterised under the assumption of four axioms [17].

Let $\Omega = \{1, 2, \ldots n\}$ be an outcome space and let $\mu = (p_1, \ldots, p_n)$ be a probability distribution over Ω. A function $\mathrm{H}(\mu)$ is said to satisfy the Shannon-Khinchin axioms if it satisfies the following properties:

- **1 Continuity** $\mathrm{H}[\mu]$ is continuous in the entries of μ.
- **2 Maximum at uniformity** $\mathrm{H}(\mu)$ is maximal when μ is the uniform distribution.
- **3 Expansibility** $\mathrm{H}(\mu)$ is invariant under the addition of zero-probability outcomes.
- **4 Separability/ Chain rule** Let $\rho = (r_1(p_1, q_1), \ldots, r_n(p_n, q_n)$ be a joint distribution ofμ and $\nu = (q_1, \ldots, q_n)$. Then the function separates as:

  $$\mathrm{H}(\rho) = \mathrm{H}(\mu) + \mathrm{H}(\nu \mid \mu),$$

  where $(\nu \mid \mu)$ is the conditional probability.

Khinchin showed that under his proposed axioms, the only functional form (up to a positive multiplicative constant) that satisfies them is the Shannon entropy. Thus, one way to assess whether entropy is the appropriate measure for a given problem is to consider whether all these axiomatic properties are required.

It is worth noting that alternative characterizations of entropy also exist, along with generalized forms such as Rényi entropy. A good starting point is [26], which is summarized in a presentation by Zennaro [32].

3.5.4. Conclusion

The entropy function has several useful properties and a natural connection to both the size and structure of sets. Therefore, it is well suited for use in combinatorics. Furthermore, one might ask whether there exists a function that performs even better than the entropy function, leading to an exploration of generalized forms of entropy. However, Khinchin already showed that there is only one function (up to a positive constant) that satisfies his four axioms. Thus, to determine whether entropy is the appropriate tool for a given problem, one can check whether all of these axioms are required.

4

Probabilities and the Union-Closed Set Conjecture

4.1. the Conjecture

As discussed in the introduction, is the Union-Closed Sets Conjecture most commonly known in the form:

Conjecture 4.1 (The Union-Closed Sets Conjecture)**.** Any finite union-closed family of sets $\mathcal{F}$, excluding the family that only contains the empty set, has an element that is contained in at least half of the member-sets.

However, this formulation doesn't directly or obviously lend itself to the tools of information theory, and it is maybe therefore not unsurprising that before Gilmer in 2022, there was no published entropy based progress on the Union-Closed Sets Conjecture.

In this section, we explore the conjecture. By first looking at examples, and later we will discuss the link between this conjecture about sets and the realm of entropy.

4.1.1. Examples from the conjecture

If we were able to find a counter example to the conjecture, we could immediately disprove it! Therefore, let us experiment with a few families.

Let's start with the Family $\mathcal{Z}$ from Example 2.4.

$$\mathcal{Z} = \Big\{\{\emptyset\}, \{1\}, \{2\}, \{1,2\}, \{1,2,3,4\}\Big\}.$$

We observe that the elements 1 and 2 both appear in 3 out of five ($> 1/2$) of the member sets. So unfortunately we haven't refuted the conjecture yet!

We can also look at one of the most famous families: *Power-sets*.

Remark 4.2. *All power-sets are union-closed families.*

Proof. let S be some nonempty set. Remember that the power-set of S ($\mathcal{P}(S)$) is the set of all subsets of a set. Then:

$$A, B \in \mathcal{P}(S) \implies A, B \subseteq S \implies A \cup B \subseteq S \implies A \cup B \in \mathcal{P}(S).$$

So, if two sets are in the power-set then their union is also in the power-set. Therefore, $\mathcal{P}(S)$ is union closed. □

Example 4.3. $S = \{1, 2, 3\}$, then:

$$\mathcal{P}(S) = \{\emptyset, \{1\},\ \{2\},\ \{3\},\ \{1, 2\},\ \{1, 3\},\ \{2, 3\},\ \{1, 2, 3\}\}\,.$$

We immediately observe that 1, 2, and 3 each appear in 4 out of the 8 sets. Since $\frac{4}{8} = \frac{1}{2}$, the conjecture holds exactly in this case.

Remark 4.4. *If S is finite, then every element $s \in S$ is contained in exactly half of the sets in $\mathcal{P}(S)$. (The conjecture is tight for power-sets.)*

Proof. Let S be a finite set with $|S| = n$. The power set $\mathcal{P}(S)$ consists of all possible subsets of S, and therefore contains exactly 2^n subsets.

Consider an arbitrary element $s \in S$. Any subset containing s can be constructed by choosing whether to include each of the remaining $n - 1$ elements. Hence, the total number of subsets containing s is exactly 2^{n-1}.

Thus, the proportion of subsets containing s is $\frac{2^{n-1}}{2^n} = \frac{1}{2}$. □

We thus see that power sets place a sharp bound on the conjecture. And in fact, power-sets are the only known separating union-closed family such that the most frequent element is in half of the sets. (We will clarify the term 'separating set' later.) This is reflected by a conjecture by Poonen.

Conjecture 4.5 (Poonen [25])**.** Let $\mathcal{A}$ be a separating union-closed family. Then it contains an element that appears in strictly more than half of it's member-sets, unless $\mathcal{A}$ is a power set.

We meet the idea of "separating family" in this formulation. We will give you the definition here, but we won't go into too much detail, so that we don't get off track.

Definition 4.6 (Separating Family)**.** A family of sets $\mathcal{F}$ over a ground set U is called *separating* if, for every pair of distinct elements $x, y \in U$, there exists a set $A \in \mathcal{F}$ such that exactly one of x and y belongs to A. That is,

$$\forall x, y \in U,\ x \neq y \Rightarrow \exists A \in \mathcal{F} \text{ such that } (x \in A \wedge y \notin A) \vee (x \notin A \wedge y \in A).$$

The notion of separating families closely aligns with the intuition of not having any redundant elements.

4.1.2. Finite families

We note that the Union-Closed Sets Conjecture doesn't hold for families with infinitely many member-sets. This is illustrated by Poonen [25].

Take, for example, the union-closed family $\mathcal{F}$ consisting of the sets $\{i, i+1, i+2, \dots\}$ for every integer i. $|\mathcal{F}|$ is infinite, but no element has infinite frequency.

Thus, all the families considered in this thesis will be finite and furthermore, we require that all the **member-sets to be finite**. This is not a restriction, because:

Lemma 4.7. *All families with a finite number of member-sets and an infinite ground set $U(\mathcal{G})$ can be reduced to a family with the same cardinality and a finite ground-set $U(\mathcal{F})$, such that the conjecture holds for $\mathcal{G}$ if and only if it holds for $\mathcal{F}$.*

Proof. The idea behind the proof is to reduce the finite family with an infinite ground set to a separating finite family, which by definition cannot have an infinite ground set. And showing that if the conjecture holds for the separating family then it also holds for the family with the infinite groundset.

We consider a family $\mathcal{G}$, which has finitely many member-sets: $|\mathcal{G}| = m$ and where $U(\mathcal{G})$ is infinite. We fix the order of the sets: $\mathcal{G} = \{A_1, A_2, \dots, A_m\}$.

We define the characteristic vector of an element $x \in U(\mathcal{G})$ as: $\vec{c_x} \in \{0,1\}^m$

$$(\vec{c_x})_i = \begin{cases} 1 & \text{if } x \in A_i, \\ 0 & \text{if } x \notin A_i. \end{cases} \tag{4.1}$$

We observe that there are only finitely many (2^m) unique characteristic vectors possible. Therefore, infinitely many elements x, y must share the same characteristic vector ($\vec{c_x} = \vec{c_y}$).

Elements x and y with identical characteristic vectors appear in exactly the same subsets of $\mathcal{G}$. Consequently, x is contained in at least half of the subsets of $\mathcal{G}$ if and only if the same holds for y.

Thus, we can remove y from all sets in $\mathcal{G}$, creating a reduced family $\mathcal{H}$ (with one less element in its universe). The conjecture holds for $\mathcal{G}$ if and only if it holds for $\mathcal{H}$. Repeating this process for all elements sharing identical characteristic vectors produces a reduced family $\mathcal{F}$ satisfying $|U(\mathcal{F})| \leq 2^m$, such that the conjecture holds for $\mathcal{G}$ if and only if it holds for $\mathcal{F}$. □

4.2. Entropy and sets

Now that we thoroughly understand the concept of entropy and have developed an intuition for the conjecture, we might begin to wonder how these two seemingly distinct ideas are related. To explore this connection, we continue building on the intuition from Example 2.11.

Let $\mathcal{F}$ be a finite family, which has a finite groundset, $|\,\mathrm{U}(\mathcal{F})| = n$. We randomly select a set $A \in \mathcal{F}$, and view its characteristic vector $x_S \in \{0,1\}^n$ as a vector of binary random variables, $(X_1, X_2, \dots, X_n)$ where $X_i = 1$ if the i^{th} element is in set A and 0 else. To demonstrate the practical applicability of this approach and the notation of $(X \cup Y)$, we present the following examples:

Example 4.8. This example illustrates the intuitive and widely used notation $X \cup Y$ to denote the pointwise maximum of two random variables, when they represent characteristic vectors of sets.

Let $A = \{1, 2\}$ and $B = \{2, 3\}$, with characteristic vectors given by:

$$x_A = (1, 1, 0),$$
$$x_B = (0, 1, 1).$$

The union of the sets is $A \cup B = \{1, 2, 3\}$, and the characteristic vector of the union is:

$$\begin{aligned} x_{A\cup B} &= (1, 1, 1) \\ &= (\max(1, 0), \max(1, 1), \max(0, 1)). \end{aligned}$$

This shows that each coordinate of $x_{A\cup B}$ is given by the maximum of the corresponding entries of x_A and x_B:

$$(x_{A\cup B})_i = \max\left((x_A)_i, (x_B)_i\right).$$

Therefore, when two random variables X and Y represent the characteristic vectors of some sets, the random variable Z representing the characteristic vector of their union is given by the pointwise maximum:

$$Z_i = \max(X_i, Y_i).$$

Thus, the union of the underlying sets is represented by the pointwise maximum of their characteristic vectors, which justifies the use of the $\cup$ notation in this context.

Example 4.9. Let $\mathcal{F}$ be:

$$\mathcal{F} = \Big\{\emptyset, \{1\}, \{2\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\Big\} = \Big\{\emptyset, A, B, C, D, E\Big\}.$$

Then the characteristic vectors of the sets are:

$$x_\emptyset = (0, 0, 0),$$
$$x_A = (1, 0, 0),$$
$$x_B = (0, 1, 0),$$
$$x_C = (1, 1, 0),$$
$$x_D = (0, 1, 1),$$
$$x_E = (1, 1, 1).$$

If we now sample randomly a set S from this family. And we will denote its characteristic vector as the random variable $X = (X_1, X_2, X_3)$. $\Pr(X_1 = 1) = 1/2$ (there are three characteristic vectors that start with a zero and three that start with a 1). $\Pr(X_2 = 1) = 2/3$ and $\Pr(X_3 = 1) = 1/3$. But again, just like in Example 2.11 the random variables X_1, X_2, X_3 are not independent: $\Pr(X_3 = 1 | X_2 = 0) = 0 \neq 1/3$.

We can now calculate the entropy of X, since it has equal probability of being any of the 6 sets, we have:

$$\mathrm{H}(X) = \log(6) \approx 2.58.$$

Next, we choose another set R from this family at random, and denote the characteristic vector of this set as $Y = Y_1, \ldots, Y_n$.

What is $\mathrm{H}(X \cup Y)$, where we define $(X \cup Y)$ as the coordinate-wise maximum of X and Y (see Example 4.8)?

We will compute this by simply counting the frequency of the possible outcomes of the two random sets S and R.

$R \cup S$	Frequency
$\emptyset$	1
$\{1\}$	3
$\{2\}$	3
$\{1,2\}$	9
$\{2,3\}$	5
$\{1,2,3\}$	15

$\{1\}$, for example, has frequency 3 because it can be formed as $\{1\} \cup \{1\}$, $\emptyset \cup \{1\}$, or $\{1\} \cup \emptyset$. The sets R and S are uniformly distributed over the original sets; consequently, the entropy is:

$$\begin{aligned} H(X \cup Y) &= -\left(\frac{1}{36}\log\frac{1}{36} + \frac{3}{36}\log\frac{3}{36} + \frac{3}{36}\log\frac{3}{36} + \frac{9}{36}\log\frac{9}{36} + \frac{5}{36}\log\frac{5}{36} + \frac{15}{36}\log\frac{15}{36}\right), \\ &\approx 2.16. \end{aligned}$$

We observe that $\mathrm{H}(X) > \mathrm{H}(X \cup Y)$. This inequality makes intuitive sense, as the new random variable $X \cup Y$ possesses additional structure, causing it to deviate from the uniform distribution. And we are then reminded that structure reduces entropy, which is exactly one of the reasons why entropy based analysis can be so fruitful in combinatorics.

Furthermore, we do not want to give the false impression that characteristic vectors, when viewed as random variables, are necessarily independent. To demonstrate this, we consider the following example.

Example 4.10. Let $\mathcal{F}$ be $\mathcal{P}([3])$. And let X be the characteristic vector of a set randomly drawn from this family. Then $\Pr(X_1 = 1) = \Pr(X_2 = 1) = \Pr(X_3 = 1) = 1/2$. This is because all elements are in exactly half of the sets. Now, if we look at the conditional probabilities, we find that these are still always $1/2$: $\Pr(X_1 = 1 | X_2 = 0) = 1/2$ for example.

Remark 4.11 (Random variables of power-sets are independent)**.** *Let $\mathcal{F}$ be a power set of $[n]$, and let the random variables X_i represent if the i-th element is in a uniformly randomly chosen set S, then: those random variables $X_1, X_2, \ldots, X_n$ are mutually independent $\mathrm{Ber}(1/2)$ random variables.*

Proof. let $\mathcal{F} = \mathcal{P}([n])$. We choose a set S uniformly at random from $\mathcal{F}$. We define the random variable $X = X_1, X_2, \ldots X_n$ as the characteristic vector of S. Note that:

$$\Pr(X_i = 1) = \frac{1}{2},$$

Because all elements are contained in exactly half of the sets from the power set (Remark 4.4).

We take an arbitrary subset of $M \subset [n]$. And we will show that all $\{X_i : i \in M\}$ are mutually independent. To that end do we fix for all elements in M if they are in the set S or not (we choose $X_i = 1$ or $X_i = 0$

for all $i \in M$. Subsequently, we consider the probability of this outcome: We take an arbitrary subset $M \subset [n]$, and we will show that the random variables $\{X_i : i \in M\}$ are mutually independent. To that end, we fix for each $i \in M$ whether i is in the set S or not, that is we choose x_i as either 0 or 1. Subsequently, we consider the probability of this outcome:

$$\Pr\big(X_i = x_i, \quad \forall i \in I\,\big).$$

So, we fix $|M|$ elements. Thus, there are exactly $n - |M|$ elements of $[n]$ not yet constrained, and each of those can be chosen freely to be in or out of S. Hence:

$$\begin{aligned}
\Pr\big(X_i = x_i, \forall i \in I\,\big) &= \frac{\text{\# possible sets where } X_i = x_i}{\text{Total number of sets}}, \\
&= \frac{2^{n-|M|}}{2^n}, \\
&= 2^{-|M|}, \\
&= \prod_{i \in M} \frac{1}{2}, \\
&= \prod_{i \in M} \Pr(X_i = x_i).
\end{aligned}$$

Thus we have showed that, the mutual probability is just the product of the individual probabilities. And since this holds for every choice of I and every choice of x_i, we can conclude that: All X_i are mutually independent. □

In this chapter, we have developed an understanding of the Union-Closed Sets Conjecture and explored how set systems can be analysed using entropy. This foundation will enable us, in the next chapter, to finally turn to a central topic of this thesis: a new bound on the Union-Closed Sets Conjecture.

5

Proving A constant bound of $\frac{3-\sqrt{5}}{2}$ for the Union-Closed Sets Conjecture

Until now, we have garnered a lot of skill and knowledge about the Union-Closed Sets Conjecture, entropy, and the link between them. And it is finally time to apply this newfound knowledge to understand the proof that there is at least 1 element in any union-closed family of sets $\mathcal{F}$ (other than the family containing only the empty set), that is in at least $\frac{3-\sqrt{5}}{2}|\mathcal{F}|$ sets.

This line of reasoning using entropy was first used by Gilmer to prove that there must be an element in at least a 0.01 fraction of the sets of a union-closed family (proving the first constant bound for the conjecture) [16]. His argument was quickly optimised by various authors to the constant of $\frac{3-\sqrt{5}}{2}$ [2, 10, 28, 24].

We will follow the outline of [10] and the blog post [1]. We will go through the proof in quite some detail to make it more accessible and comprehensible.

5.1. The structure of the proof

The main idea of the proof is to look for a contradiction between two theorems. We will start by introducing the first one and prove it, and then we continue onto the second theorem, which is the main result of Gilmer and others such as [10].

Theorem 5.1. *Let A and B be Independently and uniformly sampled at random from a union-closed family $\mathcal{F}$. Because $\mathcal{F}$ is union-closed $A \cup B \in \mathcal{F}$. And let X and Y denote the random variable representing the characteristic vector of A and B respectively, then:*

$$\mathrm{H}(X \cup Y) \leq \mathrm{H}(X). \tag{5.1}$$

This theorem follows from Theorem 3.18, which states that entropy is maximized under the uniform distribution. The underlying intuition is discussed in Example 4.9.

Proof. X is uniformly distributed over all sets in $\mathcal{F}$, so for any other samples (Z) following Theorem 3.18:

$$\mathrm{H}(X) \geq \mathrm{H}(Z). \tag{5.2}$$

Because $\mathcal{F}$ is union-closed, all unions of sets are also contained in $\mathcal{F}$ and thus $A \cup B$ is a sample on $\mathcal{F}$. Therefore, $\mathrm{H}(X \cup Y) \leq \mathrm{H}(X)$. □

The next theorem is the main result of the entropy based method applications on the Conjecture 4.1. And it is a clear contradiction with Theorem 5.1 when $p < \psi$:

Theorem 5.2. *Let A and B be i.i.d. subsets of the ground-set $[n]$ drawn according to some distribution with $\Pr(i \in A) \leq p$ and $\Pr(i \in B) \leq p$ for all $i \in [n]$. Let X and Y be the two random variables representing the characteristic vector of A and B respectively, then:*

$$\mathrm{H}(X \cup Y) \geq \frac{1-p}{1-\psi}\,\mathrm{H}(X), \tag{5.3}$$

where $\psi = (3-\sqrt{5})/2$ = $(\varphi - 1)/\varphi$ such that $\varphi = (1+\sqrt{5})/2$ is the golden ratio.

The contradiction between the two theorems implies that $\neg\left(\Pr(i \in A) \leq p \ \ \forall i \in [n]\right)$. Or in other words: there exists an element that has a bigger probability than p to be in any given set. And there is a contradiction for $p < \psi$. So there is an element that is in at least $(3-\sqrt{5})/2 \mid \mathcal{F} \mid$ sets.

Now that we have established how Theorem 5.1 and Theorem 5.2 prove that there is an item in the ground set of $\mathcal{F}$ that is in at least $\psi|\mathcal{F}|$ sets. All that remains is to prove Theorem 5.2.

5.2. Proof of the main theorem

The proof of this theorem will proceed roughly as follows: first, we separate the entropy expression into a sum with the use of the chain rule and then we apply the data processing inequality to this sum. Subsequently, we rewrite the sum in terms of the binary entropy function, after which we apply inequalities for the binary entropy to bound the entire expression in terms of $\mathrm{H}(X)$.

Let A, B be two independently sampled sets according to some distribution from a family $\mathcal{F}$ with ground-set $[n]$ and $\Pr(i \in A) \leq p$, $\Pr(i \in B) \leq p$ for all $i \in [n]$. Then let X, Y denote the characteristic vectors of A and B respectively.

Then by the chain rule (Theorem 3.21):

$$\mathrm{H}(X \cup Y) = \sum_{i=1}^{n} H\left((X \cup Y)_i \mid (X \cup Y)_{<i}\right).$$

Now, the information captured in $(X_{<i}, Y_{<i})$ is more or equal than the information in $(X \cup Y)_{<i}$. And more information implies less uncertainty and thus less entropy. So by Theorem 3.23:

$$\sum_{i=1}^{n} H\left((X \cup Y)_i \mid (X \cup Y)_{<i}\right) \geq \sum_{i=1}^{n} H\left((X \cup Y)_i \mid X_{<i}, Y_{<i}\right).$$

Next, we consider one specific term of this sum. And through the definition of conditional entropy, we

can rewrite it to:

$$\mathrm{H}((X \cup Y)_i \mid X_{<i}, Y_{<i}) = \mathbb{E}_{a,b}\left[\mathrm{H}((X \cup Y)_i \mid X_{<i} = a, Y_{<i} = b)\right],$$

where $\mathbb{E}_{a,b}$ is an expectation over all possible $X_{<i}$ and $Y_{<i}$.

To continue our analysis, we want to rewrite this expression into a single binary entropy equation. To that end, we define $\bar{p}_{ia} = \Pr(i \notin A \mid X_{<i} = a)$ and $\bar{p}_{ib} = \Pr(i \notin B \mid Y_{<i} = b)$ (we define $p_{ij} = (1 - \bar{p}_{ij}) = \Pr(i \in J \mid Y_{<i} = j)$).

Since A and B are chosen independently, the probability that i is not in $A \cup B$ given $X_{<i} = a$ and $Y_{<i} = b$ is $\bar{p}_{ia}\bar{p}_{ib}$. We also note that the random variable '$(X \cup Y)_i$' is a binary random variable, so we can use the binary entropy h_2 to obtain:

$$\mathrm{H}((X \cup Y)_i \mid X_{<i} = a, Y_{<i} = b) = h_2(\bar{p}_{ia}\bar{p}_{ib}).$$

So far, we have obtained:

$$\begin{aligned}
\mathrm{H}(X \cup Y) &= \sum_i^n \mathrm{H}\left((X \cup Y)_i \mid (X \cup Y)_{<i}\right), \\
&\geq \sum_i^n \mathrm{H}\left((X \cup Y)_i \mid X_{<i}, Y_{<i}\right), \\
&= \sum_i^n \mathbb{E}_{a,b}\left[\mathrm{H}((X \cup Y)_i \mid X_{<i} = a, Y_{<i} = b)\right], \\
&= \sum_i^n \mathbb{E}_{a,b}[(h_2(\bar{p}_{ia}\bar{p}_{ib}))].
\end{aligned}$$

Lemma 5.3 ([1, 5, 10])**.**

$$h_2(\bar{p}_{ia}\bar{p}_{ib}) \geq \frac{\varphi}{2}\left(\bar{p}_{ia}h_2(\bar{p}_{ib}) + \bar{p}_{ib}h_2(\bar{p}_{ia})\right),$$

where $\varphi = \dfrac{1+\sqrt{5}}{2}$.

This lemma is rather involved, so its proof is presented in the section 5.3 for readability. However, as it forms a crucial part of the overall argument, it has not been relegated to the appendix.

We apply Lemma 5.3 to the previous result, to obtain:

$$\sum_i^n \mathbb{E}_{a,b}[(h_2(\bar{p}_{ia}\bar{p}_{ib}))] \geq \sum_i^n \mathbb{E}_{a,b}[\frac{\varphi}{2}\left(\bar{p}_{ia}h_2(\bar{p}_{ib}) + \bar{p}_{ib}h_2(\bar{p}_{ia})\right)]$$

As A and B are i.i.d., we have that $\mathbb{E}_a[\bar{p}_{ia}] = \mathbb{E}_b[\bar{p}_{ib}]$. And $\bar{p}_{ib}$ is precisely the probability that the i-th element is not in B given $Y_{<i} = b$. Following the sum notation of the expectation and the hypothesis

that every element has a lower probability than p to be in any given set:

$$\begin{aligned}\mathbb{E}_b\big[\bar{p}_{ib}\big] &= \sum_b \Pr\big(Y_{<i} = b\big)\ \Pr\big(i \notin B \mid Y_{<i} = b\big) \\ &= \Pr\big(i \notin B\big) \\ &= 1 - \Pr\big(i \in B\big) \\ &\geq 1 - p.\end{aligned}$$

$\mathbb{E}_b[p_{ib}] \leq p$. We obtain:

$$\mathbb{E}_b[\bar{p}_{ib}] = \mathbb{E}_b[1 - p_{ib}] \geq (1 - p).$$

Then, remembering that $h_2(p) = h_2(1-p)$ and thus:

$$\mathbb{E}_a\left[h_2(\bar{p}_{ia})\right] = \mathbb{E}_a\left[h_2(p_{ia})\right] = \mathbb{E}_a\left[\mathrm{H}(X_i \mid X_{<i} = a)\right] = \mathrm{H}(X_i \mid X_{<i}).$$

putting this together we arrive at:

$$\begin{aligned}\mathrm{H}(X \cup Y) &\geq \sum_i^n \frac{\varphi}{2}\left[\mathbb{E}_a[(\bar{p}_{ia} h_2(\bar{p}_{ib}))] + \mathbb{E}_b[(\bar{p}_{ib} h_2(\bar{p}_{ia}))]\right], \\ &\geq \sum_i^n \frac{\varphi}{2}(1-p)\left[\mathrm{H}(X_i|X_{<i}) + \mathrm{H}(Y_i|Y_{<i})\right].\end{aligned}$$

since A and B are i.i.d.: $\mathrm{H}(X_i|X_{<i}) = \mathrm{H}(Y_i|Y_{<i})$. Thus:

$$\begin{aligned}\mathrm{H}(X \cup Y) &\geq \sum_i^n \frac{\varphi}{2}\left[\mathbb{E}_a[(\bar{p}_{ia} h_2(\bar{p}_{ib}))] + \mathbb{E}_b[(\bar{p}_{ib} h_2(\bar{p}_{ia}))]\right], \\ &\geq \sum_i^n \varphi(1-p)\left[\mathrm{H}(X_i|X_{<i})\right], \\ &= \varphi(1-p)\,\mathrm{H}(X), \\ &= \frac{1-p}{1-\psi}\,\mathrm{H}(X),\end{aligned}$$

where the second to last step is the reversed chain rule and $\varphi = \frac{1}{1-\psi}$.

This concludes the proof of the main theorem (leaving the proof of Lemma 5.3 for the next section)

5.3. The proof of Lemma 5.3

Our next goal is to prove:

$$h_2(\bar{p}_{ia}\bar{p}_{ib}) \geq \frac{\varphi}{2}\left(\bar{p}_{ia} h_2(\bar{p}_{ib}) + \bar{p}_{ib} h_2(\bar{p}_{ia})\right). \tag{5.4}$$

However, we will proceed in multiple steps. First, we first prove a lemma from Chase and Lovett, and then combine it with a result from Boppana to obtain Lemma 5.3.

5.3.1. Lemma from Chase and Lovett

Lemma 5.4. *[10] We define the function $f : [0,1]^2 \to \mathbb{R}_{\geq 0}$ as:*

$$f(x,y) = \begin{cases} 1 & \text{if } x \in \{0,1\} \text{ or } y \in \{0,1\}, \\ \dfrac{h_2(xy)}{h_2(x)y + h_2(y)x} & \text{otherwise.} \end{cases}$$

The function $f(x,y)$ achieves a minimum on the diagonal $x = y$, where x is an interior point of $[0,1]^2$.

Proof. We begin with showing that this is indeed a continuous function and then we will show that the minimum is achieved on the diagonal. This is proven by a clever analysis of the function f and related functions.

This is a continuous function because in the interior it is made out of continuous functions and has no asymptotes and on the boundary:

$$\begin{aligned} \lim_{x\to 0} f(x,y) &= \lim_{x\to 0} \frac{-xy\log(xy) - (1-xy)\log(1-xy)}{y(-x\log x - (1-x)\log(1-x)) + x(-y\log y - (1-y)\log(1-y))}, \\ &= 1 \end{aligned}$$

and

$$\begin{aligned} \lim_{x\to 1} f(x,y) &= \frac{-y\log(y) - (1-y)\log(1-y)}{(1-x)\log(1-x) + (-y\log y - (1-y)\log(1-y))}, \\ &= 1, \end{aligned}$$

using the fact that $\lim_{x\to 1}(1-x)\log(1-x) = 0$. We note that f is a symmetrical function for x and y, thus $\lim_{y\to 0} f(x,y) = \lim_{y\to 1} f(x,y) = 1$. Furthermore, we have included a proof in the appendix, that $\lim_{x,y\to(0,0)} f(x,y) = \lim_{x,y\to(1,1)} f(x,y) = 1$. at Remark A.5.

Therefore, the function f is continuous and has value 1 along the boundary.

$f(0.5, 0.5) \approx 0.811278 \leq 1$, so there exist a minimum of f in the interior of $(x,y) \in [0,1]^2$. Let this minimum be at the point (x^*, y^*) and the minimal value: $f(x^*, y^*) = \alpha$. We note here that $\alpha > 0$ because both the denominator and numerator are positive on the domain.

Define $g(x) := \dfrac{h_2(x)}{x}$. Then:

$$f(x,y) = \frac{g(xy)}{g(x) + g(y)} \geq \alpha. \tag{5.5}$$

Define

$$F(x,y) = g(xy) - \alpha(g(x) + g(y)). \tag{5.6}$$

Note that, $F(x,y) \geq 0$, following Equation 5.5. Furthermore, the derivative of $F(x,y)$ exists and is continuous on $(0,1)^2$ and $F(x^*, y^*) = 0$ (and $x^*, y^* \neq 0$). Therefore, $F(x^*, y^*)$ must be some kind of minimum and consequently it's derivative must be zero:

$$\nabla F(x^*, y^*) = 0.$$

But we can also calculate the derivative of F by taking the derivative of Equation 5.6 and evaluating it at (x^*, y^*):

$$\nabla F = (yg'(xy) - \alpha g'(x), xg'(xy) - \alpha g'(y)) = (0, 0).$$

Which implies that:

$$y^* g'(x^* y^*) = \alpha g'(x^*),$$
$$x^* g'(x^* y^*) = \alpha g'(y^*).$$

Now rearrange both equations:

$$\alpha = \frac{y^* g'(x^* y^*)}{g'(x^*)} = \frac{x^* g'(x^* y^*)}{g'(y^*)} \quad \Rightarrow \quad \frac{y^*}{g'(x^*)} = \frac{x^*}{g'(y^*)} \quad \Rightarrow \quad x^* g'(x^*) = y^* g'(y^*). \tag{5.7}$$

Define:

$$G(x) := xg'(x) = x\frac{d}{dx}\left(\frac{h_2(x)}{x}\right).$$

So from Equation 5.7: $G(x^*) = G(y^*)$ and:

$$G(x) = x\frac{d}{dx}\left(\log x + \frac{1-x}{x}\log(1-x)\right) = -\frac{\log(1-x)}{x}.$$

This function $G(x)$ is strictly increasing on $(0, 1)$ as can be seen in a picture of the plot of $G(x)$ using GeoGebra. However, we will also prove that $G(x)$ is strictly increasing in the Appendix A at remark A.6.

Since $G(x^*) = G(y^*)$ and G is strictly increasing, we find that $x^* = y^*$. Therefore:

$$\alpha = f(x^*, y^*) = f(x^*, x^*) = \frac{h_2(x^{*2})}{2x^* h_2(x^*)}.$$

So the minimum of $f(x, y)$ is at $f(x^*, x^*)$. □

So, we have proven that:

$$\frac{h_2(xy)}{h_2(x)y + h_2(y)x} \geq \frac{h_2((x^*)^2)}{2(x^*)h_2(x^*)} = \alpha.$$

it remains for us to prove that: $\alpha \geq \varphi/2$

5.3.2. The lemma from Boppana

Lemma 5.5 ([5]). *For all* $x \in [0, 1]$,

$$h_2(x^2) \geq \varphi x h_2(x).$$

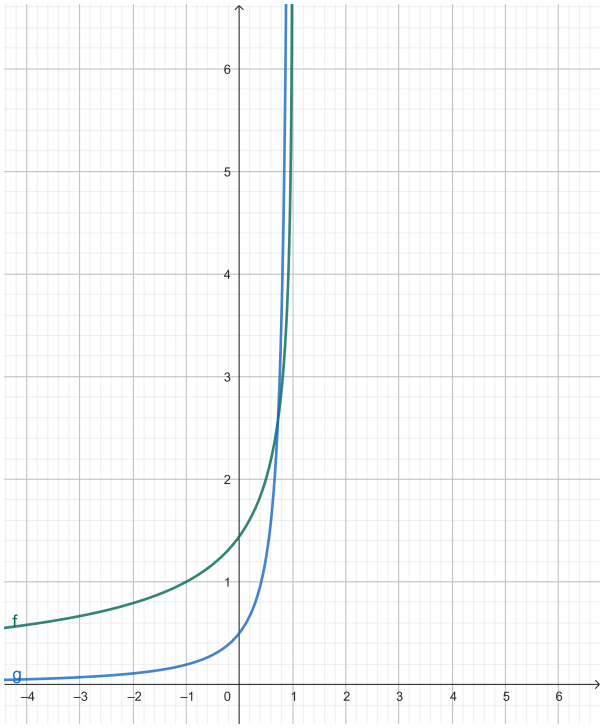


Figure 5.1: Plot of $G(x) = -\frac{\log(1-x)}{x}$ (green line) and its derivative $G'(x)$ (blue line). This plot illustrates that G(x) is strictly increasing on the interval $(0, 1)$.

where h_2 is the binary entropy function and φ is the golden ratio: $\dfrac{\sqrt{5}+1}{2}$.

Proof. We define the following function on $\mathbb{R}$:

$$h(x) = \begin{cases} \ln(2)h_2(x) = -x\ln|x| - (1-x)\ln|1-x| & \text{if } x \neq 0 \text{ and } x \neq 1, \\ 0 & \text{if } x = 0 \text{ or } x = 1. \end{cases} \tag{5.8}$$

(This is the extended natural log variation of the binary entropy expression, which simplifies derivatives here.)

We then define the function $v(x)$:

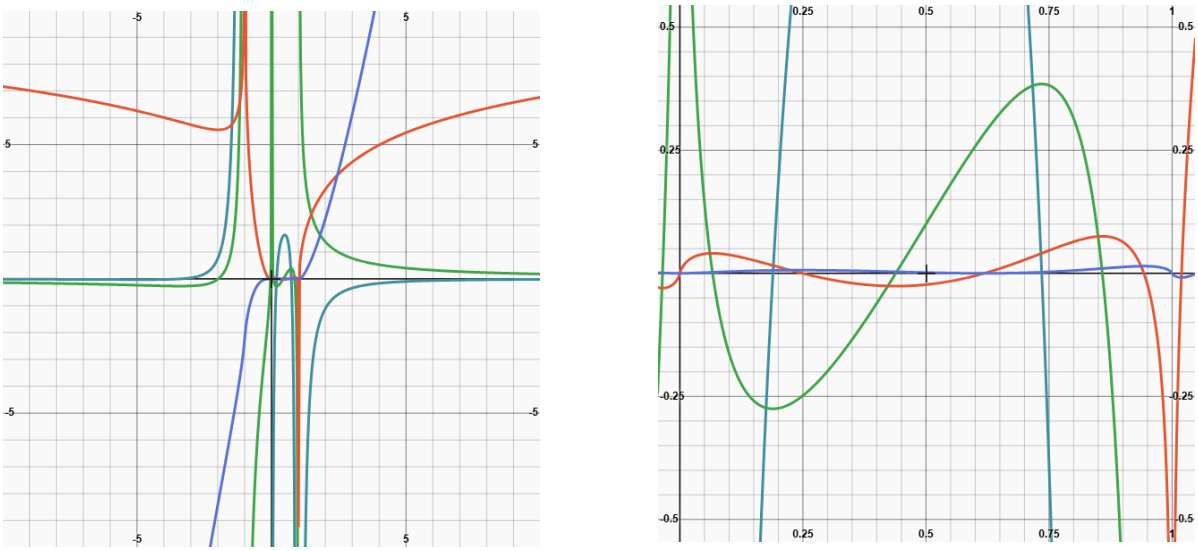
$$v(x) = h(x^2) - \varphi x h(x).$$

We note that to prove the Lemma 5.5 it is sufficient to show that $v(x)$ is nonnegative on $[0, 1]$.

To that end, we compute derivatives of $v(x)$ until we get rid of the logarithms:

$$\begin{aligned} v'(x) &= -2x\ln\left(x^2\right) + 2x\ln\left(1-x^2\right) + 2\varphi x\ln(x) - 2\varphi\ln(1-x)\,x + \varphi\ln(1-x), \\ v''(x) &= -2\ln\left(x^2\right) + 2\ln\left(1-x^2\right) + 2\varphi\ln(x) - \frac{4x^2}{1-x^2} + \frac{2\varphi x}{1-x} - \frac{\varphi}{1-x} - 2\varphi\ln(1-x) + 2\varphi - 4, \\ v'''(x) &= -\frac{\varphi x^3 + 4x^2 - 3\varphi x - 2\varphi + 4}{x^5 - 2x^3 + x}. \end{aligned}$$

We can plot these derivatives for a more intuitive picture, but we want to emphasize that these graphs are not used in the proof.



(a) Zoomed in

Figure 5.2: Two graphs showing the behaviour of the first three derivatives of the function $v(x) = \ln(2)(h_2(x^2) - \varphi x h_2(x))$, where h_2 is the binary entropy function and φ is the golden ratio. 5.2a is a zoomed in version of the graph. The flat **blue** line is the *original function*, the **orange** line is the *first derivative*, the **green** line is the *second derivative* and the **dark cyan** is the *third derivative* of $v(x)$.

We now examine the third derivative v''' to determine how many zeros it can have, since this will bound the number of zeros of v itself and ultimately allow us to show that $v(x) > 0$ for all $x \in (0, 1)$. First, consider the denominator of v''', which factorizes as:

$$x^5 - 2x^3 + x = x(1 - x^2)^2. \tag{5.9}$$

This equation is bigger than zero for $x \in (0, 1)$, as $x > 0$ and $(1 - x^2)^2 > 0$. Consequently, the denominator doesn't have any roots on $(0, 1)$.

We continue by analysing the numerator of v''': $p(x) = -\varphi x^3 - 4x^2 + 3\varphi x + 2\varphi - 4$.

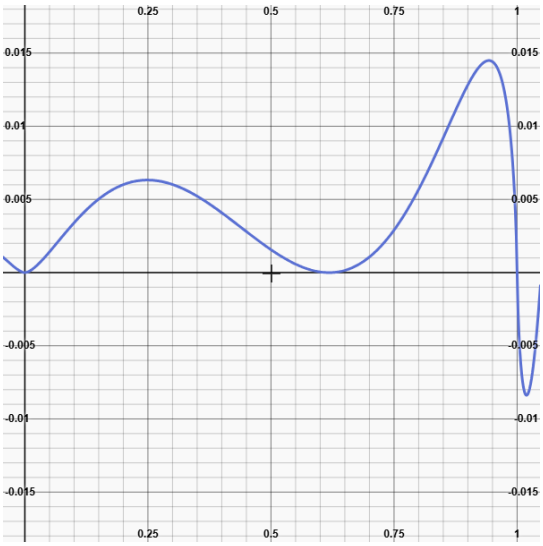
We note that this is a continuous third order polynomial, thus it crosses the x-axis at most 3 times. The leading coefficient is negative. Hence as $x \to -\infty$, $p(x) \to +\infty$. We also observe that $p(0) = 2\phi - 4 \approx -0.7639 < 0$, thus $p(x)$ must have crossed the x-axis at least once before $x = 0$.

This implies that $p(x)$ has at most 2 positive roots. subsequently, v''' has at most 2 roots in $[0, 1]$.

We then apply the reverse form of Rolle's theorem (see Theorem A.7 in the appendix), which asserts that if a function g has k zeros on (a, b), then any antiderivative G of g can have at most $k + 1$ zeros on the same interval.

In our case, on the interval $(0, 1)$, we have already established that $v'''(x)$ admits at most two roots on $(0, 1)$. Therefore:

1. v'' can have at most $2 + 1 = 3$ zeros,
2. v' can have at most $3 + 1 = 4$ zeros,



(a) $v(x)$, zoomed in (especially over the y-axis)

(b) $v'''(x)$

Figure 5.3: Two graphs showing the behaviour of the functions $v(x)$ and $v'''(x)$, where $v(x) = \ln(2)(h_2(x^2) - \varphi x h_2(x))$ and h_2 is the binary entropy function. Sub-figure 5.3a is a zoomed in plot of $v(x)$ (the y and x-axis are scaled differently). And 5.3b is a zoomed out graph of the third derivative of $v(x)$

.

3. v can have at most $4 + 1 = 5$ zeros.

Thus v has at most 5 zeros on $[0, 1]$, when we count multiplicity. We will just state these zeros here, but the calculations are in Appendix A at Remark A.8. $v(x)$ has double roots at 0 and φ^{-1} and a single root at 1, as can be seen in the Sub-Figure 5.3a.

So, we have found all 5 nonnegative roots. Furthermore, v has a double root at φ^{-1}, so v must be all nonnegative on [0,1] or all non-positive. Because, the equation can't change signs in the middle because there is no single root.

We evaluate v at 0.5 to obtain the value: $0.00156 > 0$, consequently $v(x)$ is all nonnegative on $[0, 1]$. Therefore:

$$\ln(2)\big(h_2(x^2) - \varphi x h_2(x)\big) \geq 0 \qquad 0 \leq x \leq 1.$$

And therefore:

$$h_2(x^2)\ln(2) \geq \varphi x h_2(x)\ln(2),$$
$$h_2(x^2) \geq \varphi x h_2(x).$$

□

Now we can put the lemma from Chase and Lovett together with the lemma from Boppana to arrive at:

$$\frac{h_2(xy)}{h_2(x)y + h_2(y)x} \geq \frac{h_2((x^*)^2)}{2(x^*)h_2(x^*)} = \alpha \geq \frac{\varphi}{2} \qquad \text{for } x, y \in [0, 1].$$

And substituting $x = \bar{p}_{ia}$ and $y = \bar{p}_{ib}$:

$$h_2(\bar{p}_{ia}\bar{p}_{ib}) \geq \frac{\varphi}{2}\left(\bar{p}_{ia}h_2(\bar{p}_{ib}) + \bar{p}_{ib}h_2(\bar{p}_{ia})\right).$$

Which concludes the proof of Lemma 5.3.

Now that we understand the lemmas we will go back one last time to recap the whole proof of the main theorem.

$$
\begin{aligned}
\mathrm{H}(X \cup Y) &= \sum_{i}^{n} \mathrm{H}\left((X \cup Y)_i \mid (X \cup Y)_{<i}\right), \\
&\geq \sum_{i}^{n} \mathrm{H}\left((X \cup Y)_i \mid X_{<i}, Y_{<i}\right), \\
&\geq \sum_{i}^{n} \mathbb{E}_{a,b}\left[\mathrm{H}((X \cup Y)_i \mid X_{<i} = a, Y_{<i} = Y)\right], \\
&= \sum_{i}^{n} \mathbb{E}_{a,b}[(h_2(\bar{p}_{ia}\bar{p}_{ib}))], \\
&\geq \sum_{i}^{n} \mathbb{E}_{a,b}[\frac{\varphi}{2}\left(\bar{p}_{ia}h_2(\bar{p}_{ib}) + \bar{p}_{ib}h_2(\bar{p}_{ia})\right)] \\
&= \sum_{i}^{n} \frac{\varphi}{2}\left[\mathbb{E}_a[(\bar{p}_{ia}h_2(\bar{p}_{ia}))] + \mathbb{E}_b[(\bar{p}_{ib}h_2(\bar{p}_{ib}))]\right] \\
&\geq \sum_{i}^{n} \varphi(1-p)\left[\mathrm{H}(X_i|X_{<i})\right] \\
&= \varphi(1-p)\,\mathrm{H}(X) \\
&= \frac{1-p}{1-\psi}\,\mathrm{H}(X)
\end{aligned}
$$

And this leads again to the contradiction with Theorem 5.1 that we discussed earlier.

6

Further improvements on the lower bound and Sawin

6.1. Further improvements on the lower bound

In Chapter 5 we proved that there must be at least one element in at least $\frac{3-\sqrt{5}}{2}|\mathcal{F}|$ sets in the union-closed family $\mathcal{F}$. This was for many mathematicians not the end of the entropy approach on the Union-Closed Sets Conjecture. Sawin demonstrated in his article [28] that this bound $\frac{3-\sqrt{5}}{2} \approx 0.38197$ could be improved by considering a combination of i.i.d. random variables and coupled random variables. This laid out a framework for improving the bound. Thereafter, Yu and Cambie [31, 8] evaluated and expanded this framework to obtain a slightly higher bound of 0.38234. Then Liu considered another class of coupling under which the two binary sequences are i.i.d. conditioned on an auxiliary random variable, and he improved the constant to: 0.38271 [19].

6.2. Sawin

Sawin does two major things: He optimizes the bound from Gilmer to $\frac{3-\sqrt{5}}{2}$. And he also gives an idea how this bound can be pushed even further. Additionally, while optimizing Gilmer's approach, Sawin proves a very useful lemma, that is later also used by Das and Wu (next chapter).

In this Section we will look at the idea from Sawin to push the bound further than $\frac{3-\sqrt{5}}{2}$ (which was eventually done by Cambie and Yu) and we will go through the proof of lemma 3 from Sawin's paper.

6.2.1. Sawin's idea for further improvement

The main idea of Sawin is to introduce correlated random samples in addition to independent ones. Intuitively, independent draws A, B from $\mathcal{F}$ give a certain amount of entropy increase when taking unions. But Sawin suggests mixing in another pair of draws (A, C) that are not independent but have a controlled correlation chosen to maximize the entropy gain when taking the union. Sawin shows that by cleverly coupling of A and C, one can force a larger entropy jump from A to $A \cup C$. And he shows that while the correlation might introduce some entropy loss in later steps, the right choice can ensure

a bigger gain in the union step, than the loss in later steps. He sketches the proof in the paper, and this should be read before one continues on to the eventual optimizations by Cambie or Yu.

6.3. Notation and preliminaries

Since the initial step of the proof we are about to discuss relies more heavily on measure theory than we have encountered thus far, it is helpful to briefly review the relevant notation and preliminaries. However, as measure theory is not the main focus of this thesis, we will keep these definitions short (and therefore less rigorous). For readers interested in a more thorough treatment, we refer to [4]. For readers already familiar with the basics of measure theory, or those willing to accept certain measure-theoretic claims used in the proof, we recommend proceeding directly to section 6.4.

Definition 6.1 (Topology)**.** Let A be a set. A *topology* on A is a collection $\tau \subseteq \mathcal{P}(A)$ of subsets of A, such that these sets can be defined as 'open sets', which have some properties. The pair (A, τ) is called a *topological space*.

Definition 6.2 (Borel σ-algebra)**.** The *Borel σ-algebra* $\mathcal{B}(D)$ is the smallest collection of subsets of D that contains every open set and is closed under countable unions, countable intersections, and complements.

Definition 6.3 (Borel probability measure)**.** A Borel probability measure μ on D is a function

$$\mu : \mathcal{B}(D) \longrightarrow [0, 1],$$

satisfying:

1. $\mu(\emptyset) = 0$,
2. $\mu(D) = 1$,
3. If $\{A_i\}$ is a countable family of pairwise disjoint Borel sets, then:

$$\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i).$$

The intuition behind a Borel probability measure is that it assigns to each Borel set $A \subseteq D$ a number $\mu(A)$ in $[0, 1]$, with total mass 1 on D.

Definition 6.4 (Expectation)**.** If μ is a Borel probability measure on D and $f : D \to \mathbb{R}$ is bounded and continuous, then the *expectation* of f under μ is:

$$\mathbb{E}_{p\sim\mu}[f(p)] = \int_0^1 f(p) d\mu(p).$$

Definition 6.5 (Weak convergence of probability measures)**.** Let $\{\mu_n\}_{n=1}^{\infty}$ be a sequence of Borel probability measures on $[0, 1]$, and let μ be another such measure. We say

$$\mu_n \xrightarrow{\text{weakly}} \mu$$

if and only if for every bounded continuous function $f : [0,1] \to \mathbb{R}$,

$$\lim_{n\to\infty} \int_0^1 f(p)d\mu_n(p) = \int_0^1 f(p)d\mu(p).$$

The intuition behind this is that weak convergence means that averages of any "nice" function f converge.

Definition 6.6 (Compactness)**.** A topological space X is said to be *compact* if every open cover has a finite sub cover.

Remark 6.7. *If a metric space K is* compact*, then every sequence $\{x_n\} \subseteq K$ has a subsequence $\{x_{n_k}\}$ that converges to a point in K.*

Theorem 6.8 (Heine-Borel Theorem)**.** *All bounded and closed intervals in $\mathbb{R}^n$ are compact.*

Theorem 6.9 (Prokhorov's theorem)**.** *If D is compact, the space of all Borel probability measures on D is also compact.*

By combining these theorems, we conclude that on the space of all probability measures on $[0,1]$, equipped with the weak topology, every sequence has a convergent subsequence.

6.4. lemma 3 from Sawin

In this section, we discuss a key lemma from Sawin's article. He uses this lemma to refine Gilmer's argument to obtain the improved bound of $\frac{3-\sqrt{5}}{2}$. Our goal is to understand the proof both as a foundation for its application in the work of Das and Wu (see next chapter), and as a starting point for the subsequent improvements on the bound developed by Cambie and Yu.

Lemma 6.10 (Lemma 3 from Sawin[28])**.** *Let $u \in [0,1]$. Let p, q be i.i.d. $[0,1]$-valued random variables with $\mathbb{E}[p] = \mathbb{E}[q] \leq u$. Then:*

$$\mathbb{E}[h_2(p+q-pq)] \geq \mathbb{E}[h_2(p)] \cdot \begin{cases} \dfrac{h_2(2u-u^2)}{h_2(u)} & \text{if } u \leq \frac{3-\sqrt{5}}{2}, \\ (1-u)\cdot\dfrac{2}{\sqrt{5}-1} & \text{if } u \geq \frac{3-\sqrt{5}}{2}. \end{cases}$$

6.4.1. Overview and idea

Sawin's proof is not the most straightforward, and at times it is not immediately clear why certain steps are taken. Therefore, it is helpful to first outline the structure of the proof before we dive in.

He begins by rewriting the formula from Lemma 6.10. He introduces a parameter λ in place of the function of u, since λ is not yet determined. He then moves all terms to one side of the inequality so that he can analyse whether a certain expression is positive, let's call this *expression A*.

Then he argues that to prove that Expression A is always positive, it suffices to show that it is positive for the probability distribution μ that minimizes A.

Next, Sawin shows that if μ minimizes Expression A, then it must also minimize a second expression, *expression B*. He then proves that any μ minimizing Expression B must have a specific form.

Using this knowledge of the form of μ, he simplifies Expression A and then identifies conditions on λ that ensure Expression A is positive.

This condition on lambda is exactly the function of u that we talked about earlier.

6.4.2. The start of the proof

To start the proof, we write the lemma in an equivalent form:

$$C(\mu) = \mathbb{E}_{(p,q)\sim\mu\times\mu}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{p\sim\mu}\left[h_2(p)\right] \geq 0, \tag{6.1}$$

where μ is some probability measure on $[0,1]$, $\mathbb{E}[\mu] \leq u$. And he replaces the function dependent on u with λ, but leaves λ to be defined later.

Then the minimum is attained by some measure (the infimum is reached), this is true because $[0,1]$ is compact, see Theorem 6.8, thus $\mathcal{P}([0,1])$ is compact (see Theorem 6.9). Furthermore, is this expression made out of continuous functions, so the expectation over a continuous function of a sequence of probability measures from a compact space converges.

$C(\mu)$ is very hard to analyse and therefore, We want to analyse instead the following expression:

$$D(\nu) := 2\mathbb{E}_{(p,q)\sim\mu\times\nu}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{q\sim\nu}\left[h_2(q)\right] = \mathbb{E}_{q\sim\nu}\left[2\mathbb{E}_{p\sim\mu}\left[h_2(p+q-pq)\right] - \lambda h_2(q)\right]. \tag{6.2}$$

Later we will see that this expression works rather well for Sawin, because he will be able to split the expectation and the derivatives behave rather nicely. The equality in Equation 6.2 is because in general:

$$\mathbb{E}_{X,Y}[f(X,Y)] = \mathbb{E}_X\left[\mathbb{E}[f(X,Y) \mid X]\right] = \mathbb{E}_Y\left[\mathbb{E}[f(X,Y) \mid Y]\right]. \tag{6.3}$$

And because p and q are independent: $\mathbb{E}[f(p,q) \mid q]$ = $\mathbb{E}_p[f(p,q)]$.

6.4.3. If μ minimizes $C(\mu)$ it also minimizes $D(\mu)$

To be able to use Equation 6.2 we need to prove first that if μ minimizes Equation 6.1, then it will also minimize $D(\nu)$ for all probability measures on $[0,1]$, with $\mathbb{E}[\nu] \leq u$.

To prove this, we proceed by contradiction. Suppose that μ minimizes Equation 6.1, but does not minimize D. Instead, assume that D is minimized by some measure ν. We then consider a convex combination of μ and ν, defined as $\mu' = (1-\varepsilon)\mu+\varepsilon\nu$, where $\varepsilon > 0$ is small. By construction, ν minimizes D, so we expect that $D(\nu) < D(\mu)$ and $C(\mu) < C(\mu')$.

Then we note that expectation is linear, because expectation is defined as: $\mathbb{E}_\rho = \int f(x)d\rho(x)$ and integrals are linear. Thus: $\mathbb{E}_{\mu'}[f] = (1-\varepsilon)\mathbb{E}_\mu[f] + \varepsilon\mathbb{E}_\nu[f]$.

We fill this in Equation 6.1 to obtain:

$$\begin{aligned}
&\mathbb{E}_{(p,q)\sim\mu'\times\mu'}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{p\sim\mu'}\left[h_2(p)\right] \\
&= (1-\varepsilon)^2\mathbb{E}_{(p,q)\sim\mu\times\mu}\left[h_2(p+q-pq)\right] + 2\varepsilon(1-\varepsilon)\mathbb{E}_{(p,q)\sim\mu\times\nu}\left[h_2(p+q-pq)\right] \\
&\quad + \varepsilon^2\mathbb{E}_{(p,q)\sim\nu\times\nu}\left[h_2(p+q-pq)\right] - \lambda(1-\varepsilon)\mathbb{E}_{p\sim\mu}\left[h_2(p)\right] - \lambda\varepsilon\mathbb{E}_{p\sim\nu}\left[h_2(p)\right].
\end{aligned}$$

Then we expand products and organize this equation on powers of ε, to obtain:

$$
\begin{aligned}
&(1-\varepsilon)^2\mathbb{E}_{(p,q)\sim\mu\times\mu}\left[h_2(p+q-pq)\right] + 2\varepsilon(1-\varepsilon)\mathbb{E}_{(p,q)\sim\mu\times\nu}\left[h_2(p+q-pq)\right] \\
&\quad + \varepsilon^2\mathbb{E}_{(p,q)\sim\nu\times\nu}\left[h_2(p+q-pq)\right] - \lambda(1-\varepsilon)\mathbb{E}_{p\sim\mu}\left[h_2(p)\right] - \lambda\varepsilon\mathbb{E}_{p\sim\nu}\left[h_2(p)\right] \\
&= \mathbb{E}_{(p,q)\sim\mu\times\mu}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{p\sim\mu}\left[h_2(p)\right] \\
&\quad + \varepsilon\Big(2\mathbb{E}_{(p,q)\sim\mu\times\nu}\left[h_2(p+q-pq)\right] - 2\mathbb{E}_{(p,q)\sim\mu\times\mu}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{p\sim\mu}\left[h_2(p)\right] + \lambda\mathbb{E}_{p\sim\nu}\left[h_2(p)\right]\Big) \\
&\quad + \varepsilon^2\Big(\mathbb{E}_{(p,q)\sim\mu\times\mu}\left[h_2(p+q-pq)\right] + \mathbb{E}_{(p,q)\sim\nu\times\nu}\left[h_2(p+q-pq)\right]\Big).
\end{aligned}
$$

Now ε was very small, so we simplify the ε^2 term into the error term $\mathcal{O}(\varepsilon^2)$ and we focus on the ε term:

$$
\begin{aligned}
&\mathbb{E}_{(p,q)\sim\mu'\times\mu'}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{p\sim\mu'}\left[h_2(p)\right] \\
&= \mathbb{E}_{(p,q)\sim\mu\times\mu}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{p\sim\mu}\left[h_2(p)\right] \\
&\quad + \varepsilon\Big(2\mathbb{E}_{(p,q)\sim\mu\times\nu}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{p\sim\nu}\left[h_2(p)\right] - 2\mathbb{E}_{(p,q)\sim\mu\times\mu}\left[h_2(p+q-pq)\right] + \lambda\mathbb{E}_{p\sim\mu}\left[h_2(p)\right]\Big) \\
&\quad + \mathcal{O}(\varepsilon^2).
\end{aligned}
$$

We observe that the first ε term is: $\Delta D = D(\nu) - D(\mu)$. $D(\mu)$ is not the minimum, but $D(\nu)$ is, so $\Delta D = D(\nu) - D(\mu) < 0$. But this applies for a small (positive) ε:

$$
\begin{aligned}
&\mathbb{E}_{(p,q)\sim\mu'\times\mu'}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{p\sim\mu'}\left[h_2(p)\right] = \\
&\mathbb{E}_{(p,q)\sim\mu\times\mu}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{p\sim\mu}\left[h_2(p)\right] + \varepsilon\Delta D \\
&< \mathbb{E}_{(p,q)\sim\mu\times\mu}\left[h_2(p+q-pq)\right] - \lambda\mathbb{E}_{p\sim\mu}\left[h_2(p)\right].
\end{aligned}
$$

But this is a contradiction, because μ minimizes $C(\mu)$ but we see here that $C(\mu') < C(\mu)$. Hence, our assumption that μ doesn't minimize $D(\nu)$ must be wrong.

6.4.4. Analysing $F_\mu(q)$

Now that we have proven We define and study now the function:

$$
F_\mu(q) = 2\mathbb{E}_{p\sim\mu}\big[h_2(p+q-pq) - \lambda h_2(q)\big]. \tag{6.4}
$$

And in specific we will analyse:

$$
\frac{d}{dq}G(q) = \frac{d}{dq}\Big(q(1-q)\frac{d^2}{dq^2}F_\mu(q)\Big). \tag{6.5}
$$

Sawin introduces this as a deliberate trick that offers several advantages, without altering the sign of the function. First, it regularizes the boundary behaviour of $h_2''(q) = -\frac{1}{q} - \frac{1}{1-q}$. Second, it yields a derivative with a definite sign. We will see how this becomes useful as we proceed with the proof.

We will calculate $G(p)$ and $\frac{d}{dq}G(q)$, by first calculating the derivatives of all the separate terms. (and

remember $h_2(q) = -q\log(q) - (1-q)\log(1-q)$)

$$\begin{aligned}
\frac{d}{dq}h_2(q) &= -\log(q) + \log(1-q),\\
\frac{d^2}{dq^2}h_2(q) &= -\frac{1}{q} - \frac{1}{1-q},\\
\frac{d}{dq}h_2(p+q-pq) &= (1-p)\ln\!\left(\frac{(1-p)(1-q)}{p+q-pq}\right),\\
\frac{d^2}{dq^2}h_2(p+q-pq) &= -\frac{1-p}{\big(p+q-pq\big)(1-q)}.
\end{aligned}$$

So then:

$$F_\mu(q) = 2\mathbb{E}_{p\sim\mu}\big[h_2(p+q-pq) - \lambda h_2(q)\big], \tag{6.6}$$

$$\frac{d^2}{dq^2}F_\mu(q) = 2\mathbb{E}_{p\sim\mu}\left[\frac{1-p}{(p+q-pq)(1-q)}\right] + \lambda\frac{1}{q(1-q)}, \tag{6.7}$$

$$G(q) = q(1-q)\frac{d^2}{dq^2}F_\mu(q) = -2\mathbb{E}_{p\sim\mu}\left[\frac{(1-p)q}{p+q-pq}\right] + \lambda, \tag{6.8}$$

$$\frac{d}{dq}G(q) = -2\mathbb{E}_{p\sim\mu}\left[\frac{(1-p)p}{(p+q-pq)^2}\right]. \tag{6.9}$$

We now observe that $\frac{d}{dq}G(q) \le 0$. This holds because both $(1-p)p \ge 0$ and $(p+q-pq)^2 \ge 0$. If we disregard the special case where μ is supported only on $\{0,1\}$, then in fact $\frac{d}{dq}G(q) < 0$. This follows from the observation that $p(1-p)$ is a quadratic function with roots at $p=0$ and $p=1$, and attains a maximum at $p=\frac{1}{2}$, where $p(1-p) = \frac{1}{4} > 0$. Therefore, for all $p \in (0,1)$, we have $p(1-p) > 0$. Moreover, the term $(p+q-pq)$ is non-zero for all $p, q \in (0,1)$, except when $p = q = 0$.

Thus $\frac{d}{dq}G(q) < 0$ and therefore $G(q)$ is strictly decreasing. In particular, it either:

- takes positive values on some interval $[0,a)$, zero at a, and negative values on $(a,1]$,
- is positive on all of $[0,1]$,
- or is negative on all of $[0,1]$.

The key is that $\frac{d^2}{dq^2}F_\mu(q)$ is strictly positive or negative when G is positive or negative ($q(1-q) > 0$, when q is not only supported on 1 and 0).

If $F''_\mu(q) > 0$ on an open interval $(0,a)$, then F_μ is strictly convex on the closed interval $[0,a]$. And along the same line: if $F''_\mu(q) < 0$, then F_μ is strictly concave. So we distinguish three cases:

- $F_\mu(q)$ is **strictly convex** on some interval $[0,a]$ and is **strictly concave** on $[a,1]$,
- $F_\mu(q)$ is **strictly concave** on the whole interval $[0,1]$,
- $F_\mu(q)$ is **strictly convex** on the whole interval $[0,1]$.

We then apply Jensen's inequality (see Theorem 3.16) to determine the distribution of the measure's mass.

6.4.5. The strictly concave section (a,1)

For the strictly concave section $[0, a]$, we have: $F(\mathbb{E}[\mu]) > \mathbb{E}[F(\mu)]$. And the expectation of μ is fixed ($\leq u$) so that means that the expectation of F would decrease if we move all the weight of the measure to the sides of the interval, while preserving $\mathbb{E}[\mu] \leq u$. And thus the probability measure can't assign weight to the strictly concave open interval. We provide a proof of this fact.

Remark 6.11 (A probability distribution that minimizes an expression can not assign any mass to an open interval where the expression is strictly concave)**.** *Let F be a strictly concave function on $(a, 1)$. For some $a \in [0, 1)$. And suppose μ is a probability measure on $[0, 1]$ that minimizes:*

$$\mathbb{E}_{q\sim\mu}\big[F(q)\big],$$

subject to the constraint $\mathbb{E}_{q\sim\mu}[q] \leq u$.

Then, μ cannot assign positive mass to any point $q_0 \in (a, 1)$.

Proof. Assume, looking for a contradiction, that there exists a point in the open interval that has mass:

$$q_0 \in (a, 1) \quad \text{with} \quad \mu\big(\{q_0\}\big) > 0.$$

As $a < q_0 < 1$, there is a $t \in (0, 1)$ such that we can write q_0 as a combination of a and 1:

$$q_0 = ta + (1 - t)1.$$

Next, we define a new probability measure:

$$\mu' = \mu - \epsilon\delta_{q_0} + \epsilon\big[t\delta_a + (1 - t)\delta_1\big],$$

where $0 < \epsilon \leq \mu(\{q_0\})$. Clearly, μ' is still a probability measure on $[0, 1]$. Moreover, its mean is:

$$\mathbb{E}_{q\sim\mu'}[q] = \mathbb{E}_{q\sim\mu}[q] - \epsilon q_0 + \epsilon\big(ta + (1 - t)1\big) = \mathbb{E}_{q\sim\mu}[q].$$

So, μ' satisfies the same constraint $\mathbb{E}_{q\sim\mu'}[q] \leq u$. Hence, μ' is an admissible competitor to μ to minimize the expectation.

Next, because F is strictly concave on $(a, 1)$ and $q_0 = ta + (1 - t)1$, we have:

$$F(q_0) > tF(a) + (1 - t)F(1).$$

Therefore, the expectation of F under μ' is:

$$\begin{aligned}
\mathbb{E}_{q\sim\mu'}\big[F(q)\big] &= \mathbb{E}_{q\sim\mu}\big[F(q)\big] - \epsilon F(q_0) + \epsilon\big[tF(a) + (1 - t)F(1)\big], \\
&= \mathbb{E}_{q\sim\mu}\big[F(q)\big] - \epsilon\Big[F(q_0) - \big(tF(a) + (1 - t)F(1)\big)\Big], \\
&< \mathbb{E}_{q\sim\mu}\big[F(q)\big].
\end{aligned}$$

This strictly smaller expectation contradicts the minimality of μ. Hence, no such $q_0 \in (a, 1)$ with $\mu(\{q_0\}) > 0$ can exist. In other words,

$$\mu\big((a, 1)\big) = 0.$$

□

6.4.6. The convex section [0,a]

For the strictly convex portion it holds that: $F(\mathbb{E}[\mu]) \leq \mathbb{E}[F(\mu)]$. Therefore, it can be proven that to minimize the expectation one should put all the mass in one point ($\mathbb{E}[\mu]$). This proof can be found in the appendix at Remark A.9, but the approach is similar to that of the concave section.

6.4.7. Combining the sections

Combining the two arguments, gives that the probability measure can't put weight in the interval $(a, 1)$ and it can only put weight on one point in $[0, a]$. So in total there will be at most mass at 1 and at one other point v in $[0, a]$. We then define w such that μ mass w at v and $1 - w$ at 1, then:

$$\mu(p) = w\delta_v + (1 - w)\delta_1. \tag{6.10}$$

6.4.8. Recap and back to equation

We have shown that if μ minimizes 6.1, then it also minimizes 6.2, and any minimizer of 6.2 must take the form given in 6.10. In what follows, we will use that we know the form of μ in Equation 6.1. Remember the Equation 6.1:

$$C(\mu) = \mathbb{E}_{(p,q)\sim\mu\times\mu}\big[h_2(p + q - pq)\big] - \lambda\mathbb{E}_{p\sim\mu}\big[h_2(p)\big]. \tag{6.11}$$

Then:

$$\mathbb{E}_{p\sim\mu}[h_2(p)] = wh_2(v) + (1 - w)h_2(1) = wh_2(v),$$

$$\begin{aligned}\mathbb{E}_{(p,q)\sim\mu\times\mu}\big[h_2(p + q - pq)\big] &= \sum_{x,y\in\{v,1\}} \mu(x)\mu(y)h_2(x + y - xy),\\ &= w^2h_2(2v - v^2) + 2w(1 - w)h_2(1) + (1 - w)^2h_2(1),\\ &= w^2h_2(2v - v^2),\end{aligned}$$

since $h_2(1) = 0$.

When we substitute this into equation 6.11, we obtain:

$$w^2h_2\big(2v - v^2\big) \geq \lambda wh_2(v) \quad \Longrightarrow \quad w\frac{h_2\big(2v - v^2\big)}{h_2(v)} \geq \lambda.$$

This inequality is increasing in w and still depends on v, so we must perform a worst-case analysis to eliminate these dependencies.

6.4.9. Analysing w and v.

We will first analyse w: The expectation of μ is given by $w \cdot v + (1-w) \cdot 1$. Since the original assumption was that $\mu \leq u$, we have:

$$1 - w(1-v) \leq u \quad \Longrightarrow \quad w \geq \frac{1-u}{1-v}.$$

For a worst-case analysis, we choose the smallest possible value of w, which is $w = \frac{1-u}{1-v}$. Note that $w \geq 0$ since the expectation $u \geq v$, as there is no mass assigned to values less than v. Substituting this expression for w yields:

$$\frac{w h_2(2v - v^2)}{h_2(v)} \geq (1-u) \cdot \frac{h_2(2v-v^2)}{h_2(v)(1-v)}.$$

Next we will analyse v: We again use that $u \geq v$ so v can take any value in $[0, u]$. To ensure that our expression holds for all those possible values of v, we require that the minimal value of the expression over all $v \in [0, a]$ is at least λ:

$$(1-u) \cdot \min_{v \in [0,u]} \left(\frac{h_2(2v-v^2)}{h_2(v)(1-v)} \right) \geq \lambda. \tag{6.12}$$

So far, we have seen how we apply the form of μ to our original inequality, and next we use the worst case analysis twice to delete dependencies.

Sawin then uses another of his lemma's:

Lemma 6.12 (lemma 8 from Sawin)**.** *[28] The function* $\frac{h_2(2v-v^2)}{h_2(v)(1-v)}$ *defined on the interval* $(0,1)$ *is increasing for*

$$v > \frac{3-\sqrt{5}}{2} \quad \textit{and it is decreasing for} \quad v < \frac{3-\sqrt{5}}{2}.$$

So, by Lemma 6.12, Equation 6.12 achieves its minimum on $[0,1]$ at $v^* = \frac{3-\sqrt{5}}{2}$. However, $v \leq u$ and u could be lower than $\frac{3-\sqrt{5}}{2}$.

Therefore, we have two cases: 1. If $u \geq \frac{3-\sqrt{5}}{2}$, then:

$$\begin{aligned} \min_{v \in [0,u]} \left(\frac{h_2(2v-v^2)}{h_2(v)(1-v)} \right) &= \frac{h_2(2v^* - v^{*2})}{h_2(v^*)(1-v^*)}, \\ &= \frac{1}{1-v^*}, \\ &= \frac{1}{1-(3-\sqrt{5})/2} = \frac{2}{2-3+\sqrt{5}} = \frac{2}{\sqrt{5}-1}, \end{aligned}$$

because: $2v^* - v^{*2} = 1 - v^*$, so $h_2\big(2v - v^2\big) = h_2(1-v^*) = h_2(v^*)$.

2. If $u \leq \frac{3-\sqrt{5}}{2}$, then since the function

$$\frac{h_2(2v-v^2)}{h_2(v)(1-v)}$$

is decreasing on the interval $(0, u]$, it attains its minimum value at $v = u$. So:

$$\min_{v\in[0,1]}\left(\frac{h_2(2v-v^2)}{h_2(v)(1-v)}\right)=\frac{h_2(2u-u^2)}{h_2(u)(1-u)}. \tag{6.13}$$

Putting this all together:

$$\min_{v\in[0,u]}\left(\frac{h_2(2v-v^2)}{h_2(v)(1-v)}\right)=\begin{cases}\dfrac{h_2(2u-u^2)}{h_2(u)(1-u)} & \text{if } u\le\frac{3-\sqrt{5}}{2}=v^*,\\ \dfrac{2}{\sqrt{5}-1} & \text{if } u\ge\frac{3-\sqrt{5}}{2}=v^*.\end{cases}$$

We start from the general inequality derived in the proof:

$$\mathbb{E}[h_2(p+q-pq)]\ge\mathbb{E}[h_2(p)]\cdot(1-u)\cdot\min_{v\in[0,1]}\left(\frac{h_2(2v-v^2)}{h_2(v)(1-v)}\right).$$

Substituting this into the earlier inequality gives:

$$\mathbb{E}[h_2(p+q-pq)]\ge\mathbb{E}[h_2(p)]\cdot\begin{cases}\dfrac{h_2(2u-u^2)}{h_2(u)} & \text{if } u\le\frac{3-\sqrt{5}}{2},\\ (1-u)\cdot\dfrac{2}{\sqrt{5}-1} & \text{if } u\ge\frac{3-\sqrt{5}}{2}.\end{cases}$$

7

The kth most frequent element

The Union-Closed Sets Conjecture states that the most frequent element appears in at least half of the sets in a nonempty union-closed family. We thus talk only about the most frequent element. But a natural generalization of this conjecture considers not just the most frequent element, but also the kth most frequent one. In 2022 Nagel proposed such a conjecture, stating that the kth most frequent element appears in at least $\frac{1}{2^{k-1}+1}|\mathcal{F}|$ sets [22].

Das and Wu prove that this conjecture holds for $k \geq 3$ or if $k = 2$ and $|\mathcal{F}| \leq 44$ or $|\mathcal{F}| \geq 114$. The approach from Das and Wu used the entropic method from Gilmer and combinatorial arguments from Knill [18] to achieve these results. But as we have discussed in the previous chapter; there have been made improvements on Gilmer's approach, and we wanted to research if it was possible to apply these improvements on the structure from Das and Wu to tighten the gap between 44 and 114. However, we didn't have enough time to understand the improved argument, and therefore we couldn't apply it to the kth frequent conjecture.

In this chapter we will discuss how Das and Wu use Gilmer's approach to bound the kth most frequent conjecture.

7.1. Notation

We denote by f_l the frequency of the l-th most frequent element in the ground set of the family $\mathcal{F}$.

7.2. The kth most frequent conjecture

Das and Wu word the conjecture as follows:

Conjecture 7.1. *For any union-closed set family $\mathcal{F}$ with $|\mathcal{F}| = m$ and $|\bigcup_{S\in\mathcal{F}} S| \geq k$, the k-th most frequent element lies in at least $\frac{m}{2^{k-1}+1}$ sets in $\mathcal{F}$.*

We observe that this conjecture is the same as conjecture 4.1 for $k = 1$.

7.2.1. A tight bound

We now might ask ourselves why this conjecture takes the specific value $\frac{1}{2^{k-1}+1}$. As just the naive repeated application of conjecture 4.1 would lead to the fraction $\frac{1}{2^k}$. We will now discuss why this fraction is at most $\frac{1}{2^{k-1}+1}$.

We construct a certain family, which Das and Wu call *near-k-cubes*:

Definition 7.2. A **near-k-cube** $\mathcal{C}_k$ is the full power-set of $k-1$ elements together with one additional set that contains at least one other element:

$$\mathcal{C}_k = \mathcal{P}([k-1]) \cup S, \quad \text{s.t. } [k-1] \subsetneq S$$

To give quickly some clarification about the definition of near-k-cubes, we will give a couple of example's:

Example 7.3. Examples of near-k-cubes are:

$$\mathcal{C}_2 = \{\emptyset, \{1\}, \{2\}, \{1,2\}, \{1,2,3\}\} \tag{7.1}$$

$$\mathcal{D}_3 = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}, \{1,2,3,4,5,100\}\} \tag{7.2}$$

Remark 7.4. *Near-k-cubes are a tight bound on the kth most frequent conjecture*

Proof. If the set S has $l = k-1$ elements than the power set $\mathcal{P}(S)$ has 2^{k-1} sets. Hence, a near-k-cube has $2^{k-1}+1$ sets. In this family there are at least k elements and all elements in $[k-1]$ are in exactly $2^{k-2}+1$ sets (see Remark 4.4). The kth most frequent element however is only in 1 set. And thus is the kth most frequent element (which is k) in a fraction $\frac{1}{2^{k-1}+1}$. □

There is no other construction known that gives a tight bound on Nagel's conjecture.

7.2.2. the Union-Closed Sets Conjecture implies the kth most frequent conjecture

In the beginning of this chapter, we talked about the naive repeated application of the Union-Closed Sets Conjecture to obtain the bound $\frac{1}{2^k}$. But there is a way that obtains the fraction $\frac{1}{2^{k-1}+1}$ using conjecture 4.1. And this means that the Union-Closed Sets Conjecture implies Nagel's conjecture. To prove this, we will follow the proof from by Das and Wu [12].

Remark 7.5. *Assuming the Union-Closed Sets Conjecture, Nagel's conjecture is true for all k.*

Before we begin the proof we will sketch the outline of the proof. Given a union□closed family $\mathcal{F}$, such that the groundset is $[n]$ and $n \geq k$. We will first delete the $k-1$ elements that occur most often in $\mathcal{F}$. Removing these elements doesn't change the fact that $\mathcal{F}$ is union closed, so by Frankl's conjecture there is some element e that appears in at least half of these reduced sets.

Subsequently, each reduced set can arise from at most 2^{k-1} original sets (by re□adding any subset of all the $k-1$ deleted elements). Consequently, every reduced set containing e contributes at least one original set with e, while each reduced set not containing e contributes at most 2^{k-1} such originals. A straightforward worst case count yields that e lies in at least a $\frac{1}{1+2^{k-1}}$ fraction of all sets in $\mathcal{F}$.

Proof. We fix $k \in \mathbb{N}$ and let $\mathcal{F}$ be a union-closed family of m sets ($|\mathcal{F}| = m$), which has a ground set $[n]$ such that $n \geq k$. Then we define the set S as the set of the $k-1$ most frequent elements. (and when there are multiple elements that have the same frequency, and they straddle the cutoff for being among those first $k-1$ elements, one may decide which tied elements to include in S and which to leave out arbitrarily.)

Now we consider the 2^{k-1}-to-1 map $g : 2^{[n]} \to 2^{[n]\setminus S}$ defined by $g(F) = F \setminus S$. The family $g(\mathcal{F})$ is still a nonempty union-closed family, because:

It contains a nonempty set because the groundset of $\mathcal{F}$ is bigger than $k-1 = |S|$. So there are still elements 'left' in $\mathcal{F}$ when you take away the first $(k-1)$ most frequent elements. And $g(\mathcal{F})$ **is union-closed**, because for all $A, B \in g(\mathcal{F})$ there exist $F_1, F_2 \in \mathcal{F}$ such that $A = F_1 \setminus S$ and $B = F_2 \setminus S$. And because $\mathcal{F}$ is union-closed: $(F_1 \cup F_2) \in \mathcal{F}$. Then $g(F_1 \cup F_2) = (F_1 \cup F_2) \setminus S = (F_1 \setminus S) \cup (F_2 \setminus S) = A \cup B$. Thus for all $A, B \in g(\mathcal{F})$, $A \cup B \in g(\mathcal{F})$.

Getting back to the proof, we apply Frankl's conjecture to this union-closed family $g(\mathcal{F})$. Evidently, there exist an element e that is at least in $\frac{1}{2}|g(\mathcal{F})| = f_e$. Now we trace the sets in $g(\mathcal{F})$ back to their preimage in $\mathcal{F}$. Then we find the following:

1. For at least half of the sets $A \in g(\mathcal{F})$, $e \in A$, then $g^{-1}(A) = \{F \in \mathcal{F} : F \setminus S = A\} \subseteq \mathcal{F}$. And $|g^{-1}(A)| \geq 1$.
2. For the other sets $B \in g(\mathcal{F})$, which do not contain e, the preimage is: $g^{-1}(B) = \{F \in \mathcal{F} : F \setminus S = B\}$. When we want to bound the size of this preimage we can think of all the different ways subsets of S ($T \subseteq S$) can be added to B such that $(B \cup T) \setminus S = B$. The possibilities for T are maximally all the subsets of S and because $|S| = k-1$, S has 2^{k-1} subsets. So to summarize:

$$|g^{-1}(B)| = |\{F \in \mathcal{F} : F \setminus S = B\}| \leq |\{B \cup T : T \subseteq S\}| = 2^{k-1}.$$

Now we have that at least half of the sets in $g(\mathcal{F})$ have a preimage with at least size 1 that contains e. And at most half of the sets in $g(\mathcal{F})$ have a preimage with at most size 2^{k-1} which do not contain e.

Worst case counting then gives us that element e is in at least:

$$\frac{1}{1+2^{k-1}}|\mathcal{F}| \text{ sets in } \mathcal{F}.$$

And element e is at most the kth most frequent element (because $e \notin S$). Hence, Nagel's conjecture is true. □

7.3. How Das and Wu use the entropy method for the kth most frequent conjecture

When we look back at the proof in Chapter 5, and we take the union-closed family $\mathcal{F}$ and two characteristic vectors (X and Y) of two sets i.i.d. chosen from $\mathcal{F}$ following some distribution. And if for each element $i \in U(\mathcal{F})$, $\Pr(i \in A) \leq \frac{3-\sqrt{5}}{2}$ then:

$$\mathrm{H}(X \cup Y) \geq \mathrm{H}(X).$$

Das and Wu use that this statement can be generalised to a per element form for the infrequent ele-

ments (elements that occur in less than the fraction $\frac{3-\sqrt{5}}{2}$ of sets):

$$\mathrm{H}((X \cup Y)_i | X_{<i}, Y_{<i}) - \mathrm{H}(X_i | X_{<i}) = d \geq 0.$$

However, this inequality doesn't hold for more frequent elements, then d might be negative. A key insight from Das and Wu is to quantify this d to show that the gains made by infrequent elements are bigger than the losses made by frequent elements.

Theorem 7.6. *[12] For any $0 \leq \alpha < \frac{3-\sqrt{5}}{2}$, there exists a constant $c_\alpha \geq 0$ such that if $k \geq 1$ and $|\mathcal{F}| = m \geq 2^{c_\alpha (k-1)}$, then there are always at least k elements in the ground set that appear in at least $\alpha|\mathcal{F}|$ sets in $\mathcal{F}$.*

Proof. [12] Let $\mathcal{F}$ be a union-closed family with at least k elements in its groundset ($|\mathcal{F}| = n \geq k$). Define the fraction of the sets out of $\mathcal{F}$ that contain the k-th most frequent element as f_k. Suppose the opposite of Theorem 7.6: $f_k \leq \alpha$. And we relabel the elements such that $[k-1]$ is the set of the first $k-1$ most frequent elements.
Then we let A, B be two i.i.d. uniformly at random chosen sets out of $\mathcal{F}$ and let $X = X_1, X_2, \ldots, X_n$ and $Y = Y_1, Y_2, \ldots, Y_n$ be random variables representing the characteristic vectors of A and B respectively. So, with the ordering of frequent elements and the hypothesis $f_k \leq \alpha$, we have for all $i \geq k$, $\Pr(X_i = 1) \leq \alpha$.
Having assumed that $f_k \leq \alpha$, we now want to prove a contradiction by proving $\mathrm{H}(X \cup Y) \geq \mathrm{H}(X)$, which contradicts the uniform bound (just like in Chapter 5).

We start similar to the other proof in Chapter 5:

$$\mathrm{H}(X \cup Y) = \sum_{i=1}^{n} \mathrm{H}((X \cup Y)_i \mid (X \cup Y)_{<i}) \geq \sum_{i=1}^{n} \mathrm{H}((X \cup Y)_i \mid X_{<i}, Y_{<i}) \tag{7.3}$$

Then we will split this sum into two parts: the part of the sum with *frequent elements* (the first $k-1$ elements) and the part of the sum with *infrequent elements* (the other elements from k to n):

$$\begin{aligned}\mathrm{H}(X \cup Y) &\geq \sum_{i=1}^{n} \mathrm{H}((X \cup Y)_i \mid X_{<i}, Y_{<i}), \\ &= \sum_{i=1}^{k-1} \mathrm{H}((X \cup Y)_i \mid X_{<i}, Y_{<i}) + \sum_{i=k}^{n} \mathrm{H}((X \cup Y)_i \mid X_{<i}, Y_{<i}).\end{aligned}$$

On these two different parts we will use two different lemmas, the first lemma is from Sawin which we (partly) proved in Chapter 6. The second lemma is from Das and Wu.

Lemma 7.7. *[28] Let X, Y be i.i.d. distributions on the family $2^{[n]}$. Suppose further that $\mathbb{E}[X_i] \leq \alpha < \frac{3-\sqrt{5}}{2}$. Then:*

$$\mathrm{H}((X \cup Y)_i \mid X_{<i}, Y_{<i}) - \lambda_\alpha \, \mathrm{H}(X_i \mid X_{<i}) \geq 0,$$

where $\lambda = \frac{\mathrm{H}(2\alpha - \alpha^2)}{\mathrm{H}(\alpha)} > 1$.

Lemma 7.8. *[12] Let X, Y be i.i.d. distributions on the family $2^{[n]}$. Then for any $\lambda \geq 0$:*

$$\mathrm{H}((X \cup Y)_i \mid X_{<i}, Y_{<i}) - \lambda \, \mathrm{H}(X_i \mid X_{<i}) \geq - \max_{0 \leq x \leq 1} \left(\lambda h_2(x) - h_2(x^2) \right).$$

We now apply Lemma 7.7 to elements k up to n and we apply Lemma 7.8 to elements 1 up to $k-1$:

$$\begin{aligned}
\mathrm{H}(X \cup Y) &\geq \sum_{i=1}^{k-1} \mathrm{H}((X \cup Y)_i \mid X_{<i}, Y_{<i}) + \sum_{i=k}^{n} \mathrm{H}((X \cup Y)_i \mid X_{<i}, Y_{<i}), \\
&\geq \sum_{i=1}^{k-1} \left(\lambda_\alpha \, \mathrm{H}(X_i \mid X_{<i}) - \max_{0 \leq x \leq 1} (\lambda_\alpha h_2(x) - h_2(x^2)) \right) + \sum_{i=k}^{n} \lambda_\alpha \, \mathrm{H}(X_i \mid X_{<i}), \\
&= -\sum_{i=1}^{k-1} \max_{0 \leq x \leq 1} (\lambda_\alpha h_2(x) - h_2(x^2)) + \lambda_\alpha \sum_{i=1}^{n} \mathrm{H}(X_i \mid X_{<i}), \\
&= -(k-1) \max_{0 \leq x \leq 1} (\lambda_\alpha h_2(x) - h_2(x^2)) + \lambda_\alpha \, \mathrm{H}(X).
\end{aligned}$$

We need $\mathrm{H}(X \cup Y) \geq \mathrm{H}(X)$, thus we need:

$$-(k-1) \max_{0 \leq x \leq 1} (\lambda_\alpha h_2(x) - h_2(x^2)) + \lambda_\alpha \, \mathrm{H}(X) \geq \mathrm{H}(X)$$

or (rewritten)

$$(\lambda_\alpha - 1) \, \mathrm{H}(X) \geq (k-1) \max_{0 \leq x \leq 1} (\lambda_\alpha h_2(x) - h_2(x^2)).$$

X is uniformly distributed over all possible characteristic vectors and there are $|\mathcal{F}|$ characteristic vectors, so $\mathrm{H}(X) = \log |\mathcal{F}|$ (Remark 3.17).

So, if $|\mathcal{F}| \geq 2^{c_\alpha (k-1)}$, where $c_\alpha = \max_{x \in [0,1]} (\lambda_\alpha h_2(x) - h_2(x^2))/(\lambda_\alpha - 1)$ then:

$$\begin{aligned}
(\lambda_\alpha - 1) \, \mathrm{H}(X) &\geq (\lambda_\alpha - 1) \log(2^{c_\alpha (k-1)}), \\
&= (\lambda_\alpha - 1) c_\alpha (k-1), \\
&= (k-1) \max_{0 \leq x \leq 1} (\lambda_\alpha h_2(x) - h_2(x^2)).
\end{aligned}$$

And $c_\alpha \geq 0$, because $c_\alpha = \max_{x \in [0,1]} (\lambda_\alpha h_2(x) - h_2(x^2))/(\lambda_\alpha - 1) \geq (\lambda_\alpha h_2(0) - h_2(0^2))/(\lambda_\alpha - 1) = 0$. (because $\lambda_\alpha > 0$).

And thus: $\mathrm{H}(X \cup Y) \geq \mathrm{H}(X)$. Even though as discussed earlier we also have: $\mathrm{H}(X \cup Y) < \mathrm{H}(X)$ so we have a contradiction. Hence, our hypothesis $f_k \leq \alpha$ is wrong. on the conditions that:

$$|\mathcal{F}| \geq 2^{c_\alpha (k-1)},$$

$$c_\alpha = \frac{\max_{x \in [0,1]} (\lambda_\alpha h_2(x) - h_2(x^2))}{\lambda_\alpha - 1}.$$

□

We can now use this to calculate the bounds numerically.

8

Conclusion and Further Research

The Union-Closed Sets Conjecture remains one of the most intriguing and challenging open problems in combinatorics. In this thesis, we explored the novel entropy-based approach pioneered by Gilmer, which has led to major breakthroughs on the conjecture itself as well as the related Nagel's conjecture. Specifically, the entropy method establishes a constant lower bound on the frequency of the most frequent element in any finite union-closed family, and it improved the bound on the size of a family for Nagel's conjecture to be true.

This thesis was written for undergraduate students in mathematics, as well as physicists with a strong mathematical interest. Accordingly, we aimed to provide thorough explanations of all key concepts. We introduced the notion of entropy, examined its applications in combinatorics, and applied it to the Union-Closed Sets Conjecture. In doing so, we developed an intuitive understanding of how entropy can be leveraged as a powerful tool in discrete mathematics.

In particular, we proved that in every finite union-closed family of sets, there exists at least one element that appears in at least a $\frac{3-\sqrt{5}}{2} \approx 0.382$ fraction of the sets.

We then briefly discussed Sawin's idea to use coupling between random variables to further improve the bound. This approach was refined by Cambie and Yu, who achieved a slightly stronger bound of approximately 0.38234, and later improved by Liu to approximately 0.38271 by considering a different form of coupling. Though the full details of these proofs fall beyond the scope of this thesis.

We did, however, present an important lemma from Sawin, which taught us an alternative method for using entropy in combinatorics. This lemma was later applied in our discussion of Nagel's conjecture.

Finally, we explored this conjecture posed by Nagel, which asserts that the kth most frequent element in a union-closed family appears in at least a $\frac{1}{2^{k-1}+1}$ fraction of the sets. We examined how Das and Wu utilized the entropy method and Sawin's lemma to improve bounds on this conjecture.

8.1. Further Research

Although this thesis does not fully encompass the most recent progress on the Union-Closed Sets Conjecture, it raises several interesting directions for further research:

- We have seen that the Union-Closed Sets Conjecture implies the kth most frequent element conjecture. It remains an open question whether the reverse implication also holds.
- While entropy has proven to be a remarkably effective tool in these proofs, it is natural to ask: does it necessarily have to be the entropy function? This thesis briefly discussed why entropy is suited for the work in combinatorics. However, it would be interesting to research if the defining properties are truly essential for the arguments used in the proof by Gilmer and others.
- In our discussion of Nagel's conjecture, we used Sawin's lemma and the bound of $\frac{1}{2^{k-1}+1}$. However, the improved constants obtained by Cambie, Yu, and Liu were not incorporated into this approach. It would be valuable to investigate:
    1. Whether even a modest improvement in the constant significantly improves the bound in Nagel's setting.
    2. How the improved bounds could be applied to the structure used by Das and Wu.

References

[1] Mark Aldridge. *The Union-Closed Sets Conjecture*. Accessed: 2025-05-24. 2019. URL: https://mpaldridge.github.io/blog/union-closed.html.

[2] Ryan Alweiss, Brice Huang, and Mark Sellke. *Improved Lower Bound for Frankl's Union-Closed Sets Conjecture*. 2024. arXiv: 2211.11731 [math.CO]. URL: https://arxiv.org/abs/2211.11731.

[3] John C. Baez, Tobias Fritz, and Tom Leinster. "A Characterization of Entropy in Terms of Information Loss". In: *Entropy* 13.11 (Nov. 2011), pp. 1945–1957. ISSN: 1099-4300. DOI: 10.3390/e13111945. URL: http://dx.doi.org/10.3390/e13111945.

[4] Patrick Billingsley. *Probability and Measure*. 3rd. John Wiley & Sons, 1995.

[5] Ravi B. Boppana. *A Useful Inequality for the Binary Entropy Function*. 2023. arXiv: 2301.09664 [math.CO]. URL: https://arxiv.org/abs/2301.09664.

[6] L. M. Bregman. "Some properties of nonnegative matrices and their permanents". In: *Soviet Mathematics Doklady* 14 (1973), pp. 945–949.

[7] Henning Bruhn and Oliver Schaudt. *The journey of the union-closed sets conjecture*. 2013. arXiv: 1309.3297 [math.CO]. URL: https://arxiv.org/abs/1309.3297.

[8] Stijn Cambie. *Better bounds for the union-closed sets conjecture using the entropy approach*. 2025. arXiv: 2212.12500 [math.CO]. URL: https://arxiv.org/abs/2212.12500.

[9] Tom Carter et al. "An Introduction to Information Theory and Entropy". In: (June 2000).

[10] Zachary Chase and Shachar Lovett. *Approximate union closed conjecture*. 2022. arXiv: 2211.11689 [math.CO]. URL: https://arxiv.org/abs/2211.11689.

[11] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd. Wiley-Interscience, 2006. Chap. 5.

[12] Shagnik Das and Saintan Wu. *Frequent elements in union-closed set families*. 2024. arXiv: 2412.03862 [math.CO]. URL: https://arxiv.org/abs/2412.03862.

[13] Richard M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

[14] Emory University Math Center. *Proof of Rolle's Theorem*. Accessed: 2025-06-03. URL: https://mathcenter.oxford.emory.edu/site/math111/proofs/rollesTheorem/.

[15] David Galvin. *Three tutorial lectures on entropy and counting*. 2014. arXiv: 1406.7872 [math.CO]. URL: https://arxiv.org/abs/1406.7872.

[16] Justin Gilmer. *A constant lower bound for the union-closed sets conjecture*. 2022. arXiv: 2211.09055 [math.CO]. URL: https://arxiv.org/abs/2211.09055.

[17] Aleksandr Yakovlevich Khinchin. *Mathematical Foundations of Information Theory*. Translation by Silverman, R. A. and Friedman, M. D. of two papers originally published in *Uspekhi*; 120 pp. New York: Dover Publications, 1957. ISBN: 9780486604343. URL: https://archive.org/details/khinchin-mathematical-foundations-of-information-theory.

[18] Emanuel Knill. *Graph generated union-closed families of sets*. 1994. arXiv: math/9409215 [math.CO]. URL: https://arxiv.org/abs/math/9409215.

[19] Jingbo Liu. *Improving the Lower Bound for the Union-closed Sets Conjecture via Conditionally IID Coupling*. 2023. arXiv: 2306.08824 [cs.IT]. URL: https://arxiv.org/abs/2306.08824.

[20] David J.C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press, 2003, p. 81.

[21] MIT OpenCourseWare. *Lecture Notes, Chapter 10*. Lecture Notes for Course 18.218, Spring 2019. Accessed via MIT DSpace. Massachusetts Institute of Technology, 2019.

[22] Nicolas Nagel. *Notes on the Union Closed Sets Conjecture*. 2023. arXiv: 2208.03803 [math.CO]. URL: https://arxiv.org/abs/2208.03803.

[23] Daniel Naylor. *Electromagnetism and Classical Field Theory*. Lecture Notes. Accessed online. University of Cambridge, 2024.

[24] Luke Pebody. *Extension of a Method of Gilmer*. 2022. arXiv: 2211.13139 [math.CO]. URL: https://arxiv.org/abs/2211.13139.

[25] Bjorn Poonen. "Union-closed families". In: *Journal of Combinatorial Theory, Series A* 59.2 (1992), pp. 253–268. ISSN: 0097-3165. DOI: https://doi.org/10.1016/0097-3165(92)90068-6. URL: https://www.sciencedirect.com/science/article/pii/0097316592900686.

[26] José C. Principe. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. 1st ed. Information Science and Statistics. Springer New York, 2010. ISBN: 9781441915696. DOI: 10.1007/978-1-4419-1570-2.

[27] Jaikumar Radhakrishnan. "An entropy proof of Bregman's theorem". In: *Journal of Combinatorial Theory, Series A* 77.1 (1997), pp. 161–164. DOI: 10.1006/jcta.1996.2813.

[28] Will Sawin. *An improved lower bound for the union-closed set conjecture*. 2023. arXiv: 2211.11504 [math.CO]. URL: https://arxiv.org/abs/2211.11504.

[29] Martin Schlather and Carmen Ditscheid. "An Intrinsic Characterization of Shannon's and Rényi's Entropy". In: *Entropy* 26.12 (2024). ISSN: 1099-4300. DOI: 10.3390/e26121051. URL: https://www.mdpi.com/1099-4300/26/12/1051.

[30] Claude E. Shannon. "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3 (1948), pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[31] Lei Yu. "Dimension-Free Bounds for the Union-Closed Sets Conjecture". In: *Entropy* 25.5 (May 2023), p. 767. ISSN: 1099-4300. DOI: 10.3390/e25050767. URL: http://dx.doi.org/10.3390/e25050767.

[32] Fabio Massimo Zennaro. *Shannon's Entropy, Rényi's Entropy, and Minimum Error Entropy*. Talk, Information□Theoretic Learning. Available at https://fmzennaro.github.io/talks/ITL.pdf.

A

Additional proofs

A.1. Limit of xlog(x)

Remark A.1.

$$\lim_{x \to 0^+} x \log x = 0,$$

where we use 0^+ *because the logarithm function is only defined for positive values.*

To evaluate $\lim_{x \to 0^+} x \log x$*, we rewrite it as:*

$$x \log x = \frac{\log x}{1/x}.$$

Next we apply L'Hôpital's Rule, which states that if a limit yields an indeterminate form, we may take derivatives of the numerator and denominator:

$$\lim_{x \to 0^+} \frac{\log x}{1/x} = \lim_{x \to 0^+} \frac{\frac{d}{dx} \log x}{\frac{d}{dx} 1/x} = \lim_{x \to 0^+} \frac{1/x}{-1/x^2} = \lim_{x \to 0^+} -x = 0.$$

Hence, $\lim_{x \to 0^+} x \log x = 0$.

A.2. Entropy of fully dependent random variables

Remark A.2. *Two fully dependent random variables share the same entropy and their joint entropy is equal to this.*

Proof. Let X and Y be two random variables that are fully dependent on each other, i.e. there exists a bijective function f such that:

$$Y = f(X) \quad \text{and} \quad X = f^{-1}(Y).$$

Next, by the chain rule for entropy:

$$\mathrm{H}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y \mid X).$$

However, since $Y = f(X)$ is a deterministic function of X, we have $\mathrm{H}(Y \mid X) = 0$, so:

$$\mathrm{H}(X, Y) = \mathrm{H}(X).$$

We can then use the same procedure to obtain:

$$\begin{aligned}\mathrm{H}(X, Y) &= \mathrm{H}(Y) + \mathrm{H}(X \mid Y),\\ &= \mathrm{H}(Y) + 0.\end{aligned}$$

Combining these two expressions for $\mathrm{H}(X, Y)$ yields:

$$\mathrm{H}(X) = \mathrm{H}(Y) = \mathrm{H}(X, Y).$$

□

A.3. proof of the concavity of entropy

Although the main text provided a concise overview of the proof's structure, we now present a detailed argumentation. We will shortly restate the remark and the beginning of the proof.

Remark A.3. ***Concavity of entropy*** *Let X_1 and X_2 be discrete random variables with respective probabilities distributions μ_{X_1} and μ_{X_2}, and let $\lambda \in [0, 1]$. Define a new random variable X with exactly $\mu_X = \lambda\mu_{X_1} + (1 - \lambda)\mu_{X_2}$ Then the entropy of X satisfies:*

$$H(X) \geq \lambda H(X_1) + (1 - \lambda)H(X_2),$$

with equality if and only if X_1 and X_2 have the same distribution.

Proof. The entropy of a random variable X with finite outcome space $\mathbf{E}$ is defined as:

$$\begin{aligned}\mathrm{H}(X) &= -\sum_{x \in \mathbf{E}} p_x \log p_x,\\ &= \sum_{x \in \mathbf{E}} f(p_x),\end{aligned}$$

where $f : [0, 1] \to \mathbb{R}$ s.t. $f(p_x) = -p_x \log(p_x)$. We differentiate $f(p_x)$ twice with respect to p_x to obtain on the interval $p_x \in (0, 1]$:

$$\frac{d^2}{d(p_x)^2} f(p_x) = -\frac{1}{p_x} < 0.$$

We note that the second derivative is strictly negative on the interval $(0, 1]$ and therefore f is a strictly concave function. Thus, for every $x \in \mathbf{E}$, $p_x, q_x \in (0, 1]$ s.t. $p_x \neq q_x$ and $\forall\lambda \in (0, 1)$:

$$f\big(\lambda p_x + (1 - \lambda)q_x\big) > \lambda\, f(p_x) + (1 - \lambda)\, f(q_x).$$

And trivially: $\forall x \in \mathbf{E}$, $p_x, q_x \in (0, 1]$ s.t. $p_x = q_x$,

$$f\big(\lambda p_x + (1 - \lambda)q_x\big) = \lambda\, f(p_x) + (1 - \lambda)\, f(q_x).$$

Now what if $p_x = 0$ or $q_x = 0$? WLOG we assume that $q_x = 0$, then:

$$f\big(\lambda p_x + (1-\lambda)q_x\big) = f\big(\lambda p_x\big) = -\lambda p_x \log(\lambda p_x) = -\lambda p_x \log(p_x) - \lambda p_x \log(\lambda) > \lambda p_x \log(p_x) = \lambda\, f(p_x) + (1-\lambda)\, f(q_x).$$

Where are reminded that $f(0) = 0$ by continuity, see Remark A.1. And $-\lambda p_x \log(\lambda) > 0$, because, $\lambda < 1$ and $\lambda p_x > 0$. So if $\exists x \in \mathbf{E}$ s.t. $p_x \neq q_x$ we have:

$$\sum_{x\in\mathbf{E}} f\big(\lambda p_x + (1-\lambda)q_x\big) > \sum_{x\in\mathbf{E}} \big(\lambda\, f(p_x) + (1-\lambda)\, f(q_x)\big). \tag{A.1}$$

Equivalently, in terms of random variables and entropy, let X_1 and X_2 be discrete random variables taking values in $\mathbf{E}$ with probability mass functions

$$p_x = \Pr(X_1 = x), \quad q_x = \Pr(X_2 = x).$$

For $\lambda \in (0, 1)$, define the mixture variable X by

$$\Pr(X = x) \;=\; p_X(x) \;=\; \lambda\, p_x \;+\; (1-\lambda)\, q_x.$$

Then, setting $f(u) = -u \log u$, we have

$$\mathrm{H}(X) = \sum_{x\in\mathbf{E}} f\big(\lambda p_x + (1-\lambda)q_x\big) > \sum_{x\in\mathbf{E}} \big(\lambda\, f(p_x) + (1-\lambda)\, f(q_x)\big) = \lambda\, \mathrm{H}(X_1) + (1-\lambda)\, \mathrm{H}(X_2).$$

Hence $H(X)$ is strictly concave as a function of the underlying distribution. □

A.4. Jensen's inequality

Remark A.4. ***Jensen's inequality****. Let f be a concave function and let X be a real-valued random variable with finite expectation, then:*

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)],$$

where the inequality flips for convex functions. And for strictly concave functions h and random variable Y, that isn't certain (it has at least two outcomes with non-zero probability):

$$h(\mathbb{E}[Y]) > \mathbb{E}[h(Y)],$$

where the inequality flips for convex functions.

Proof. Let X be a bounded random variable s.t. $X : \Omega(\omega) \to \mathbf{D}$ and let f be a continuous strictly concave function: $\mathbf{D} \to \mathbb{R}$, then by definition: for all points ($x_0 \in D$) the tangent at that point is always above the function f:

$$f(x) < f(x_0) + f'(x_0)(x - x_0) \qquad \forall x \in \mathbf{D} \setminus \{x_0\}.$$

Now we can choose x_0 freely so we choose $x_0 = \mathbb{E}[X]$, consequently:

$$f(x) < f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(x - \mathbb{E}[X]) \qquad \forall x \in D \setminus \mathbb{E}[X].$$

This inequality holds for all $x \in \mathbf{D} \setminus \{\mathbb{E}[X]\}$. Thus, this inequality holds for all values that X might take except if it takes the value of its own expectation ($x = \mathbb{E}[X]$), then the equation is clearly equal. The

expectation over $f(x)$ can be written as: $\sum_{x\in\mathbf{D}} \Pr(X=x)f(x)$. We have shown that for all $x \neq \mathbb{E}[X]$, $f(x) < f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(x - \mathbb{E}[X])$. Thus, if $\exists x \in \mathbf{D} \setminus \mathbb{E}[X]$ s.t. $\Pr(X=x) > 0$:

$$\sum_{x\in\mathbf{D}\setminus\mathbb{E}[X]} \Pr(X=x)f(x) < \sum_{x\in\mathbf{D}\setminus\mathbb{E}[X]} \Pr(X=x)(f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(x-\mathbb{E}[X])).$$

At $x = \mathbb{E}[X]$, $f(x) = f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(x - \mathbb{E}[X])$. So if we add that equation to both sides, the inequality doesn't change the inequality.

$$\sum_{x\in\mathbf{D}} \Pr(X=x)f(x) < \sum_{x\in\mathbf{D}} \Pr(X=x)(f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(x-\mathbb{E}[X])).$$

We can write this equation alternatively as the following:

$$\begin{aligned}\mathbb{E}[f(X)] &< \mathbb{E}\big[f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(X-\mathbb{E}[X])\big], \qquad \forall X \neq \mathbb{E}[X],\\ &= f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(\mathbb{E}[X]-\mathbb{E}[X]),\\ &= f(\mathbb{E}[X]).\end{aligned}$$

And quite logically if $X = \mathbb{E}[X]$, then:

$$\mathbb{E}[f(X)] = f(\mathbb{E}[X]).$$

For a convex function, $g = -f$ it is clear that the inequality flips. □

A.5. The limit of f(x,y) along the corner points of the boundary

Remark A.5. $\lim_{x,y\to(0,0)} f(x,y) = \lim_{x,y\to(1,1)} f(x,y) = 1$, *where* $f(x,y) : [0,1]^2 \to \mathbb{R}_{\geq 0}$ *is defined as follows:*

$$f(x,y) = \frac{h_2(xy)}{h_2(x)y + h_2(y)x} = \frac{h(xy)}{h(x)y + h(y)x},$$

where h_2 *is the base two binary entropy function. And* h *is the binary entropy function with the natural logarithm.*

Proof. Substituting the Taylor expansion

$$\ln(1-u) \;=\; -u \;-\; \tfrac{1}{2}u^2 \;-\; \tfrac{1}{3}u^3 \;-\; \cdots$$

into the formula for $h(u)$, we get:

$$h(u) = -u\ln u - (1-u)\big(-u - \tfrac{1}{2}u^2 - \cdots\big) = -u\ln u \;+\; u \;+\; \tfrac{1}{2}u^2 \;+\; \mathcal{O}(u^3).$$

which is equal to:

$$h(u) = -u\ln u(1 + \mathcal{O}(u)).$$

As $(x,y) \to (0,0)$, it follows that:

$$f(x,y) = \frac{h(xy)}{xh(y) + yh(x)} = \frac{-xy\ln(xy)\,(1+\mathcal{O}(u))}{-xy(\ln x + \ln y)\,(1+\mathcal{O}(u))} = \frac{\ln(xy)}{\ln x + \ln y}\,(1+\mathcal{O}(u)) \;\longrightarrow\; 1,$$

Therefore, $\lim_{(x,y)\to(0,0)} f(x,y) = 1$.

For $\lim_{(x,y)\to(1,1)} f(x,y)$ it can be shown using similar logic that $\lim_{(x,y)\to(1,1)} f(x,y) = 1$. □

A.6. G(x) is strictly increasing

Remark A.6 ($G(x)$ is strictly increasing)**.** *The function*

$$G(x) = -\frac{\ln(1-x)}{x}$$

is strictly increasing on $(0,1)$*.*

Proof. We compute $G'(x)$ using the quotient rule.

$$G'(x) = \frac{N'(x)D(x) - N(x)D'(x)}{\big[D(x)\big]^2} = \frac{\frac{x}{1-x} + \ln(1-x)}{x^2}, \tag{A.2}$$

$$= \frac{1}{x^2}\Big(\frac{x}{1-x} + \ln(1-x)\Big). \tag{A.3}$$

Because $1/x^2 > 0$ for $x \in (0,1)$, the sign of $G'(x)$ is determined by:

$$l(x) = \frac{x}{1-x} + \ln(1-x).$$

We will prove for $x \in (0,1)$ that $l(x) > 0$ and thus $G'(x) > 0$.

$$l'(x) = \frac{x}{(1-x)^2} > 0 \quad x \in (0,1).$$

Thus l is strictly increasing. And $l(0) = 0$, so $l(x) > 0$ for $x \in (0,1)$. It follows that:

$$G'(x) > 0 \quad \forall x \in (0,1),$$

So, G is strictly increasing on $(0,1)$. □

A.7. the reverse direction Rolle's theorem

Rolle's theorem is the idea that for a continuous function $g : \mathbb{R} \to \mathbb{R}$:

$$\exists a, b \in \mathbb{R} \text{ s.t. } g(a) = g(b) \implies \exists c \in (a,b) \text{ s.t. } g'(c) = 0.$$

This is a quite intuitive theorem, but a formal proof can be found here: [14].

Now if we want to use it in reverse order on this problem we can rephrase Rolle's theorem into:

Theorem A.7. *If $g(x)$ has k zero's on the interval, (a,b) then the antiderivative $G(x)$ has at most $k+1$ zero's.*

Proof. This is a proof by contradiction: Suppose, that G has $k+2$ distinct zeros in the interval I, such

that:

$$a_1 < a_2 \cdots < a_{k+2}, \quad \text{with } G(a_i) = 0.$$

Then G vanishes at $k+2$ distinct points. By Rolle's theorem, between each pair (a_i, a_{i+1}), there exists a point $c_i \in (a_i, a_{i+1})$ such that

$$G'(c_i) = 0.$$

Since there are $k+1$ such pairs, we obtain $k+1$ distinct points $c_0, \ldots, c_k$ in I where $G'(c_i) = g(c_i) = 0$. However, $g(x)$ had only k zeros. This is a contradiction. □

A.8. The roots of v

Remark A.8. *The function $v(x)$:*

$$\begin{aligned} v(x) &= h(x^2) - \varphi x h(x) = \ln(2) h_2(x^2) - \varphi x \ln(2) h_2(x), \\ &= \big[-x^2 \ln(x^2) - (1-x^2)\ln(1-x^2) \big] - \varphi x \big[-x \ln x - (1-x)\ln(1-x) \big], \end{aligned}$$

has double roots at 0 and φ^{-1} and a single root at 1.

Proof. $v(x)$ has a double zero at (0) as $v(0) = 0$ (as $h(0) = 0$) and:

$$\begin{aligned} \lim_{x\to 0} v'(x) &= \lim_{x\to 0} 2x \ln(|x^2-1|) + 2\varphi x \ln(|x|) - 2x \ln(x^2) - 2\varphi \ln(|x-1|)\, x + \varphi \ln(|x-1|), \\ &= \lim_{x\to 0} \quad 0 \quad + 2\varphi x \ln(|x|) - 4x \ln(x) \; - \quad 0 \quad + \quad 0, \\ &= 0. \end{aligned}$$

v has a double root at φ^{-1}. Some useful identities are: $1-\varphi^{-2} = \varphi^{-1}$, $1-\varphi^{-1} = \varphi^{-2}$ and $1-\varphi = -\varphi^{-1}$. We can use these to evaluate $v(\varphi^{-1})$ and $v'(\varphi^{-1})$.

$$\begin{aligned} v(\varphi^{-1}) &= -\varphi^{-2}\ln(\varphi^{-2}) - (1-\varphi^{-2})\ln(1-\varphi^{-2}) \; + \; \varphi\varphi^{-1}\left[\varphi^{-1}\ln\varphi^{-1} + (1-\varphi^{-1})\ln(1-\varphi^{-1})\right], \\ &= -\varphi^{-2}\ln(\varphi^{-2}) - \varphi^{-1}\ln(\varphi^{-1}) \; + \; \varphi^{-1}\ln(\varphi^{-1}) + \varphi^{-2}\ln(\varphi^{-2}), \\ &= 0. \\ v'(\varphi^{-1}) &= 2\varphi^{-1}\ln(|\varphi^{-2}-1|) + 2\varphi\varphi^{-1}\ln(|\varphi^{-1}|) - 2\varphi^{-1}\ln(\varphi^{-2}) - 2\varphi\ln(|\varphi^{-1}-1|)\,\varphi^{-1} + \varphi\ln(|\varphi^{-1}-1|), \\ &= 2\varphi^{-1}\ln(|\varphi^{-2}-1|) + 2\ln(|\varphi^{-1}|) - 2\varphi^{-1}\ln(\varphi^{-2}) - 2\ln(|\varphi^{-1}-1|) + \varphi\ln(|\varphi^{-1}-1|), \\ &= 2\varphi^{-1}\ln(\varphi^{-1}) + 2\ln(\varphi^{-1}) - 2\varphi^{-1}\ln(\varphi^{-2}) - 2\ln(\varphi^{-2}) + \varphi\ln(\varphi^{-2}), \\ &= -2\varphi^{-1}\ln(\varphi) - 2\ln(\varphi) + 4\varphi^{-1}\ln(\varphi) + 4\ln(\varphi) - 2\varphi\ln(\varphi), \\ &= 2\varphi^{-1}\ln(\varphi) + 2\ln(\varphi) - 2\varphi\ln(\varphi), \\ &= 2\varphi^{-1}\ln(\varphi) + (2-2\varphi)\ln(\varphi) = 2\varphi^{-1}\ln(\varphi) - 2\varphi^{-1}\ln(\varphi), \\ &= 0. \end{aligned}$$

And lastly, does v have a single root at 1, as $h(1) = 0$. (and $v'(1) \neq 1$) □

A.9. Probability distribution on a strictly convex section [0,a]

Remark A.9 (A probability distribution that minimizes a convex expression over a closed interval will assign all its mass to one value)**.** *Let F be a strictly convex function on $[0, a]$. For some $0 < a \leq 1$. Suppose μ is a probability measure on $[0, 1]$ that minimizes*

$$\mathbb{E}_{q\sim\mu}\big[F(q)\big]$$

subject to the constraint $\mathbb{E}_{q\sim\mu}[q] \leq u$. Then, $\mu\big([0, a]\big)$ must be concentrated at a single point (i.e. μ must be an atom in $[0, a]$).

Proof. Assume, for contradiction, that μ places positive mass on two distinct points. That is, there exist:

$$p_1, p_2 \in [0, a], \quad p_1 \neq p_2, \quad \text{and} \quad \mu\big(\{p_1\}\big), \mu\big(\{p_2\}\big) > 0.$$

Then we define λ as:

$$\lambda = \frac{\mu\big(\{p_1\}\big)}{\mu\big(\{p_1\}\big) + \mu\big(\{p_2\}\big)}.$$

And we define the waited average of the two points p_1 and p_2 as m.

$$m = \frac{p_1(\mu\big(\{p_1\}\big)) + p_2(\mu\big(\{p_2\}\big))}{\mu\big(\{p_1\}\big) + \mu\big(\{p_2\}\big)} = \lambda p_1 + (1 - \lambda)p_2.$$

Since F is strictly convex on $[0, a]$, we have:

$$F(m) < \lambda F(p_1) + (1 - \lambda)F(p_2).$$

Thus we define a new probability measure μ' by merging p_1 and p_2 into the single point m. Such that:

$$\mu' = (\mu\big(\{p_1\}\big) + \mu\big(\{p_2\}\big))\delta_m.$$

Clearly μ' is still a probability measure on $[0, 1]$. Moreover, its mean is:

$$\mathbb{E}_{q\sim\mu'}[q] = m = \mathbb{E}_{q\sim\mu}[q].$$

Hence, μ' satisfies the same constraint $\mathbb{E}_{q\sim\mu'}[q] \leq u$, so it is an admissible competitor.

Next, we compare the expectations of F under μ and μ', while remembering the convexity inequality.

$$\mathbb{E}_{q\sim\mu'}\big[F(q)\big] = (\mu\big(\{p_1\}\big) + \mu\big(\{p_2\}\big))F(m) < \mu\big(\{p_1\}\big)F(p_1) + \mu\big(\{p_2\}\big)F(p_2) = \mathbb{E}_{q\sim\mu}\big[F(q)\big]. \tag{A.4}$$

This contradicts the assumption that μ minimizes $\mathbb{E}[F(q)]$. Therefore, it is impossible for μ to assign positive mass to two distinct points in $[0, a]$. □

B

The proof that the logarithmic form of entropy is characterized by four basic properties

There are multiple characterizations of the entropy function, some more complex than others [3], [29]. We found the approach from [9] the most intuitive and easy to follow, but we realize that there is a lot more depth to the characterization of entropy than is given here.

Firstly, we define a function G associated with one event having probability p. The function $G(p)$ represents the amount of information gained when observing an event with probability p. We will show that if this function satisfies four fundamental properties, it must take the form of a **logarithm**:

The four fundamental properties are:

1. $G(p)$ is a monotonically decreasing function in p: as the probability of an event increases, the information associated with observing it decreases, and vice versa.
2. $G(1) = 0$: Events that always occur with certainty provide no new information.
3. $G(p_1 \cdot p_2) = G(p_1) + G(p_2)$: The information from observing independent events is additive.
4. The function $G(p)$ is assumed to be continuous and differentiable for mathematical convenience.

Proof. **Characterization of Entropy**

Starting from property 3, we have:

$$G(p_1 p_2) = G(p_1) + G(p_2).$$

We differentiate this expression with respect to p_1 and p_2:

$$G'(p_1 p_2) + p_1 p_2 G''(p_1 p_2) = 0.$$

We set $p_1 \cdot p_2 = u$. And rewrite this as a derivative of a product.

$$G'(u) + uG''(u) = 0,$$
$$(uG'(u))' = 0.$$

Integrating once with respect to u, we obtain a constant k:

$$uG'(u) = k.$$

Solving this differential equation gives:

$$G'(u) = \frac{k}{u} \quad \Longrightarrow \quad G(u) = k\ln(u) + c, \quad k, c \in \mathbb{R}.$$

Applying property 2, we have $[G(1) = 0] \implies k\ln(1) + c = c = 0$ And because following property 2 we have that $G(u)$ needs to increase when p decreases and thus for all $0 \leq p \leq 1$, $k\ln(p) \geq 0$. Because we know that $\ln(x) \leq 0$ for $0 \leq x \leq 1$, $k \leq 0$

$$G(u) = k\ln(u).$$

The choice of k corresponds to a specific scaling of information units (bits if $k = -1/(\ln(2))$ is used, nats if base $k = -1$).

Thus, these four fundamental properties characterize a logarithmic form of Shannon's entropy. □

So G, which is a function of only one event probability is a logarithm. It is then not absurd to link this to H which is a sum of these logarithms, because it is associated with a sum of events (not only one event like G).