A STUDY OF ALTERNATIVE AXIOMATIC JUSTIFICATIONS FOR THE ENTROPY FUNCTION AND OF ITS USE IN UNCERTAIN REASONING

A dissertation submitted to the University of Manchester for the degree of Master of Science in the Faculty of Engineering and Physical Sciences

2007

Elizabeth Howarth
School of Mathematics
Contents

Abstract 5

Declaration 6

Copyright Statement 7

Acknowledgements 8

1 Introduction 9

2 The nature of information 12
  2.1 What is information? 12
  2.2 How can we measure information? 14
  2.3 Summary 18

3 Proposed properties of an information measure 20
  3.1 Monotonicity 20
  3.2 Maximal value for the uniform distribution 21
  3.3 Symmetry 21
  3.4 Expansibility 22
  3.5 Branching properties 23
    3.5.1 The branching property 23
    3.5.2 Recursivity 24
    3.5.3 Multiplicative Recursivity 25
    3.5.4 Linear Recursivity 25
3.5.5 Strong Linear Recursivity ........................................ 26
3.6 Additivity properties .............................................. 27
  3.6.1 Additivity ......................................................... 28
  3.6.2 Generalized Additivity .......................................... 29
  3.6.3 Subadditivity ..................................................... 30
3.7 Relationships between Branching and Additivity
  properties ............................................................ 31
  3.7.1 Derivability between the Branching Properties ............... 31
  3.7.2 Derivability between Branching and Additivity Properties ........................................ 32
3.8 Continuity .......................................................... 34
3.9 Being small for small probabilities .................................. 35
3.10 Normalization ........................................................ 35
3.11 Summary ............................................................ 36

4 Some characterizations of measures of information .......... 37
  4.1 Characterizations of $H$ from desiderata ......................... 37
    4.1.1 Shannon, Jaynes and Paris .................................. 37
    4.1.2 Khinchin ......................................................... 38
    4.1.3 Feinstein and Rényi ............................................ 39
    4.1.4 Aczél, Forte & Ng ............................................. 40
  4.2 Other justifications for $H$ ...................................... 41
  4.3 Characterizations of other measures .............................. 42
    4.3.1 Hartley’s information measure ............................... 42
    4.3.2 Rényi’s generalized distributions ........................... 42
    4.3.3 Entropies of order $\alpha$ .................................... 46
    4.3.4 Entropies of degree $\beta$ .................................... 48
  4.4 Other characterization results ................................... 49
    4.4.1 Sum Property .................................................... 49
    4.4.2 Properties of $I_2(1-p,p)$ .................................... 50
The nature of information is examined. The axiom of information theory: that the object of a measure of information should be a probability distribution, is discussed, as is the relationship of information to uncertainty. Several desiderata of an information measure from the literature are presented and discussed, together with justifications and criticisms of these.

Various axiomatizations of Shannon’s information measure

\[ H(p_1, p_2, \ldots, p_n) = - \sum_{i=1}^{n} p_i \log p_i \]

are considered and compared, together with alternative justifications and characterizations of alternative measures. Arguments for extending the domain of an information measure to include partial (non-exhaustive) probability distributions are examined and found to be convincing.

Various justifications of the Maximum Entropy Principle (MEP) are similarly discussed and compared. These include the information theoretic justification which relies on the justification of \( H \) as a measure of information, as well as others which do not.

The Minimum Gain Inference Process, is introduced in order to explore the consequences of allowing assignment of zero probabilities (in contrast to the MEP). Its behaviour is explored through a simple case-study. It is shown to satisfy certain principles of inductive reasoning, and is also shown not to satisfy certain others. It is found unlikely to be of great use due to certain undesirable properties.
Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright Statement

i. Copyright in text of this dissertation rests with the author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author. Details may be obtained from the appropriate Graduate Office. This page must form part of any such copies made. Further copies (by any process) of copies made in accordance with such instructions may not be made without the permission (in writing) of the author.

ii. The ownership of any intellectual property rights which may be described in this dissertation is vested in the University of Manchester, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the University, which will prescribe the terms and conditions of any such agreement.

iii. Further information on the conditions under which disclosures and exploitation may take place is available from the Head of the School of Mathematics.
Acknowledgements

I would like to thank George Wilmers for his thought-provoking supervision, and my family for their ongoing support.
Chapter 1

Introduction

For a finite, discrete, complete probability distribution

\[ P = (p_1, p_2, \ldots, p_n), \quad p_i \geq 0 \quad i = 1 \ldots n, \quad \sum_{i=1}^{n} p_i = 1 \]

Shannon’s [25] information function (also called the ‘entropy’ or ‘uncertainty’ function)

\[ H(P) = -K \sum_{i=1}^{n} p_i \log p_i \]  \hspace{1cm} (1.1)

where \( K \) is some positive constant and \( 0 \log 0 \) is taken to be \( \lim_{x \to 0} x \log x = 0 \),
gives a non-negative real number which may be interpreted as a measure of how much information is contained, on average, in the outcome of an experiment with distribution \( P \).

Various justifications of the entropy expression (1.1) have been given (Aczél et al. [1], Feinstein [4], Jaynes [7], Khinchin [13], Lee [15], Paris [18], Rényi [23], [24], Shannon [25], Tverberg [30]) and it is widely accepted and used, but not universally or uniquely so (Daróczy [2], Hobson [5], Kahre [11], Rényi [23], [24], Uffink [31]). Several such justifications take the form of a list of desiderata for a measure of information, followed by a proof that \( H \) is the only measure to satisfy them. The weak point of such an argument is of course that there may be disagreement about whether the given desiderata specify exactly what is required. The fact that different information measures exist indicates that there is no universal agreement about what constitute the desirable properties of a measure of information. However, the range
of different characterizations which all uniquely entail Shannon’s function, shows that it, uniquely, satisfies several combinations of ‘desirable properties’ identified by those working in the field.

One important use of Shannon’s function (1.1) is as a tool in inductive reasoning, through the Maximum Entropy Principle (MEP) (Jaynes [7], [9], [10], Klir & Weirman [14], Paris [18], [19], Paris & Vencovská [20], [21], [22], Uffink [31]), to assign unknown probabilities consistently with known constraints by choosing, from all possible distributions consistent with the given constraints, the unique distribution $P$ for which $H(P)$ is maximal. The MEP has applications in many diverse fields such as statistical mechanics [7], [8], image reconstruction [33], artificial intelligence [18], [19], [20], [21], and others [27].

This dissertation presents a review of the literature regarding the justification of Shannon’s function (1.1) and its use in the Maximum Entropy Principle. Chapter 2 examines the nature of information through the comments of various authors. It discusses what information is, how it might be measured, its relationship to probability and the justification of the conventional domain for a measure of information:

$$\Gamma = \bigcup_{n \in \mathbb{N}^+} \Gamma_n$$  \hspace{1cm} (1.2)

where $\mathbb{N}^+ = \{n \in \mathbb{N} \mid n \geq 1\}$ and

$$\Gamma_n = \{P = (p_1, p_2, \ldots, p_n) \mid p_i \geq 0 \hspace{0.5cm} i = 1 \ldots n, \hspace{0.5cm} \sum_{i=1}^{n} p_i = 1\}$$  \hspace{1cm} (1.3)

Chapter 3 discusses several properties that have been proposed as ‘desirable’ for an information measure to possess, many of which have been used as axioms, in various combinations, in order to characterize Shannon’s function (1.1). Justifications and criticisms of these are presented and discussed. Chapter 4 compares characterizations of Shannon’s function, including Shannon’s original [25] as well as alternatives by Aczél et al. [1], Feinstein [4], Jaynes [7], Khinchin [13], Paris [18], and Rényi [23], [24], and of other measures (Aczél et al. [1], Daróczy [2], Rényi [23], [24]). Chapter 5 reviews some of the justifications given for the Maximum Entropy Principle, as

\[\text{\footnotesize\textsuperscript{1}Certain assumptions are necessary to ensure that such a unique solution exists, see §5.2}\]
well as some criticisms of these. Chapter 6 presents an alternative inference process and an investigation of its behaviour in a simple case and of certain of its properties. Chapter 7 presents a summary of conclusions drawn throughout the study.

In this study the expression \( H(p_1, p_2, \ldots, p_n), H(P) \) or simply \( H \) refers to Shannon’s information function (1.1), and similarly \( aH(P), H_a(P) \) and \( H^b(P) \) refer to specific measures as defined in the text, whereas the expression \( I(p_1, p_2, \ldots, p_n), I(P) \) or \( I \) refers to an unspecified measure of information. A Roman subscript e.g. \( H_n \) or \( I_n \) indicates a restriction of the domain to \( \Gamma_n \) (1.3). Where not otherwise accredited, results and working are due to the author.

This study will, for the sake of simplicity, only examine the case of finite, discrete probability distributions, although extensions of both Shannon’s function (1.1) and other measures of information exist to deal with continuous distributions (Jaynes [7], [9], Shore and Johnson [27], Uffink [31]).

The optimality or otherwise of Shannon’s entropy function (1.1) as a measure of information, and of the Maximum Entropy Principle as a method of inductive inference, are controversial topics and this study does not adopt the ambitious aim to settle either question. Rather, it aims to present a balanced account of the debate and a reasoned consideration of the arguments in order to reach personal conclusions and to allow readers to reach theirs.
Chapter 2

The nature of information

2.1 What is information?

The word ‘information’ is widely used in everyday language and has a range of meaning depending on the context. Contemporary dictionaries show that the word ‘information’ can refer to objective ‘facts’ or ‘data’, or the more subjective ‘knowledge’, ‘intelligence’ or ‘news’ obtained from these or otherwise. Information may be, or be represented by, signals or characters; it may be conveyed, communicated, provided, obtained, received or learned; it may be ‘about’ a subject, and it can justify change in belief or behaviour.

The Stanford Encyclopedia of Philosophy [28] gives a ‘General Definition of Information’ which is widely used in the philosophy of information. It is expressed in terms of form and meaning as follows.

Definition: $\sigma$ is an instance of information, understood as semantic content, if and only if:

- $\sigma$ consists of one or more data;
- the data in $\sigma$ are well-formed;
- the well-formed data in $\sigma$ are meaningful.

This indicates that information consists of data with two aspects of interest: a form and a meaning. Of these, it may seem that to quantify the amount of meaning
would be the most interesting area for investigation. However, this is surely a highly subjective quantity depending in various ways on the receiver, and as such does not seem to lend itself well to a mathematical study. The form of data, while objective, is also a challenging area for mathematical investigation: how to compare the information contained in a sight with that in a smell for example? It seems that any system of measurement must be based on a more abstract feature of information, separable from both its meaning and its form.

Mathematical information theory began with Shannon [25], a telecommunications engineer, whose work originated in the consideration of the communication of natural language messages. Any message which may be expressed in English, for example, may be broken down into a sequence of letters and punctuation marks, members of a finite, alphanumeric character set which may or may not be further codified into some smaller (e.g. binary) symbol set. In this way any natural language message can be transmitted, one character or symbol at a time, and reconstructed, one character or symbol at a time, by a receiver.

While such messages are usually created for the purpose of conveying information (in the sense of meaning), from the engineer’s point of view, the amount of information ‘contained in’ a message is a function of the statistical properties of the message, a quite different kind of information. Thus, he is not concerned with the meaning of the message or in which language or alphabet it is expressed.

Shannon [25] proposed that the receipt of each successive character of a message may be considered to be a source of information, an information-carrying event, and that a measurement of how much information it conveys can be derived using some underlying probability distribution for which, of the set of possible symbols, will come next in the message. These statistical properties depend on the meaning of the message in the sense that the message was presumably created to convey its meaning, and its meaning dictates the symbols chosen to convey it. However, once the message exists and its statistical properties are known, the message itself, including its semantic content or meaning, and even the form (symbol set) used to convey it, can be disregarded as irrelevant.
To generalize from the receipt of characters in a message to a broader notion of information-carrying events seems intuitively reasonable. We feel from experience that the process of obtaining information, whereby we gain knowledge (or reduce ignorance), is achieved through tangible events, for example seeing a familiar face, hearing a door slam or opening the curtains to discover that it is raining. Feinstein [4] asserts that “We receive information whenever we are informed of an event whose occurrence was not previously certain”. Happily, mathematics has a long-established way to model uncertain events, using probability theory.

2.2 How can we measure information?

Thus Shannon [25] proposed that the object of a measure of information should be a probability distribution, representing some random experiment or other random system of exhaustive and mutually exclusive events. One advantage of this model is that it makes information potentially quantifiable, provided that one has a probability distribution for an event and its possible alternatives. Another is that assuming the information content of an event to be a function only of its underlying distribution means that the quantity may be treated as independent of both the context or meaning of the event and of any property of the receiver. A third is that it renders comparable, in terms of information content, any number of seemingly disparate situations, provided that they can each be modelled by a probability distribution.

Rényi [23] supports Shannon’s proposal, asserting that the information contained in an event is indeed independent of both its context and its form; and that it is an intrinsic and entirely objective property, independent even of whether or not it reaches the perception of an observer. He gives the example: if a random variable $\xi$ takes values $x_1, x_2, \ldots x_n$ with probability distribution $P = (p_1, p_2, \ldots p_n)$ where $p_i = Pr(\xi = x_i)$, then we can say that the information $I(\xi)$ contained in the value of $\xi$ is actually a function $I(P)$ of $P$, and does not depend on $x_1, x_2, \ldots x_n$. Furthermore, any

\footnote{Aczél et al. note that the terminology of ‘experiments’ and ‘outcomes’ refers to one interpretation of partitions of a set. Other interpretations are possible, and information theory may thus be re-interpreted in other terms, as explored by Ingarden and Urbanik [6].}
random variable with the same distribution, regardless of its real-world significance, will have the same measure of information. This makes explicit what Shannon had implied.

There are other information measures which use a different domain, for example the measure of Algorithmic Complexity [32], defined on strings (which may represent messages, as above). This is a measure of the incompressibility of a string, with the interpretation that a message which cannot be compressed has a high information content, while one which is less succinct has a lower information content. However, there is a similarity to Shannon’s measure in that the measure is a function of statistical properties of a ‘message’ (string).

Most authors studied here do not examine the nature of information, or the question of the domain for an information measure, but take as the starting point for their work the question of how to measure the information associated with a given distribution. Even less is said regarding the appropriate range for an information measure. All characterizations considered here produce information measures with range $\mathbb{R}_0^+$, the non-negative real numbers. The property of non-negativity is cited by some authors (Shannon [25], Khinchin [13]), as appropriate for an information measure, (although without explicit justification). However, justification of the non-negative real numbers, or consideration of alternatives, is not discussed.

The purpose of a measure is to allow comparison, suggesting that the range of any useful measure should be a partial ordering. Certain proposed desiderata considered below (Monotonicity (§3.1) and Maximal Value for the Uniform Distribution (§3.2)) express this implicitly by requiring that the information measures associated with certain distributions are comparable in size. Furthermore, it seems appealing to require, if possible, that any two objects to which a measure can be applied should be comparable by it. If accepted, this requirement would mean that the range must be a linear ordering.

While not explicitly assumed in any characterization, it could perhaps be argued that range $\mathbb{R}_0^+$ is implicitly assumed in the statement of the desiderata. All characterizations considered here assume continuity in one form or another, and all implicitly
CHAPTER 2. THE NATURE OF INFORMATION

assume a single-valued output. This combination, together with the intuitive appeal of a linear ordering, strongly suggests the real numbers; and it seems difficult to propose any suitable alternative.

Certain other proposed desiderata (the branching and additivity properties (§3.5, §3.6)) require that the information values associated with different distributions may be added to obtain that associated with a different related distribution. This places further restrictions on the units of the measure, meaning that more is required of the measure than a simple method of comparison.

From this point on, unless otherwise stated, Shannon’s axiom: that the domain of an information measure \( I \) is a set of probability distributions, will be adopted here. For the sake of simplicity only the set \( \Gamma \) (1.2) of finite, discrete distributions will be considered. The question of range is not pursued further, since the explicit assumption of a given range is not necessary to any of the characterizations considered. However, this may be an interesting area for further study.

Given, then, that an information measure applies to a probability distribution, what should it actually express? Shannon [25] described its intended purpose as being to measure how much choice is involved in the selection of an event or how uncertain we are of the outcome. The idea of choice applies, for example, in the case of a natural language message, where each successive character is chosen by the sender but is not known in advance by the receiver. The idea of information being related to uncertainty is more widely applicable, as well as being intuitively appealing, and is adopted and explored by several authors.

Khinchin [13] asserts that every finite, discrete probability distribution describes a state of uncertainty, using the following example. Let \( A = (0.5, 0.5) \), \( B = (0.99, 0.01) \). Then an experiment with distribution \( A \), where each outcome is equally likely, describes a state of greater uncertainty than one with distribution \( B \), where an observer (aware of the underlying distribution) could be much more confident of observing one outcome over the other. Distribution \( C = (0.3, 0.7) \) represents a state of uncertainty somewhere between those of \( A \) and \( B \).

He elaborates, reasoning that performing an experiment, whose possible outcomes
are described by a given distribution, yields information and simultaneously elimin- 
ates the prior uncertainty which existed regarding the outcome. Hence, information 
consists in removing the uncertainty that existed prior to the experiment. He extends 
this reasoning to propose that information therefore increases with prior uncertainty 
and so can be expressed as some increasing function of uncertainty. He suggests that 
information and uncertainty may be taken to be equal for the sake of convenience.

Rényi [23] presents a similar argument: that when one receives some information, 
"...the previously existing uncertainty will be diminished. The meaning of information 
_{is precisely this diminishing of uncertainty."}_ He agrees that the uncertainty regarding 
an event, prior to its occurrence, may be taken to be equal to the information yielded 
by the occurrence of the event, and asserts that uncertainty and information are 
essentially the same thing, relevant respectively before and after the related event.

Feinstein [4] takes a slightly different approach by considering the information 
conveyed in an event _independently of its underlying distribution_, as a decreasing 
function of its probability, saying “the more likely an event is, the less information 
is conveyed by knowledge of its occurrence”. This seems intuitively appealing, and 
consistent with the everyday notion of information.

Feinstein extends this reasoning to assert that the information content of a com-
plete scheme of events represented by a probability distribution \( P = (p_1, p_2, \ldots, p_n) \in \Gamma_n \) should be the “weighted average” of the information contents of each individual 
possible outcome: 
\[
I(P) = p_1 I(p_1) + p_2 I(p_2) + \ldots + p_n I(p_n),
\]
i.e. the average amount of information yielded by a trial of such an experiment, taking into account all possible outcomes. Implicit in his argument is the assumption that the object of a measure 
of information need _not_ necessarily be a complete distribution but may _also_ be an 
individual event, a single component of a distribution.

Rényi [23], [24] extends Feinstein’s consideration of the information content of 
individual events with the notion of a ‘generalized distribution’: a partial probability 
distribution where the events represented are mutually exclusive but not necessarily 
exhaustive, and considers the information content of such. These related approaches 
will be discussed further in chapter 4.
Shannon [26] noted: “The word ‘information’ has been given different meanings by various writers in the general field of information theory. It is likely that at least a number of these will prove sufficiently useful in certain applications to deserve further study and permanent recognition. It is hardly to be expected that a single concept of information would satisfactorily account for the numerous possible applications of this general field.” This may be interpreted as a proviso regarding the applicability of his work and later work based on it; a warning that it has strictly defined applicability and should not be expected to deal with every kind of information.

The Marriot-Webster dictionary [16] includes the following definition of information as:

a quantitative measure of the content of information; specifically: a numerical quantity that measures the uncertainty in the outcome of an experiment to be performed.

This is the definition most widely used in the field of information theory, and that on which most work in the field, and this study, is based.

2.3 Summary

In the majority of the literature on the theory of information, the object of a measure of information is taken to be a probability distribution describing the outcomes of a random experiment. The reasoning behind this assumption is that the process of gaining information can be modelled by learning which of a set of possible outcomes actually occurs.

Some authors describe an information measure in terms of the uncertainty which exists regarding the outcome of a trial before it is carried out and claim that the information gained as the outcome is learned is equivalent to the elimination of the prior uncertainty. Furthermore, the information content of an experiment is thus assumed to be independent of its real-world outcomes, depending only on the probability distribution of these outcomes, and of any feature of any observer of the experiment.
In this study it will be assumed, for the sake of simplicity and unless otherwise stated, that the domain of a measure of information is the set, $\Gamma (1.2)$, of finite, discrete probability distributions.
Chapter 3

Proposed properties of an information measure

Given the assumption, discussed in the previous chapter, that a measure of information $I$ should have domain $\Gamma$ (1.2) (i.e. the set of all finite, discrete, complete probability distributions), various authors have identified properties which they judge to be ‘desirable’ for such a measure to possess. Various combinations from among these have been taken as axioms for characterizations of information measures, some of which will be considered in the next chapter.

This chapter will examine some of these properties individually, presenting selected justifications and criticisms found in the literature, and a discussion of how convincing these arguments are. It will also examine relationships of derivability between these properties, in order to prepare for a comparison, in the following chapter, of various characterizations of information measures, including Shannon’s (1.1) and others.

3.1 Monotonicity

Some authors reason that, for a uniform distribution $P = (p_1, p_2, \ldots, p_n)$ where $p_i = \frac{1}{n}, i = 1 \ldots n$, $I(P)$ should be a monotonic increasing function of $n$ since, as Shannon [25] says: “with equally likely events there is more choice, or uncertainty, when there
are more possible events”. This is intuitively very appealing, matching everyday experience closely, and no criticism of this proposed property has been found in the literature.

### 3.2 Maximal value for the uniform distribution

A related but distinct property postulated for an information measure $I$ is that, for a given $n$, $I(p_1, p_2, \ldots, p_n)$ should take its maximal value at the uniform distribution, i.e. when $p_i = \frac{1}{n}$, $i = 1 \ldots n$. To justify this, Khinchin [13] asserts that “for fixed $n$ it is obvious that the scheme with the most uncertainty is the one with equally likely outcomes”. This seems intuitively reasonable and is consistent with his motivating example considered in section 2.2.

Shannon [25] does not assume but derives this property from $H (1.1)$ and claims that it substantiates $H$ as a reasonable measure of information, since the situation modelled by the uniform distribution is “intuitively the most uncertain situation”. No criticism of this desideratum has been found in the literature.

### 3.3 Symmetry

**Definition:** $I$ is $m$-symmetric if

$$I(p_1, p_2, \ldots, p_m) = I(p_{\pi(1)}, p_{\pi(2)}, \ldots, p_{\pi(m)})$$

for $\pi$ any permutation of $\{1, 2, \ldots, m\}$, $m \in \mathbb{N}$, $m \geq 2$.

**Definition:** $I$ is symmetric if $I$ is $m$-symmetric for every $m \in \mathbb{N}$, $m \geq 2$.

These definitions express the desideratum, proposed by Feinstein [4], Rényi [23], [24], and Aczél et al. [1], that the information content of an experiment should not depend on the ordering or labelling of its possible outcomes.

This is intuitively compelling where $P$ is interpreted as representing a random experiment, since all of its constituent probabilities share the status of being known
possible events. Therefore, the experimental distribution may be equivalently represented by $P = (p_1, p_2, \ldots, p_n)$ or by $P' = (p_{\pi(1)}, p_{\pi(2)}, \ldots, p_{\pi(n)})$ where $\pi$ is any permutation of $1, \ldots, n$. Hence, to do otherwise than assign equal values to $I(P)$ and $I(P')$ would seem inconsistent.

However, Shannon’s measure is used in the Maximum Entropy Inference Process, where the distributions to which it is applied refer to the atoms of a propositional language, which do not all share the above symmetry of status, so the desirability of the Symmetry property in this case is less clear. This is discussed further below (§5.3.2).

### 3.4 Expansibility

**Definition:** $I$ is **expansible** if

$$I(p_1, p_2, \ldots, p_n, 0) = I(p_1, p_2, \ldots, p_n)$$

for $n = 1, 2, \ldots$

Aczél et al. [1] and Khinchin [13] assume this property, asserting (without explicit justification) that adding an impossible event to a scheme does not alter its uncertainty. This seems intuitively reasonable given the interpretation of probability zero as a representation of impossibility, and no criticism of this desideratum has been found.

However, it is interesting to consider an alternative interpretation where probability zero is taken to represent the absolute least likelihood, than which nothing can be less likely, but without the special connotation of certainty given by the term ‘impossible’. With this alternative interpretation, the probability of such an event would still make no contribution to a sum (or to a sum of terms $p_i \log p_i$ by the convention stated in the introduction), but the event itself could not be discounted from consideration as ‘impossible’. This would make the issue of expansibility less clear, since if one accepts monotonicity then one accepts that the number of possible cases affects the information content of an event, so it may be desirable that adding
a zero-probability event does alter (increase) the information of the system. Some of the implications of this alternative interpretation for uncertain reasoning will be further examined in chapter 6.

### 3.5 Branching properties

The following proposed properties all describe how a measure of information behaves when applied to different groupings of probabilities (representing groupings of events) from the same underlying distribution. Each property in this section (after the first) is a special case of the preceding one. Thus, where any is assumed, all preceding it are also implicitly assumed.

#### 3.5.1 The branching property

The following definition of the branching property is taken from Ebanks et al. [3], who consider it to be “one of the most fundamental of all axioms which can be postulated for an information measure”.

**Definition:** $I$ has the branching property if there exists a real-valued function $G$ such that

$$I(p_1, p_2, \ldots, p_n) = I(p_1 + p_2, p_3, \ldots, p_n) + G(p_1, p_2)$$

(3.3)

$\forall P \in \Gamma_n, \ n \geq 3$

Ebanks et al. [3] offer the following interpretation. If an experiment $A$ with possible outcomes $A_1, A_2, \ldots, A_n$ and associated distribution $(p_1, p_2, \ldots, p_n)$ is compared with experiment $A'$ with possible outcomes $A_1 \cup A_2, A_3, \ldots, A_n$ and associated distribution $(p_1 + p_2, p_3, \ldots, p_n)$ (i.e. identical to $A$ except that outcomes $A_1$ and $A_2$ are grouped together as a single outcome), then the amount of information ‘lost’ by observing the outcome of $A'$ depends only on the probabilities $p_1$ and $p_2$. They call such a function $G$ a generating function for $I$.

This is appealing since, for a given experiment, it asserts both that there is greater information to be gained where the outcomes are considered in individual detail than
where summarized or categorized in groups, and that the difference in information depends on the probabilities of the grouped outcomes. Both of these assertions seem reasonable and no criticism of this proposed property has been found in the literature.

### 3.5.2 Recursivity

Ebanks et al. [3] define the recursivity property as follows.

**Definition:** $I$ is recursive of type $M$ for some real-valued function $M$ if

$$I(p_1, p_2, \ldots, p_n) = I(p_1 + p_2, p_3, \ldots, p_n) + M(p_1 + p_2)I\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$

(3.4)

for $n \geq 3$

i.e. if $I$ is branching (3.3) with generating function

$$G(x, y) = M(x + y)I\left(\frac{x}{x + y}, \frac{y}{x + y}\right)$$

for $x, y \in (0, 1), \ x + y \leq 1$.

This strengthens the assumption of the Branching property, i.e. that the amount of information lost when the outcome of $A'$ is $A_1 \cup A_2$ is some (arbitrary) function of $p_1$ and $p_2$, to the assumption that it is the information of the experiment consisting of only possible outcomes $A_1$ and $A_2$, weighted by some function of the probability, $p_1 + p_2$, that this loss occurs.

Ebanks et al. [3] offer an intuitive interpretation of recursivity; it expresses the idea that the amount of information lost when outcomes $A_1$ and $A_2$ are grouped can be ‘recaptured’ by performing another experiment with only possible outcomes $A_1$ and $A_2$. The weighting represents the idea that this further experiment need only be performed $p_1 + p_2$ of the time.

Clearly, since it is a stronger assumption than that of the Branching property, one would in general be less willing to accept it, although it does have intuitive appeal and is accepted by many authors (Feinstein [4], Jaynes [7], Paris [18], Rényi [23], [24], Shannon [25], Tverberg [30]), including all of those who assume Linear Recursivity (§3.5.4) or Strong Linear Recursivity (§3.5.5), which are special cases.
However, the implicit assumption of the interpretation of Ebanks et al. [3]; that one could in fact perform such a ‘recapturing’ experiment with restricted sample space $A_1$ and $A_2$ only seems somewhat questionable. For example, when throwing a die one could group outcomes ‘lands on 3’ and ‘lands on 6’ as ‘lands on a multiple of 3’ but it is not then possible to perform an experiment where the same die may only land on 3 or 6 in order to re-capture the lost information. This is a criticism of the analogy, not necessarily of the property itself, but it makes the argument for its assumption less compelling, and applies also to all recursivity properties discussed below.

### 3.5.3 Multiplicative Recursivity

Ebanks et al. [3] define the property of Multiplicative Recursivity as follows.

**Definition:** $I$ is recursive of multiplicative type $M$ if it is recursive of type $M$ where $M$ is multiplicative, i.e. $M(pq) = M(p) \cdot M(q)$ $\forall p, q \in \mathbb{R}$.

They then prove the following.

**Lemma 1.** (Ebanks et al. [3]). If $I$ is recursive and

$$I(p_1, p_2, p_3, p_4) = I(p_1 + p_2 + p_3, p_4)$$

$$+ M(p_1 + p_2 + p_3) \cdot I\left(\frac{p_1}{p_1 + p_2 + p_3}, \frac{p_2}{p_1 + p_2 + p_3}, \frac{p_3}{p_1 + p_2 + p_3}\right)$$  \hspace{1cm} (3.5)

then $I$ is recursive of multiplicative type.

The property (3.5) could be interpreted similarly to that of recursivity, and justified on similar grounds. Ebanks et al. [3] presumably judge it to be no less reasonable that the assumption of recursivity, since they consider only multiplicative recursivity thereafter.

### 3.5.4 Linear Recursivity

The special case of Multiplicative Recursivity where $I$ is recursive with $M(x) = x$ will be called Linear Recursivity; defined as follows.
CHAPTER 3. PROPERTIES OF AN INFORMATION MEASURE  

Definition: $I$ is *linearly recursive* if

$$I(p_1, p_2, \ldots, p_n) = I(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2)I\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$ (3.6)

∀$P \in \Gamma_n$, $n \geq 3$.

Many authors (Feinstein [4], Jaynes [7], Paris [18], Rényi [23], [24], Shannon [25]) adopt the assumption of Linear Recursivity (or its special case - Strong Linear Recursivity (§3.5.5)), where the identity function is used as the weighting. This is presumably because this is the simplest function $M$ consistent with the definition of recursivity and so captures the desired dependence on $p_1$ and $p_2$ in the simplest possible way. It also has the natural interpretation of expressing how proportionally often the ‘re-capturing’ exercise of described above (§3.5.2) is applicable.

However, Aczél et al. [1] question the justification of the linear weighting in this assumption, observing that different weighting functions produce different measures which may also be considered as information measures, for example with $M(x) = x^\alpha$ ($\alpha \in \mathbb{R}, \alpha > 0, \alpha \neq 1$) one obtains the infinite family of entropies degree $\beta$ (§4.3.4, Daróczy [2]).

3.5.5 Strong Linear Recursivity

Jaynes [7], Paris [18] and Shannon [25] assume the following property of an information measure, which is sometimes referred to as the Composition Law (Jaynes [7]). In this study it will be called ‘Strong Linear Recursivity’ to highlight the relationship with Linear Recursivity.

Definition: An information measure $I$ is *strongly linearly recursive* if

$$I(p_1, p_2, \ldots, p_n) = I(w_1, w_2, \ldots, w_r) + \sum_{i=1}^{r} w_i I\left(\frac{p_{m_i-1+1}}{w_i}, \ldots, \frac{p_{m_i}}{w_i}\right)$$ (3.7)

where $w_1 = p_1 + \ldots + p_{m_1}, \ w_2 = p_{m_1+1} + \ldots + p_{m_2}, \ldots, \ w_r = p_{m_{r-1}+1} + \ldots + p_n, \ 1 \leq m_1 < m_2 < \ldots < m_r = n, \ 1 \leq r \leq n, \ n \geq 3$.

This property extends the idea of Linear Recursivity to apply to groupings of arbitrary size across the original distribution. It includes Linear Recursivity as a special case, where $w_1 = p_1 + p_2$, $r = n - 1$ and $w_i = p_{i+1}$ for $i = 2, \ldots, r$. 
The definition refers to the conditional distributions associated with each group: 

\[ P_i = \left( \frac{p_{mi-1}+1}{w_i}, \ldots, \frac{p_{mi}}{w_i} \right), \]

i.e. given that an outcome belongs to group \( i \), the probability that it is event 1, 2, \ldots, \( m_i \) within that group. The uncertainty of the full distribution is thus required to be equal to the uncertainty of the distribution of composite events \( W = (w_1, w_2, \ldots, w_r) \) added to the uncertainties of the conditional distributions, each weighted linearly according to how proportionally often it is applicable.

An intuitive interpretation given by Jaynes [7] is that, instead of considering every individual event of a sample space, one might group the first \( m_1 \) events as a single composite event with probability \( w_1 = p_1 + \ldots + p_{m_1} \), the next \( m_2 \) as a single composite event with probability \( w_2 \), and so on. Then, learning of a particular, individual outcome is equivalent to learning first which composite event the outcome belongs to, and then learning which individual outcome within that group has occurred. (Note that one must also assume symmetry in order for this property to apply to groups of non-consecutively labelled probabilities). Jaynes [7] argues that this property is a consistency requirement of an information measure, since without it, different values of uncertainty may be given to a single experiment through different groupings of its events.

### 3.6 Additivity properties

The properties considered in this section all describe how an information measure behaves when applied to a combined distribution, (i.e. representing a combination of two or more separate experiments). Firstly, some notation (taken from Rényi [23], [24]) will be introduced.

For distributions \( P = (p_1, p_2, \ldots, p_m) \), \( Q = (q_1, q_2, \ldots, q_n) \), for \( m, n \in \mathbb{N}^+ \), define \( P \ast Q \) to be the direct product of \( P \) and \( Q \), i.e.

\[
P \ast Q = (p_1 q_1, \ldots, p_1 q_n, p_2 q_1, \ldots, p_2 q_n, \ldots, p_m q_1, \ldots, p_m q_n)
\]

(3.8)

Where \( P \) and \( Q \) represent respectively experiment \( A \) with possible outcomes...
$A_1, \ldots, A_m$ and $B$ with possible outcomes $B_1, \ldots, B_n$, and where $A$ and $B$ are statistically independent, i.e. $Pr(A_i \cap B_j) = Pr(A_i)Pr(B_j) = p_iq_j$, then $P \ast Q$ may be interpreted as representing their combined experiment $AB$ with outcomes $A_i \cap B_j$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$.

In the case where $A$ and $B$ are not independent, if a trial of $A$ is followed by a trial of $B$, the outcomes for $B$ may have a different probability distribution in light of the outcome of $A$. For a given outcome $A_i$ of $A$, the conditional distribution $Q_i$ of $B$ gives the revised probabilities of outcomes for $B$, denoted

$$Q_i = (q_{1|i}; q_{2|i}; \ldots, q_{n|i}) = (Pr(B_1|A_i), Pr(B_2|A_i), \ldots, Pr(B_n|A_i)) \quad (3.9)$$

where $Pr(B_j|A_i)$ is the conditional probability that experiment B yields outcome $B_j$ given that experiment A yields outcome $A_i$.

In this case, the combined distribution\(^1\) of $A$ and $B$ is

$$(P, Q) = (p_1q_{1|1}; \ldots, p_1q_{n|1}, p_2q_{1|2}; \ldots, p_2q_{n|2}, \ldots, p_mq_{1|m}; \ldots, p_mq_{n|m}) \quad (3.10)$$

The quantity

$$I_P(Q) = \sum_{i=1}^{m} p_iI(Q_i) \quad (3.11)$$

may thus be interpreted as the average information obtained from observing experiment $B$, given the outcome of experiment $A$.

### 3.6.1 Additivity

The property of additivity (also sometimes referred to as the ‘decomposition rule’ [11]) is defined as follows.

**Definition:** $I$ is additive if

$$I(P \ast Q) = I(P) + I(Q) \quad (3.12)$$

\(^1\)This representation may be interpreted as the probabilities for a trial of $A$ followed by a trial of $B$, but since

$$p_iq_{ji} = Pr(A_i)Pr(B_j|A_i) = Pr(A_i \cap B_j) = Pr(B_j)Pr(A_i|B_j) = q_jp_{i|j}$$

it follows that $(P, Q)$ is a permutation of $(Q, P)$, the distribution for a trial of $B$ followed by one of $A$. Only the given representation will be used, for the sake of simplicity.
i.e. if
\[
I(p_1 q_1, \ldots, p_1 q_n, \ldots, p_m q_1, \ldots, p_m q_n) = I(p_1, p_2, \ldots, p_m) + I(q_1, q_2, \ldots, q_n)
\] (3.13)
for all \( P \in \Gamma_m, \ Q \in \Gamma_n, \ m, n = 1, 2, \ldots \)

Rényi [23] claims that this is “one of the most important properties of an entropy function” and Aczél et al. [1] and Khinchin [13] both assert that it is “natural” to expect this property of an information measure, although neither gives explicit justification. It seems intuitively appealing since it is difficult to imagine how a combined experiment could yield more information (in the sense under consideration) than the individual experiments which it combines, or why it should yield less given that the experiments are uncorrelated. No criticism of this proposed property has been found in the literature, although there exist information measures which do not obey additivity, such as the entropies of degree \( \beta \) discussed below (§4.3.4).

### 3.6.2 Generalized Additivity

The property of Generalized Additivity is proposed and defined by Khinchin [13] as follows.

**Definition:** \( I \) is generally additive if
\[
I(P, Q) = I(P) + I_P(Q)
\] (3.14)
i.e. if
\[
I(p_1 q_{1|1}, \ldots, p_1 q_{n|1}, \ldots, p_m q_{1|m}, \ldots, p_m q_{n|m}) = I(p_1, \ldots, p_m) + \sum_{i=1}^{m} p_i I(q_{1|i}, \ldots, q_{n|i})
\] (3.15)
for all \( P \in \Gamma_m, \ Q, Q_i \in \Gamma_n, \ i = 1 \ldots m, \ m, n = 1, 2, \ldots \)

In terms of the statistical interpretation given above, this expresses the idea that the information of the combined experiment is equal to the information of \( A \) added to the sum, for each possible outcome of \( A \), of the information of the corresponding conditional distribution of \( B \), each weighted by \( p_i \), to account for how proportionally often it is applicable.
Khinchin [13] reasons that the uncertainty regarding the outcome of the combined experiment is equal to that where experiment $A$ is followed by experiment $B$ (which now has conditional distribution associated with knowledge of the outcome of $A$). Since the outcomes are not known in advance, all possible combinations of outcomes must be considered, each weighted according to its likelihood, so the above expression represents the average expected information from the joint experiment.

This reasoning is plausible, but the linear weighting here may be questioned in the same way as that in the Linear Recursivity property (§3.5.4).

### 3.6.3 Subadditivity

The property of subadditivity is defined by Aczél et al. [1] as follows.

**Definition:** $I$ is **subadditive** if

$$I(P, Q) \leq I(P) + I(Q)$$

(3.16)

i.e. if

$$I(p_1q_1|1, \ldots, p_1q_n|1, \ldots, p_mq_1|m, \ldots, p_mq_n|m) \leq I(p_1, \ldots, p_m) + I(q_1, \ldots, q_n)$$

(3.17)

for all $P \in \Gamma_m, \ Q \in \Gamma_n, \ m, n = 1, 2, \ldots$ for all $P, Q \in \Gamma$.

This may be interpreted as the requirement that the total information contained in the joint distribution of two experiments cannot be greater than the sum of the information contained in each individually. Shannon [25] and Rényi [24] deduce Subadditivity from $H$ and cite it as a desirable property of an information measure. It is also used as an axiom by Aczél et al. [1].

It is closely related to Additivity, and may be justified by an extension of the same argument: If, when two variables are independent their joint information is equal to the sum of their individual informations, then surely from this value it can only decrease with increasing correlation. No criticism of this desideratum has been found in the literature, although the entropies of order $\alpha$ (§4.3.3) do not have this property.
3.7 Relationships between Branching and Additivity properties

This section presents some derivability relationships between the various branching and additivity properties discussed above. These relationships may be useful in comparing the relative merits of various characterizations, all of which assume at least one of these properties, in the next chapter.

3.7.1 Derivability between the Branching Properties

It is clear from the definitions given above in section 3.5 that each branching property is a special case of those which precede it (in the given order). Thus, where any is assumed, all more general cases are also implicitly assumed. Rényi [24] provides a proof that Linear Recursivity together with Symmetry imply Strong Linear Recursivity. Otherwise it has not been discovered in the literature under what conditions the derivations might be reversed. Thus the relationships known (by this author) are shown in the following diagram.

```
                Strong Linear Recursivity
                ↓     ↑ + symmetry (Rényi [24])
                Linear Recursivity
                ↓
        Multiplicative Recursivity
                ↓
        Recursivity
                ↓
        Branching
```
3.7.2 Derivability between Branching and Additivity

Properties

Strong Linear Recursivity and Generalized Additivity

Paris [18] uses the following notation to express the property of Strong Linear Recursivity

$$I(w_1y_1, w_1y_2, \ldots, w_1y_{m_1}, \ldots, w_ry_1, w_ry_2, \ldots, w_ry_{m_r}) = I(w_1, w_2, \ldots, w_r) + \sum_{i=1}^{r} w_iI(y_1, \ldots, y_{m_i})$$ (3.18)

(let $y_{ij} = \frac{p_m - 1 + j}{w_i}$ to obtain the notation used above in section 3.5.5). From this expression it is easily shown that Generalized Additivity is a special case of Strong Linear Recursivity (where $m_1 = m_2 = \ldots = m_r = n$).

Lemma 2. If an information measure $I$ has the property of Strong Linear Recursivity then it also has that of Generalized Additivity.

Proof. Let $r = m$ and let $w_i = p_i$ and $m_i = n$ for $i = 1, \ldots, r$. Let $y_{jk} = q_{kij}$. Then the Strong Linear Recursivity of $I$ (3.18) may be expressed as follows:

$$I(p_1q_{1|1}, p_1q_{2|1}, \ldots, p_mq_{1|m}, p_mq_{2|m}, \ldots, p_mq_{n|m}) = I(p_1, p_2, \ldots, p_m) + \sum_{i=1}^{m} p_iI(q_{1|i}, \ldots, q_{n|i})$$

Therefore $I$ has Generalized Additivity by (3.15). Therefore, where Strong Linear Recursivity is assumed, Generalized Additivity is implied.

Generalized Additivity and Additivity

Lemma 3. (Khinchin [13]). If an information measure $I$ has the property of Generalized Additivity then it also has that of Additivity.

Proof. It is clear from the above definitions that Additivity (3.12) is a special case of Generalized Additivity (3.14), since if $I$ is assumed to be generally additive, then by
(3.15) we have

\[ I(p_1q_1, \ldots, p_1q_n, \ldots, p_mq_1, \ldots, p_mq_n) = I(p_1, \ldots, p_m) + \sum_{i=1}^{m} p_i I(q_i) \]

Furthermore, where \( P \) and \( Q \) are statistically independent distributions, we have \( Q_i = Q \) for \( i = 1, \ldots, n \), so this reduces to

\[ I(p_1q_1, \ldots, p_1q_n, \ldots, p_mq_1, \ldots, p_mq_n) = I(p_1, \ldots, p_m) + \sum_{i=1}^{m} p_i I(Q) \]

and further to

\[ I(P \ast Q) = I(P) + I(Q) \]

since \( \sum_{i=1}^{m} p_i = 1 \).

This argument is used by Khinchin [13] (using different notation).

**Strong Linear Recursivity and Additivity**

**Lemma 4.** (Rényi [23], [24]). If an information measure \( I \) has the property of Strong Linear Recursivity then it also has that of Additivity, although the converse is not true.

**Proof.** The implication is given by combining the above two results (although Rényi [23], [24] shows it by applying (3.7) directly to \( P \ast Q \)).

To show that the converse is not true, Rényi gives the counter-example

\[ F(p_1, \ldots, p_n) = -\log_2(p_1^2 + \ldots + p_n^2) \]

\( F \) is additive, since

\[
F(P \ast Q) = F(p_1q_1, \ldots, p_1q_n, \ldots, p_mq_1, \ldots, p_mq_n)
= -\log_2((p_1q_1)^2 + \ldots + (p_1q_n)^2 + \ldots + (p_mq_1)^2 + \ldots + (p_mq_n)^2)
= -\log_2(p_1^2q_1^2 + \ldots + q_n^2 + \ldots + p_m^2q_1^2 + \ldots + q_n^2)
= -\log_2((p_1^2 + \ldots + p_m^2)(q_1^2 + \ldots + q_n^2))
= -\log_2(p_1^2 + \ldots + p_m^2) - \log_2(q_1^2 + \ldots + q_n^2)
= F(P) + F(Q)
\]
To show that $F$ is not strongly linearly recursive, suppose the contrary. Then $F$ is linearly recursive, so by (3.6)

$$F\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right) = F\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} F\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$\Leftrightarrow -\log_2\left(\frac{3}{8}\right) = -\frac{3}{2} \log_2\left(\frac{1}{2}\right)$$

$$\Leftrightarrow \frac{3}{8} = \left(\frac{1}{2}\right)^3$$

$$\Leftrightarrow \frac{8}{3} = 2\sqrt{2}$$

This is a contradiction, therefore $F$ is not strongly linearly recursive. Thus Strong Linear Recursivity implies Additivity, although the converse is not true.

The following diagram summarizes these relationships between the additivity properties and recursivity properties.

$$\text{Strong Linear Recursivity} \uparrow \downarrow \downarrow$$

$\text{Linear Recursivity}$ $\downarrow$

$\text{Generalized Additivity}$ $\downarrow (\text{Khinchin [13]})$

$\text{Additivity}$

### 3.8 Continuity

Most authors considered here (Feinstein [4], Jaynes [7], Khinchin [13], Rényi [23], [24], Paris [18], Shannon [25]) assume either that a measure of information should be continuous in all of its arguments, i.e. that $I_n(p_1, p_2, \ldots, p_n)$ is continuous in each $p_i$, $i = 1 \ldots n$, or that $I_2(p, 1 - p)$ is a continuous function of $p$. Only Paris justifies this assumption explicitly with the argument that similar distributions contain similar amounts of information, or that small changes in argument values should not produce great changes in information content. While intuitively reasonable, this argument does not seem absolutely compelling.

Kahre [11] remarks that the assumption of continuity is the ‘weak link’ in Shannon’s [25] characterization of information, and questions what makes a continuous
measure ‘more informative’ than a discontinuous one. While the assumption of continuity may be questionable, this does not seem to be the right question for the purpose, since it is analogous to asking what makes a continuous measure of length more lengthy than a discontinuous one! He also notes the existence of discontinuous information measures, for example Hartley’s measure (§4.3.1), which seems to be a more convincing argument against the desirability of continuity.

3.9 Being small for small probabilities

Aczél et al. [1] propose the property of ‘being small for small probabilities’ (in the case $n = 2$), i.e.

$$\lim_{p \to 0^+} I(1 - p, p) = 0 \quad (3.19)$$

as a weak version of a continuity assumption. Their reasoning: that “we get very little information out of an experiment with two possible outcomes, one of which is almost certain, the other almost impossible”, is the same as Khinchin’s in his motivating example (§2.2). Interestingly, this property does not hold for an infinite category of discontinuous information measures which they characterize with a theorem discussed below (§4.4.3), although no disagreement with this proposed property has been found from any other source.

3.10 Normalization

Aczél et al. [1] and Rényi [23], [24] include a specification of units for an information measure through the normalization axiom: $I(\frac{1}{2}, \frac{1}{2}) = 1$. This fixes the units associated with a measure of uncertainty and for Shannon’s expression (1.1) it specifies that $K = 1$ and the use of logarithm base 2. The choice of logarithm base 2 corresponds to a binary symbol set (e.g. 0 and 1) to convey the information being measured and thus yields a measurement in the commonly used unit of bits (binary digits). Any choice of units is arbitrary but this choice, in the characterization of Shannon’s measure, has the advantage of convenience by simplifying the final expression.
It could be argued that a normalization requirement does not belong in a characterization of a measure, since the essential properties of a measure are separate from the units used to express it (e.g., the nature of weight is the same whether it is measured in grams, pounds or tons). If normalization is omitted from those characterizations of $H$ which use it, the resulting expression would be Shannon’s original (1.1) with variable logarithm base and term $K$, so normalization may be considered ‘optional’.

### 3.11 Summary

This chapter has presented various properties which have been identified in the literature as ‘desirable’ for an information measure to possess. Some seem to have universal agreement while others are more controversial, and various justifications and criticisms have been discussed. The assumption of these properties, taken in various combinations, has implications for the form of $I$ which will be explored through various characterization theorems in the following chapter.
Chapter 4

Some characterizations of measures of information

This chapter presents and compares various characterization results derived from the assumption of various combinations of properties considered in the previous chapter. Firstly, several different characterizations of Shannon’s function are given and compared, together with some of its alternative justifications. Characterizations of some other measures of information are then given for the purpose of comparison. Finally, some more general characterization results are given which demonstrate how various systems of assumed properties for an information measure place restrictions on its form. Justifications for the individual assumed properties presented in the previous chapter are not repeated here. Details of the proofs, which are mostly quite lengthy, are not given; the curious reader is referred to the original sources.

4.1 Characterizations of $H$ from desiderata

4.1.1 Shannon, Jaynes and Paris

Shannon [25] made the original attempt to characterize $H$ (1.1) as a measure of information, upon which all subsequent attempts presumably aim to clarify or improve. Jaynes [7] and Paris [18] present essentially the same characterization (i.e. from
the same desiderata), although their work focuses on the application of the function to inductive reasoning through the Maximum Entropy Principle, rather than on its justification.

Shannon [25] adopts the assumption discussed above (§2.2) (as do all other authors discussed in this study unless otherwise stated), that the domain of a measure of information should be \( \Gamma (1.2) \), and from this considers how to measure the information contained in a given finite, discrete probability distribution

\[ P = (p_1, p_2, \ldots, p_n) \]

(i.e. how to measure the information gained, on average, by learning the outcome of a random experiment with underlying distribution \( P \)).

He proposes the three properties:

**S1** Continuity (§3.8)

**S2** Monotonicity (§3.1)

**S3** Strong Linear Recursivity (§3.5.5)

as requirements of such a measure, \( I(P) \), and then proves that any function \( I \) satisfying these properties must be of the form

\[ I(P) \equiv H(P) = -K \sum_{k=1}^{n} p_k \log p_k \]

for some positive constant \( K \).

Of these requirements, S1 and S3 have both been subject to criticism as discussed above in the relevant sections; S1 for being simply unjustified, or unconvincingly justified, and S3 for its implicit assumption of linear weighting when different groupings of probabilities are compared.

### 4.1.2 Khinchin

Khinchin [13] presents a different characterization of \( H \) using the following properties as desiderata for an uncertainty measure \( I(P) \).
K1 Maximal value for the uniform distribution (§3.2)

K2 Generalized additivity (§3.6.2)

K3 Expansibility (§3.4)

K4 Continuity (not stated explicitly but necessary to his uniqueness proof) (§3.8)

and proves that Shannon’s measure $H$ is the unique measure of information satisfying these.

Of Khinchin’s axioms, K2 is implied by, and so may be considered preferable to, Strong Linear Recursivity (see §3.7.2). However, it may also be subject to the same criticism of an implicit assumption of linear weighting. The above criticisms of the assumption of continuity also apply here.

Khinchin assumes four properties compared to Shannon’s three. A greater number of axioms might be considered preferable where these express more ‘reasonable’ or ‘elementary’ assumptions than a few, stronger, less ‘convincing’ axioms. This is a subjective judgement, and the reader is left to form his own opinion in this and other comparisons.

4.1.3 Feinstein and Rényi

Feinstein [4] gives another characterization of $H$ from the properties of

F1 Symmetry (§3.3)

F2 Linear recursivity (§3.5.4)

F3 Continuity of $I_2(p, 1 - p)$ for $p \in [0, 1]$ (§3.8)

His proof includes the interesting intermediate results that the properties of Expansibility and Strong Linear Recursivity all follow directly from Symmetry and Linear Recursivity (without use of the obtained expression $H$).

The assumption of Linear Recursivity seems preferable to that of its special case, Strong Linear Recursivity (§3.7.1), since it allows at least as many possible solutions,
all other things being equal, but the same criticism of an implicit assumption of linearity still applies. Similarly, the assumption of the continuity of $I_2$ only (as opposed to that of $I_n$ for every $n \in \mathbb{N}$) is a technical improvement, but is still questionable in the same way.

Rényi [23], [24] gives a very similar characterization of Shannon’s measure, which he says is adapted from Fadeev. This additionally assumes

**R4 Normalization** (§3.10)

As discussed, the normalization assumption is essentially irrelevant, serving only to fix the units of the measure. Otherwise the same comments apply as to Feinstein above.

### 4.1.4 Aczél, Forte & Ng

Aczél, Forte and Ng [1] question the assumption of (Strong) Linear Recursivity precisely because it is “rather explicitly linear”. Rather then relaxing the requirement of linearity, they reject any assumption of recursivity (§3.5.2) altogether and offer the following axioms as “a set of natural conditions for a measure of expected information which do not presuppose linearity of its expression”:

**A1** Subadditivity (§3.6.3)

**A2** Additivity (§3.6.1)

**A3** Expansibility (§3.4)

**A4** Symmetry (§3.3)

**A5** Normalization (§3.10)

**A6** Being Small for Small Probabilities (§3.9)

They then prove that Shannon’s function (with $K = 1$ and logarithm base 2 due to the normalization axiom) is the only one satisfying these.\(^1\)

\(^1\)This is actually a corollary of another characterization theorem of theirs, discussed below (§4.4.3)
This characterization seems superior to those considered previously in two ways. Firstly, as Aczél et al. [1] note, A1 and A2 do not entail the assumption of linearity implicit in the recursivity and Generalized Additivity assumptions of other characterizations. Also, it is the only characterization so far considered which does not assume Continuity, one of the more questionable assumptions of other approaches.\(^2\)

Instead, Aczél et al. use assumption A6 to ensure that \(h(x) = I(1 − x, x)\) tends to 0 at both ends of its domain \([0, 1]\). A6 is, as they acknowledge, “a weak form of continuity” but it somehow seems intuitively more reasonable a requirement than explicit continuity, due to Khinchin’s compelling argument (§2.2) that only a small amount of information is yielded, on average, from an experiment, one of whose two outcomes is almost certain and the other almost impossible. However, this is clearly a subjective evaluation; what is intuitively compelling to one person may not be so to another and some may feel more convinced by Paris’ argument for explicit continuity (§3.8), or unconvinced by either.

No criticism has been found or can be provided of the assumptions of A3 and A4, A5 is essentially irrelevant as discussed above, and no criticism of this characterization has been found in the literature. Therefore, for the combination of reasons given, this characterization of Shannon’s function \(H\) by Aczél et al. seems preferable to all others considered here.

### 4.2 Other justifications for \(H\)

Rényi [24] offers a statistical interpretation of \(H(P)\) (1.1) and related theorem in support of its suitability as a measure of information, although this is beyond the scope of this study.

Tikochinsky et al. [29] claim that acceptance of the Maximum Entropy Principle (MEP) as the optimal method of inductive reasoning from a characterization of the

---

\(^2\)Some other authors have also characterized \(H\) without assuming continuity, using instead, the technically weaker assumptions of being Lebesgue integrable (Tveberg [30]) and Lebesgue measurable (Lee [15]) of \(I(1 − p, p)\) for \(p ∈ [0, 1]\) and \((0, 1)\) respectively. These analytic properties are, however, still questionable in the same way as continuity.
42

**technique**, which does not depend on properties of the entropy function, provides a justification of $H$ (1.1) as a measure of information. For example, Jaynes [8] uses a statistical argument in order to justify the MEP, which might in turn be considered to justify $H$. This reasoning will be considered in detail below (§5.3.3, §5.5).

4.3 Characterizations of other measures

4.3.1 Hartley’s information measure

The following is known (Aczél et al. [1], Rényi [24]) as Hartley’s measure of information.

$$\theta H(P) = \log N \quad (4.1)$$

where $N \geq 1$ is the number of non-zero terms in $P$. This measure predates that of Shannon [25], who states that his work is a continuation of Hartley’s.

Aczél et al. [1] give the following characterization of Hartley’s expression (with logarithm base 2 due to Normalization), showing that it may be characterized almost identically to Shannon’s measure (§4.1.4) using properties A1-A5 but omitting the property A6 of Being Small for Small Probabilities and using instead the property of Insensitivity.$^3$

**Definition:** $I$ is *insensitive* if

$$I_2(1 - p_1, p_1) = I_2(1 - p_2, p_2) \quad (4.2)$$

for at least one pair $p_1, p_2$ with $0 < p_1 < p_2 \leq \frac{1}{2}$.

This expresses an obvious property of Hartley’s measure; that it depends only on the number of non-zero terms in its argument distribution, regardless of the values of these terms.

4.3.2 Rényi’s generalized distributions

Rényi [23], [24] explores a generalization of Feinstein’s [4], (§2.2) “weighted average” interpretation of Shannon’s expression (1.1) by extending its domain to the set of

---

$^3$This is actually a corollary of another characterization theorem of theirs, discussed below (§4.4.3)
generalized probability distributions:

\[ \Delta = \{ P = (p_1, p_2, \ldots, p_n) \mid n \in \mathbb{N}^+, \quad p_i \geq 0 \quad i = 1 \ldots n, \quad W(P) = \sum_{i=1}^{n} p_i \leq 1 \} \quad (4.3) \]

where \( W(P) \) is termed the weight of the distribution.

Rényi [23], [24] offers the interpretation that an incomplete distribution \( (P \in \Delta \) where \( W(P) < 1 \) ) describes the result of an experiment whose outcome is “not always observable” (only with probability \( W(P) \)), although he does not give any example of such. In any case, this explanation of why certain probabilities are ‘missing’ from an incomplete distribution seems unnecessary; it seems quite conceivable that one might wish to consider only part of a system of events, for example in situations where only a subset of possible outcomes is known or considered ‘interesting’.

He asserts that the characterization of information measures becomes “much simpler” in the case of generalized distributions, which may be debatable. However, the use of generalized distributions does allow a different approach, hinted at but not explicitly explored by Feinstein [4]; through the expression of a ‘mean-value’ property of an information measure.

Rényi [23], [24] introduces the following notation. Let \( \{ p \} \) denote the generalized probability distribution consisting of a single probability \( p \). For \( P = (p_1, p_2, \ldots, p_m) \), \( Q = (q_1, q_2, \ldots, q_n) \in \Delta \), define \( P \ast Q \) exactly as above (3.8) . Also, where \( W(P) + W(Q) \leq 1 \), define

\[ P \cup Q = (p_1, p_2, \ldots, p_m, q_1, q_2, \ldots, q_n) \]

For \( W(P) + W(Q) > 1 \), \( P \cup Q \) is undefined. He then proposes the following desiderata of an information measure \( I_g \) with domain \( \Delta \).

**RG1** Symmetry: \( I_g(P) \) is a symmetric function of the elements of \( P \)

**RG2** Continuity: \( I_g(\{ p \}) \) is a continuous function of \( p \)

**RG3** Normalization: \( I_g(\{ \frac{1}{2} \}) = 1 \)

**RG4** Additivity: \( I_g(P \ast Q) = I_g(P) + I_g(Q) \)
RG5 Mean-value property:

\[ I_g(P \cup Q) = \frac{W(P) \cdot I_g(P) + W(Q) \cdot I_g(Q)}{W(P) + W(Q)} \]  (4.4)

From these properties he proves the unique solution

\[ I_g \equiv H_g(P) = \sum_{i=1}^{n} p_i \log_2 \left( \frac{1}{p_i} \right) \]  (4.5)

which he calls the ‘entropy of order 1 of the generalized distribution \( P \). \( H_g(P) \) is clearly equal to \( H(P) \) (1.1) (normalized) when \( W(P) = 1 \).

Properties RG1-4 seem directly analogous to their counterparts for complete distributions already discussed, and no further comment seems necessary. The mean-value property RG5 cannot be directly interpreted in the context of complete distributions, since \( P \cup Q \) is undefined for \( P, Q \in \Gamma \). However, consideration of \( P \) and \( Q \) in RG5 as a partition of a complete distribution yields the interpretation that the information contained in a complete distribution should equal the linearly-weighted sum of the information contained in each part of the distribution, no matter how it is divided.

For example, suppose \( P = (p_1, p_2, \ldots, p_n) \in \Gamma \) and let \( P_1 = (p_1, p_2, \ldots, p_m), P_2 = (p_{m+1}, p_{m+2}, \ldots, p_n) \in \Delta \), for \( 1 \leq m \leq n-1 \), form a partition of \( P \). Then \( P = P_1 \cup P_2 \) (and \( W(P_1) + W(P_2) = W(P) = 1 \)).\(^4\) Then RG5 demands that

\[ I_g(P) = \frac{W(P_1)}{W(P)} I_g(P_1) + \frac{W(P_2)}{W(P)} I_g(P_2) \]

Thus RG5 specifies, among other things, how the information measures of different groupings of values from a single complete distribution should behave. In this sense it seems intuitively similar to the property of Strong Linear Recursivity (§3.5.5) which, for the same example, would require that:

\[ I(P) = I(W(P_1), W(P_2)) + W(P_1)I(P'_1) + W(P_2)I(P'_2) \]

where \( P'_1 = (\frac{p_1}{W(P_1)}, \frac{p_2}{W(P_1)}, \ldots, \frac{p_m}{W(P_1)}) \in \Gamma \), and similarly for \( P'_2 \).

RG5 does not, however, entail recursivity of any kind, nor even the branching property, as the following lemma shows.

\(^4\)Note that the definition of \( P \cup Q \) allows this example to be extended to any partition of \( P \) into any number (\( \leq n \)) of partial distributions, including \( P = \{p_1\} \cup \{p_2\} \cup \ldots \cup \{p_n\} \)
Lemma 5. \( H_g \) does not have the branching property.

Proof. For \((p_1, \ldots, p_n) \in \Delta\)

\[
H_g(p_1, \ldots, p_n) - H_g(p_1 + p_2, p_3, \ldots, p_n) \\
= p_1 \log \frac{1}{p_1} + \ldots + p_n \log \frac{1}{p_n} - \frac{(p_1 + p_2) \log \frac{1}{p_1 + p_2} + p_3 \log \frac{1}{p_3} + \ldots + p_n \log \frac{1}{p_n}}{p_1 + \ldots + p_n}
\]

This cannot be expressed as a function of \(p_1\) and \(p_2\) only, so \(H_g\) does not have the Branching property (3.3).

Therefore, \(H_g\) does not have any of the recursivity properties discussed above (§3.5), since each is a special case of the Branching property (§3.7.1). However, \(H_g\) has property RG5 by Rényi’s characterization [23], [24], so \(H_g\) provides a counter-example to show that RG5 does not entail recursivity.

Rényi [23] cites as an advantage of the use of generalized distributions the fact that it allows the expression of the ‘mean-value property of Shannon’s measure’, i.e. \(H(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} W(\{p_i\}) I_g(\{p_i\})\). Thus, Rényi’s motivation for introducing generalized probability distributions seems to be a desire to characterize information measures in quite a different way from those considered above. Through the use of generalized distributions and the Mean-value property (RG5) he allows consideration of the information content of an individual event, and regards the information content of a scheme (exhaustive or non-exhaustive) of events as a mean value of the information contained in each constituent event, weighted according to how close the scheme is to being exhaustive. This agrees with and extends Feinstein’s view [4], (§2.2).

A further obvious advantage of the domain of incomplete distributions for a measure of information is that it is more widely applicable than the usual domain \(\Gamma\) (1.2). It allows the consideration of the information content of events out of context of their underlying distributions, for example when the full distribution is not known.

A third advantage is that it eliminates the need for the questionable assumption in the Ebanks et al. [3] interpretation of the recursivity properties (§3.5.2 - §3.5.5):
that subspaces of an experimental sample space may be taken as the full sample space for a kind of ‘sub-experiment’.

A potential disadvantage of Rényi’s generalized measure is that, due to dependence on the weight of a partial distribution, $H_g$ does not have the recursivity properties of $H$, although the mean-value property (RG5) dictates its behaviour in similar ‘branching’ situations, and might be considered an equally or more desirable alternative on further investigation. A comparison of RG5 with the various branching properties discussed above (§3.5) may be an interesting area for further study.

No criticism of this generalization of the domain of Shannon’s measure has been found in the literature, and none can be suggested by the author. It seems that one can consider an actual or potential event to be informative to a greater or lesser extent without knowing every possible alternative to it. For example, learning that it is snowing in Manchester in April would seem relatively informative compared with learning that it is overcast, which judgement may be made without consideration of all other possible weather conditions for the given day. Hence the application of a measure of information to an individual event seems intuitively at least as natural as its application to a complete distribution; perhaps more so.

The choice of the arithmetic mean in RG5 is questionable, however. Rényi [23], [24] acknowledges this and explores the consequences of allowing a more general version of RG5, as discussed in the following section.

4.3.3 Entropies of order $\alpha$

Rényi [23], [24] also characterizes a family of information measures which he calls the “entropies of order $\alpha$” (also known as the Rényi entropies [1], [31]). These are defined for generalized distributions, although the restricted form, applicable only to complete distributions, is given here for ease of comparison with other measures.

$$H_\alpha(P) = \frac{1}{1 - \alpha} \log_2 \left( \sum_{i=1}^{n} p_i^\alpha \right)$$

(4.6)

for $P = (p_1, p_2, \ldots, p_n) \in \Gamma$, $\alpha \in \mathbb{R}$, $\alpha \geq 0$, $\alpha \neq 1$. 
His characterization stems from the substitution for the arithmetic mean in (4.4) with some other mean-value. This results in the replacement of RG5 by

**RG5’** Mean-value property: There exists a strictly monotonic and continuous function \( f(x) \) such that if \( P, Q \in \Delta \) and \( W(P) + W(Q) \leq 1 \) then

\[
I_g(P \cup Q) = f^{-1} \left( \frac{W(P) f(I_g(P)) + W(Q) f(I_g(Q))}{W(P) + W(Q)} \right)
\]

(4.7)

He then proves that RG1-4 (§4.3.2) together with RG5’, where \( f(x) \) is taken to be \( f_\alpha(x) = 2^{(\alpha-1)x} \), uniquely characterize (4.6).

Rényi cites two points in support of \( H_\alpha(P) \) as an appropriate family of measures of the information contained in a distribution \( P \). Firstly, that \( H_\alpha(P) \) is characterized by the same properties as \( H_g(P) \) excepting the use of an exponential mean in RG5’ rather than the arithmetic mean in RG5. Secondly, he claims (without explicit justification) that \( H(P) \) is the limiting case of \( H_\alpha(P) \) as \( \alpha \to 1 \). He thus makes a case for a continuum of measures, characterized by the same properties, containing Shannon’s \( H \) as a special case.

Rényi does not state that a choice of parameter \( \alpha \) is a desirable feature of a measure of information. Some might take the view that such a feature is only preferable to a unique measure if a unique measure is not convincingly justified. It is not simply a matter of scaling since, for example

\[
H_2(0.5, 0.5) < H_2(0.6, 0.3, 0.1)
\]

whereas

\[
H_{50}(0.5, 0.5) > H_{50}(0.6, 0.3, 0.1)
\]

Rényi also arrives at the expression (4.6) by a different route. He gives a set of postulates to characterize a quantity he calls the ‘gain of information’ \( I_g(Q|P) \), adapts postulates RG1-4 and 5’ to apply to this quantity, and arrives by means of the characterization of \( I_g(Q|P) \) from these adapted postulates at the same expression (4.6) as was obtained by direct characterization. He also proves that these are the only two possibilities (\( H_\alpha \) for \( \alpha \neq 1 \) and Shannon’s \( H \)) for an information measure \( I_g(Q|P) \) satisfying these adapted postulates. However, further investigation of this approach is beyond the scope of this study, as it is not directly comparable to the other characterizations of \( H \) under consideration.
Perhaps different orders of information might suit different applications, although Rényi does not conjecture any specific application where a measure of order $\alpha \neq 1$ would be preferable to Shannon’s measure.

Although $H_{\alpha}$ is additive (by RG4), it is not subadditive for $\alpha \neq 1$, as Rényi [24] shows. This might be considered an objection to the suitability of $H_{\alpha}$ as a measure of information, since the argument for subadditivity of an information measure was found to be convincing (§3.6.3). Another potential objection to the entropies of order $\alpha \neq 1$ is that, like $H_{g}$, they are not recursive, and furthermore do not possess the Branching property.

**Lemma 6.** $H_{\alpha}$ does not have the branching property for $\alpha \neq 1$.

**Proof.** For $(p_1, \ldots, p_n) \in \Gamma$, $\alpha \neq 1$

\[
H_{\alpha}(p_1,\ldots,p_n) - H_{\alpha}(p_1+p_2,p_3,\ldots,p_n) = \frac{1}{1-\alpha} \log_2 \left( \frac{\sum_{i=1}^{n} p_i^\alpha}{(p_1+p_2)^\alpha + \sum_{i=3}^{n} p_i^\alpha} \right)
\]

This cannot be expressed as a function of $p_1$ and $p_2$ only, so $H_{\alpha}$ does not have the Branching property (3.3).

Therefore, $H_{\alpha}$ does not have any of the recursivity properties discussed above (§3.5), since each is a special case of the Branching property as discussed in section §3.7.1. However, like $H_{g}$, they have instead a Mean-value property (RG5′), which dictates their behaviour in ‘branching’ situations, and might be considered equally or more desirable on further investigation.

### 4.3.4 Entropies of degree $\beta$

Daróczy [2] characterizes another infinite family of information measures which he calls entropies of type $\beta$ and which are also known (Aczél et al. [1], Ebanks et al. [3]) as the entropies of degree $\beta$.

\[
H^\beta(P) = \frac{1}{2^{1-\beta} - 1} \left( \sum_{i=1}^{n} p_i^\beta - 1 \right)
\]

(4.8)
for $P = (p_1, p_2, \ldots, p_n) \in \Gamma$, $\beta \in \mathbb{R}$, $\beta > 0$, $\beta \neq 1$

He proves that the following properties are sufficient to uniquely characterize (4.8).

**D1** 3-symmetry (§3.3)

**D2** Normalization (§3.10)

**D3** Recursivity with $M(x) = x^\beta$ (which is multiplicative, §3.5.3)

Aczél et al. [1] note that the entropies of degree $\beta$ are subadditive (§3.6.3) but, as Daróczy [2] notes, they are not additive (§3.6.1). The latter might be considered an objection to the entropies of degree $\beta$, since the argument for the desirability of Additivity of an information measure was found to be compelling. Similar comments apply here as above (§4.3.3) to the feature of a choice of parameter.

Daróczy [2] notes the following relation between the entropy of degree $\beta$, $H_\beta$ and Rényi’s entropy of order $\beta$, $H_\beta$ for a given distribution $P \in \Gamma$.

$$H_\beta(P) = \frac{1}{1 - \beta} \log_2 \left( (2^{1-\beta} - 1)H_\beta + 1 \right)$$

Thus, for fixed $\beta$, each is a monotonic variant of the other, so for the purpose of maximization they are effectively equivalent.

### 4.4 Other characterization results

This section presents some characterization results which, from the assumption of various properties of an information measure, prove restrictions on its form.

#### 4.4.1 Sum Property

Ng [17] proves the following consequence of the Branching property (§3.5.1) assumed with Symmetry (§3.3) and Expansibility (§3.4).
Theorem 1. (Ng [17]). If an information measure $I$ is symmetric, expansible and branching then there exists a mapping $g : [0, 1] \rightarrow \mathbb{R}$ with $g(0) = 0$ such that

$$I(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} g(p_i) + I(1) - g(1)$$

for all $P \in \Gamma$.

Ebanks, Sahoo & Sander [3] consider an adapted version of this result (with $g$ restricted to the open domain $(0, 1)$). They call this the ‘sum property’, and stress its significance: that, for any such measure, the quantity of information contained in a distribution is bound to be a sum of values, each of which depends only on a single component of the distribution (with the possible addition of a constant term where the closed domain is admitted). They observe that “nearly all measures of information used in applications” have this property, and that it “seems to be a very fundamental property” of information.

The assumptions of the theorem seem reasonable: the Branching property is not subject to the criticism of implicit linearity as are the stronger recursivity properties, and the arguments for Symmetry and Expansibility seem compelling.

While this result applies only to the domain, $\Gamma$ (1.2), of complete distributions, it shows that each term from such a symmetric, branching, expansible distribution contributes independently to the measure of information of that distribution. Thus this result seems to give weight, especially in view of the weakness of its axioms, to Feinstein’s and Rényi’s proposal (§2.2, §4.3.2): that an information measure should be applicable to individual probabilities and non-exhaustive sets thereof.

4.4.2 Properties of $I_2(1 - p, p)$

Various authors (e.g. Aczél et al. [1], Daróczy [2], Ebanks et al. [3], Lee [15], Tverberg [30]) explore the properties of information measures through results concerning the restricted domain $\Gamma_2$ (1.3). Although this body of work is relevant to this study, a thorough investigation of it is omitted due to limitations of time. However, the following result is included to demonstrate how assumptions of properties of $I$ (on
domain $\Gamma (1.2)$) may be shown to have consequences for its restriction to $\Gamma_2$, which in turn have implications for the unrestricted measure.

Aczél, Forte & Ng [1] prove the following characterization result for $I_2$.

**Lemma 7.** (Aczél et al. [1]). If $I$ is additive, symmetric and subadditive\(^6\) then $h(x) = I(1 - p, p)$ for $p \in [0, 1]$ has the following properties:

- symmetry with respect to $\frac{1}{2}$, i.e. $h(p) = h(1 - p)$
- non-decreasing monotonicity on $[0, \frac{1}{2}]$ and non-increasing monotonicity on $[\frac{1}{2}, 1]$
- continuity on $(0, 1)$
- concavity on $[0, 1]$:
- its right and left derivatives $D^+ h$ and $D^- h$ exist everywhere on $[0, 1)$ and $(0, 1]$ respectively
- $D^+ h$ and $D^- h$ are finite on $(0, 1)$
- $D^+ h(x) \geq 0 \ \forall x \in [0, \frac{1}{2}) \ \text{and} \ D^- h(x) \geq 0 \ \forall x \in (0, \frac{1}{2}]$

Thus, these seemingly weak assumptions (§3.6.1, §3.3, §3.6.3) actually determine a great deal about the characteristics of $I_2$. These in turn can be used to further characterize the unrestricted measure $I$, as demonstrated by Aczél et al. [1] in their proof of the following theorem, which uses the above theorem as an intermediate result.

### 4.4.3 Characterization of additive measures

Aczél, Forte & Ng [1] also prove the following characterization result for additive measures.

---

\(^6\)In fact, they also show that a weaker version of subadditivity is sufficient.
Theorem 2. (Aczél et al. [1]). An information measure $I$ is subadditive, additive, expansible and symmetric if and only if it may be expressed as a linear combination of the Shannon and Hartley entropies:

$$I(P) = aH(P) + b_0H(P)$$

for $P \in \Gamma$ and some non-negative constants $a, b \in \mathbb{R}$

This seems to be a powerful theorem due to the apparent weakness of its axioms (all found to be reasonable in chapter 3 above) compared to the strength of its conclusion. If we accept its four premises as desiderata of an information measure $I(P)$, then we can be sure that any such measure is a non-negative multiple of Shannon’s measure with a non-negative constant term depending on the number of possible outcomes of $P$.

The assumptions of this theorem are also used in the Aczél et al. characterizations, both of $H$ (§4.1.4) (with the additional axioms of Normalization and Being Small for Small Probabilities), and of $_0H$ (§4.3.1) (with, additionally, Normalization and Insensitivity). (In fact, these characterizations are presented [1] as corollaries of the theorem.) Thus, as they note, it is clear that $aH(P)$ is the continuous part and $b_0H(P)$ the discontinuous part of any such subadditive, additive, expansible and symmetric measure $I$.

### 4.5 Summary

Various different characterizations and other justifications of Shannon’s function have been considered and compared. The fact that there exists such a variety of different characterizations (from different assumed properties of an information measure), all of which find Shannon’s $H$ to be the unique appropriate measure seems, paradoxically, both to support and to undermine the suitability of the measure itself. On the one hand, if one objects to a particular assumed property of one characterization, there will be another which does not require the same assumption; one can choose the characterization whose assumptions one feels to be the most reasonable. On the
other hand, the fact that the variety exists suggests that it is not clear exactly what
the desirable properties of information are, so no justification can ever be universally
convincing. This author found the characterization of $H$ by Aczél et al. [1] to be
superior to the others on the grounds of having the most reasonable assumptions.

Characterizations of alternative measures have been presented for the sake of
comparison, both of their properties and of their characterizations, with those of
Shannon’s measure. The existence of alternative information measures with different
properties from those of $H$ suggests that different people, or different applications,
require different properties of their information measures, so there will never be uni-
versal agreement on a unique ‘best’ measure. Little seems to have been said in the
literature about the suitability of any one measure or family of measures of informa-
tion compared to any other in any given situation. This may be because it is simply
a matter of personal preference: one chooses the measure with the most convenient
properties for the particular application. From a logical point of view, with no par-
ticular application in mind, one similarly chooses the measure with the properties or
characterization one feels to be the most reasonable, though these may differ from
those preferred by the next person.

The arguments for defining a measure of information for incomplete distributions,
including individual events, and a related ‘mean-value’ interpretation of the informa-
tion measure of a complete distribution were examined and found to be convincing.

More general characterization results have also been considered, showing how
the assumption of various properties of an information measure is known to place
restrictions on the form of any such measure.
Chapter 5

The Maximum Entropy Principle

5.1 Introduction

‘Ill-posed’ problems, where there is insufficient information (in the natural language sense) given to entail a unique solution by deductive reasoning, require the use of ‘inductive’ reasoning. The Maximum Entropy Principle (MEP) provides a method of dealing with a class of ill-posed problems by dictating how to assign a ‘best estimate’ of probability values where there is insufficient information to assign values with certainty.

It was first used in statistical mechanics by Boltzmann and Gibbs in the late nineteenth and early twentieth centuries and has more recently (from 1957) been championed by Jaynes [7], [8], [9], [10], whose own work is mostly in the field of statistical mechanics, but who asserts the wider applicability of the MEP. From the late twentieth century until the present day, various authors (Paris [18], [19], Paris and Vencovská [20], [21], [22], Shore and Johnson [27], Tikochinsky et al. [29]) have worked to justify the MEP as a method of uncertain inference in its own right, independently of application, and with mathematical rigour.
5.2 Applicability

In order to apply the MEP one must have identified a desired probability distribution of known size $n$, and have a set of testable (against candidate distributions) constraints, $K$, which express the entire body of available information which is relevant to the problem. In order to ensure that a solution exists and is unique, it is sufficient to assume that the solution space of possible distributions consistent with $K$ is a convex, closed region.¹ Under these assumptions, the MEP may be stated as follows.

Maximum Entropy Principle

*From all distributions $P = (p_1, p_2, \ldots, p_n)$ consistent with the given data $K$, choose that for which Shannon’s Entropy measure

$$H(P) = -\sum_{i=1}^{n} p_i \log p_i$$

(5.1)

is maximal.*

It should be noted that the use, or even the consideration, of the MEP depends on an interpretation of probabilities as representing a (subjective) state of knowledge. Jaynes [7] explains this position: “The ‘subjective’ school of thought regards probabilities as expressions of human ignorance; the probability of an event is merely a formal expression of our expectation that the event will or did occur, based on whatever information is available”.

Paris [18], from the logical perspective on uncertain reasoning, uses an interpretation of probabilities which renders them (arguably) even more subjective: that they represent the *degree of belief*, of some reasoning agent, associated with a given proposition. He acknowledges that the assumption that degree of belief can be modelled numerically is questionable, but justifies it convincingly with the argument that humans are able to express belief in such a way, e.g. on a scale modelled by the interval $[0,1]$.

There is a more restrictive interpretation held by some; that probabilities may only

¹This may be achieved by requiring certain properties of $K$, although these will not be discussed here, see Paris [18].
CHAPTER 5. THE MAXIMUM ENTROPY PRINCIPLE

refer to (actual or predicted) experimental frequencies. Under this interpretation the question of how to assign unknown probabilities is meaningless and the MEP, or any such rule, wholly unjustifiable. Jaynes [7], [9] covers this disagreement in detail, and argues keenly and convincingly for the subjective interpretation as being both perfectly valid, and the more useful of the two. The subjective interpretation will be assumed throughout this study.

5.3 Justification

Several justifications of the MEP have been proposed, some of which are considered below. These take diverse approaches based in a) information theory, b) logic, c) statistics and d) empiricism.

5.3.1 Maximal Non-commitment (an information theoretic justification)

This argument is based on the idea that, when using inductive reasoning, one should take into account all available relevant information, but avoid making any assumption regarding unavailable information. This will be referred to as the Principle of Maximal Non-commitment. According to Jaynes [9] there are ancient sources, such as the Old Testament, Herodotus and Ovennus, which cite this principle: the virtue of considering all possibilities (i.e. of not presuming information which one does not have) in order to make wise decisions.

Maximal Non-commitment seems intuitively appealing, as it expresses part of our everyday, ‘commonsense’ notion of how to reason with incomplete information. If one accepts this principle then it follows that, since assumptions reduce uncertainty (in a model), one would wish to maximize the measure of uncertainty of one’s solution, consistent with the given information. This is precisely what the MEP does using \( H \) as its measure of uncertainty.

\(^2\)For this reason, this justification seems to belong in the ‘common sense’ category discussed below, but will be treated separately to mirror its treatment in the literature.
If one also accepts Shannon’s Entropy $H$ (5.1) as the optimal measure of uncertainty then this argument gives a compelling justification of the MEP. This is that, for problems satisfying the necessary assumptions (§5.2), it selects the distribution with maximal uncertainty, consistent with the given information; which must therefore be that which is maximally non-committal regarding absent information. Of course, all comments in previous chapters regarding the relationship of information to uncertainty and the justification of Shannon’s measure apply here.

Jaynes [7] uses this two-part argument. Firstly, he presents a version of Shannon’s characterization of $H$ (§4.1.1) as the unique measure of information to satisfy the identified desiderata. He then asserts, making tacit use of the Principle of Maximal Non-commitment, that “in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have”.

Similarly, Paris [18] reworks Shannon’s characterization of $H$ as the unique measure of uncertainty from the same desiderata. He then reasons that, in a problem of uncertain inference, the solution, consistent with the given information, for which $H$ is maximal must therefore be that which “goes as little beyond the information given...as possible”.

Given the two assumptions above, of the Principle of Maximal Non-commitment and of $H$ as the optimal measure of uncertainty, this justification is compelling. No criticism has been found of the Principle of Maximal Non-commitment, although its intuitive nature makes it difficult to conclusively assert or refute that it soundly and completely describes how uncertain reasoning should be performed in situations where the MEP is applicable. As discussed in previous chapters, the assumption of $H$ as the optimal measure of information is not universally accepted.
5.3.2 Common Sense and Consistency (a logical justification)

It seems intuitively clear that ‘common sense’ plays an important role in reasoning, although exactly what common sense is and how it is employed are difficult to define (Paris [18]). Still, it would seem desirable that any formal method of reasoning should be consistent with common sense, as far as it *can* be defined. Thus, various authors have cited ways in which the MEP is consistent with common sense in order to justify its use as a method of inference.

**Jaynes**

Jaynes [10] claims that the MEP is the “natural, common sense way in which anybody does think about his problems of inference” because “it picks the solution favoured by the prior information which means, usually, that having the greatest multiplicity ($w$) or greatest entropy ($\log w$)”.

He gives the example of a driver who must make a decision at a junction based on what he can see. Rather than considering all the things he might see but does not (an allusion to the approach of sampling theorists), the driver thinks about the class of all contingencies consistent with what he does see, and acts according to which seem most likely from his previous experience. This intuitive argument seems reasonable in support of the method of the MEP as a model of human reasoning, but it lacks the weight of a mathematical argument.

**Shore and Johnson**

Shore and Johnson [27] attempt to axiomatize the generalized (to include continuous distributions) Principle of Maximum Relative Entropy, which includes the MEP as a special case. They claim to prove that the MREP is the unique correct rule of inference by showing that it, uniquely, satisfies four consistency requirements. However, their work depends on the assumption that the inference rule proceeds by maximizing some information function of the distribution under consideration. This is a stronger
version of the Principle of Maximal Non-commitment, and has been criticized as unreasonable (Paris and Vencovská [21], Uffink [31]).

Paris and Vencovská

Paris [18] and Paris and Vencovská [21], [19] formalize in mathematical terms some ‘commonsense’ principles of inductive reasoning, and show that the MEP satisfies these, in support of its use in uncertain inference.\(^3\) Furthermore, they prove that acceptance of these principles uniquely characterizes the MEP (see theorem 3 below).

They examine the MEP from a logical point of view, through the Maximum Entropy Inference Process (MEIP). This provides a framework within which any problem of uncertain reasoning, irrespective of its real-world context, may be considered, provided that it can be expressed in terms of a finite propositional language \(L = (v_1, \ldots, v_n)\) and the sentences \(SL\) built from \(L\) using the logical connectives \(-, \land\) and \(\lor\). The assumptions discussed above (§5.2) also apply. The distribution to be found \((w_1, \ldots, w_J)\) (where \(J = 2^n\)) is that of the atoms of the language, \(\alpha_1, \ldots, \alpha_J\), which maps to a probability function \(w : SL \rightarrow [0, 1]\) where \(w(\alpha_i) = w_i\). This function entails the probability of any expression \(\theta \in SL\), since \(w(\theta) = \sum_{\alpha \in S_\theta} w(\alpha)\) (Paris [18]). One obvious benefit of this approach is that it allows general results to be established, relevant to any such application of the MEP.

An essential difference between this approach and that considered above is that, previously, the distributions to which \(H (5.1)\) can be applied have been interpreted as representing random experiments or other random systems of mutually exclusive events. Under this interpretation there is an implicit symmetry of status between the individual probabilities of a distribution, since each corresponds to a known possible event. In the MEIP framework, however, the ‘given events’ are the propositional variables \(v_j\) for \(j = 1, \ldots, n\), which would seem to share this symmetry of being known

\(^3\)The mathematical formulations are not reproduced here since this would require the introduction of a considerable amount of notation and technical detail. This chapter will discuss the principles and their justifications in general terms.
possible events, while the probabilities under consideration, \( w_i \) for \( i = 1, \ldots, J \), correspond to the atoms of the language. The atoms represent combinations of propositional variables and their negations, some of which may not, in reality, be possible; although whether they are possible or not may be unknown. Therefore, it seems unclear whether the properties of \( H \) and the MEP so far discussed, particularly the Symmetry property (§3.3), are desirable with this alternative interpretation of what a distribution represents. However, an investigation of this question is beyond the scope of this study, and it will be assumed henceforth that \( H \) is applicable in this framework.

As Paris and Vencovská [21] note, another difference between this approach and that considered above is that, for an arbitrary \( \theta \in SL \), this approach does not aim to find the ‘best estimate’ of \( w(\theta) \), but rather the value \( w(\theta) \) which logically follows from the given set of constraints \( K \). So, whereas the information theoretic approach aims to approximate a ‘true’, real-world distribution, this aims to yield one which commonsense reasoning would produce from the given information. It is interesting that these two, seemingly quite different, approaches are found to have the same (ME) solution.

Paris and Vencovská [21] consider the choice of \( w \) as an instance of an inference process \( N \) mapping sets of linear constraints to weight-assigning functions, and assert that “there are natural requirements of consistency and independence on \( N \) that severely limit the possible choices of \( N(K) \) for a given set of constraints \( K \)”\]. They identify the following principles, in slightly various forms, as being necessary to any inference process that is consistent with ‘common sense’.

**Equivalence Principle:** Jaynes [8] defined the ‘basic desideratum of consistency’ in the context of uncertain reasoning to be that “in two problems where we have the same prior information, we should assign the same prior probabilities”. Paris [18], [19] and Paris and Vencovská [21] formalize this idea with the principle that two sets of constraints, \( K_1, K_2 \), which are equivalent, in the sense that they have the same solution space, should generate the same solution \( N(K_1) = N(K_2) \). This is clearly
CHAPTER 5. THE MAXIMUM ENTROPY PRINCIPLE

desirable for consistency.\footnote{Paris [18] shows that equivalence of two knowledge bases in this sense is equivalent to the requirement that the constraints of one may be derived from those of the other, so it is irrelevant how the constraints are expressed.}

**Renaming Principle:** This demands that changing the labelling of events should not affect the probabilities assigned to them. Their justification (Paris [18], [19] and Paris and Vencovská [21]) is that the atoms of a language all share the status of being possible worlds, which symmetry means that it should not matter in what configuration they are considered.

In one sense this seems intuitively appealing. However, as Paris [18] acknowledges, a permutation of the atoms of a language can cause literals (propositional variables and their negations) to be exchanged not only with literals but with more complex expressions, representing relationships between literals. In this light, it seems less clear whether or not this property is desirable.

**Irrelevant Information Principle:** This is the requirement that any information irrelevant to the problem under consideration may be included or ignored without affecting the solution. This principle is very appealing, since if it did not hold this would admit the absurd position of being able to generate different solutions to a problem by adding more and more irrelevant information to the knowledge base from which it is considered.

Paris [18], [19] and Paris and Vencovská [21] formulate the principle in terms of propositional variables. It seems that there may also be a case for formulating it in terms of atoms, since this would be more widely applicable.

**Obstinacy Principle:** This expresses the idea that new information which does not contradict existing conclusions should not cause any change to existing conclusions. Their justification (Paris [18], [19] and Paris and Vencovská [21]) is that if new information $K_2$ does not contradict existing conclusions $N(K_1)$, this means that all information in $K_2$ would be believed on the basis of $K_1$, so does not provide any grounds to change established beliefs, i.e. $N(K_1 + K_2) = N(K_1)$. This is intuitively appealing and seems to model human reasoning; having already formed an opinion about something from available evidence, one is not normally inclined to change this
opinion for anything short of a clear contradiction.

**Relativization Principle:** The essence of this principle is that the conditional probabilities one would assign *given that* some event had occurred should only depend on knowledge that one *would have if* that event had occurred. So conditional probabilities depending on its *non-occurrence* may be included or ignored, without affecting the values assigned conditional on the occurrence. This seems similar to the Irrelevant Information principle, and is intuitively appealing in the same way.

**Independence Principle:** This expresses the idea, through a special case, that where there is no evidence in a knowledge base of statistical correlation between propositional variables, statistical independence should be assumed.

This reasoning seems to depend on the assumption, necessary to apply the ME(I)P (§5.2), that the given constraints consist of *all* available relevant knowledge, i.e. that, for any propositional variables which are relevant (referred to in the knowledge base), their degree of correlation would also be relevant information, so would already be contained in the knowledge base. Therefore, absence of information regarding correlation may be taken to mean absence of correlation.

Even under this assumption, this principle does not seem as compelling as some others, since the fact that certain relevant information is *not* known is a premise of the use of inductive reasoning.

**Open-mindedness Principle:** Jaynes [7] notes an “important property” of the MEP: “that no possibility is ignored; it assigns positive weight to every situation that is not absolutely excluded by the given information”. Paris [18], [19] and Paris and Vencovská [21] formalize this with the requirement that, where a non-zero probability is consistent with the knowledge base, zero probability may not be assigned. Their justification is that to assign zero probability is to classify an event as impossible, thereby introducing unwarranted certainty and contradicting common sense.

This seems appealing from one perspective, although another argument is that to prohibit assignment of zero probability is simply a different unwarranted assumption: that each event described by an atom of $L$ really exists, when in fact some atoms may actually have probability zero. This point of view is discussed further in chapter 6,
Continuity Principle: The essence of this principle is that ‘small’ changes in the knowledge base (defined in terms of the Blaschke metric [18]) should not bring about ‘great’ changes in the solution. Paris [18], [19] gives the justification that a system of inference should be robust to small, random fluctuations in a knowledge base without producing appreciable changes to the resulting probabilities assigned.

This seems appealing in general, but not absolutely compelling. It seems conceivable that there might be special cases where a small change in premises could reasonably cause a significant change in conclusions. For example, under the interpretation of probabilities 0 and 1 as representing respectively impossibility and certainty, it seems conceivable that in some circumstances it may be desirable for changes from values 0 or 1 to values arbitrarily close to these to produce significant changes in conclusions, due to the philosophical change in status.

Discussion: Paris and Vencovská remark [21] that “while one may not be totally committed to the principles in theory it is no easy task, we believe, to be happy with them failing in a real inference situation”. This author’s conclusion is similar; that while the justifications for these principles are short of being absolutely convincing, neither can any convincing objection to the principles be provided. As remarked above, ‘common sense’ is a difficult thing to define. It is somewhat easier, given a proposition, to discuss whether it is or is not consistent with common sense, as has been attempted here with the above principles. However, it is still no easy task to say with conviction whether, taken together, they adequately formalize the essence of common sense reasoning; and such will not be attempted.

Paris and Vencovská [21] prove that acceptance of the above principles of inductive reasoning, without continuity, as necessary features of an inference process uniquely characterizes the MEIP.

Theorem 3. (Paris and Vencovská [21]). The Maximum Entropy Inference Process is the only inference process satisfying the principles of equivalence, renaming, irrelevant information, open-mindedness, relativization, obstinacy and independence.
(In a later version (Paris [19]), open-mindedness is replaced by continuity). So, if the above principles (in either given combination) are accepted as necessary to common sense, it follows that the MEIP is the only process for uncertain reasoning which is consistent with common sense, a justification indeed!

5.3.3 Maximal Multiplicity (a statistical justification)

Jaynes [8] shows on strictly combinatorial grounds that the ME distribution is the most likely in any given situation, using the following argument. Suppose that a quantity $x$ may take $n$ different values $x_1, \ldots, x_n$ and that there exist constraints on the desired distribution $(p_1, \ldots, p_n)$ where $p_i = Pr(x_i)$, in the form of mean values of several functions $g_1(x), \ldots, g_m(x)$ where $m < n$.

According to the MEP one would maximize $H$ (5.1) according to the constraints

$$\sum_{i=1}^{n} p_i = 1$$

(5.2)

$$\sum_{i=1}^{n} p_i g_k(x_i) = G_k$$

(5.3)

for $k = 1, 2, \ldots, m$, where the $G_k$ are the prescribed mean values.

Compare this to the situation where the value of $x$ is determined by some random experiment, each repetition of which yields one of the values $x_i$. Suppose the experiment is repeated $M$ times, yielding $m_1$ occurrences of $x_1$, $m_2$ occurrences of $x_2$ etc. with

$$\sum_{i=1}^{n} m_i = M$$

(5.4)

Assuming that the results are consistent with the given constraints (5.3)\(^5\) we have

$$\sum_{i=1}^{n} m_i g_k(x_i) = MG_k \quad k = 1, \ldots, m$$

(5.5)

Where $m < n - 1$, these constraints are insufficient to determine the frequencies $f_i = \frac{m_i}{M}$. However, some sets of frequencies are much more likely than others. The

---

\(^5\)Jaynes [8], [9] also deals with the case when the results are not consistent with the given constraints, which is interesting but is beyond the scope of this study.
number of possible ways of obtaining a given set of sample numbers \( \{m_1, \ldots, m_n\} \) is the multinomial coefficient
\[
W = \frac{M!}{m_1! \cdots m_n!} = \frac{M!}{(Mf_1)! \cdots (Mf_n)!}
\] (5.6)
so the set of frequencies \( \{f_1, \ldots, f_n\} \) which can be realized in the greatest number of ways is that which maximizes (5.6) subject to the constraints (5.4) and (5.5). This may be achieved by maximizing any monotonic increasing function of \( W \), such as \( \frac{1}{M} \log W \). As \( M \to \infty \), by the Stirling approximation,
\[
\frac{1}{M} \log W \to -\sum_{i=1}^{n} f_i \log f_i = H(f_1, \ldots, f_n)
\] (5.7)
This demonstrates that applying the MEP determines the distribution, consistent with the given constraints, that is most likely over a large number of trials. Jaynes [8] asserts that this argument is valid even when the constraints are not in the form of mean values, provided that they are in the form of ‘testable’ (against any proposed distribution) information.

Jaynes [8] also demonstrates that the maximum for \( W \) is extremely sharp, meaning that the ME distribution can be realized experimentally in overwhelmingly more ways than any other.\(^6\) Jaynes [9] uses this argument to great effect by comparing the ME solution to a particular problem with an alternative solution proposed by one of his contemporaries. Both models agree with the given data: a mean value of 4.5 for the result of the toss of a given die. Jaynes’ ME solution has entropy 1.61 while the alternative Urn Model solution has entropy 1.41. Jaynes calculates the multiplicity of each solution and shows that for every way in which the alternative solution could occur, his could be realized in \( 10^{86} \) different ways, with the number of trials taken to be 1000.

Thus, in ill-posed problems, the most likely distribution consistent with the given data is that selected by the MEP (that for which \( H \) is maximal). This fact certainly supports the use of the MEP in such problems, where the most likely of all possible solutions is surely exactly what is required.

\(^6\)Paris and Vencovská state [18], [20] and rigorously prove [20] refined but essentially similar theorems in the context of the Maximum Entropy Inference Process. While relevant to this study, limitations of time have meant that these have not been studied in depth by this author.
It is worth emphasizing that this argument does not depend on any assumption about the significance or interpretation of the function $H$, i.e. $H$ is not assumed here to be a measure of information. Neither is any ‘commonsense’ assumption about how to reason under uncertainty necessary to this argument. In fact, it does not require any assumption beyond those necessary to apply the MEP (§5.1). In this author’s view, this is a strong advantage of the Maximal Multiplicity justification over others considered in this chapter, and renders it superior to them.

There are two symmetric corollaries of the Maximal Multiplicity result, from which one may choose that whose premise one finds the most convincing. Firstly, if $H$ is additionally assumed to be the optimal measure of information, then this result also tells us (something which does not seem intuitively obvious to this author) that the most likely distribution consistent with given data is that which is maximally non-committal regarding missing data, i.e. has the greatest measure of uncertainty. This supports the Principle of Maximal Non-commitment as a rule for inductive reasoning.

Conversely, for those who are prepared to assume that the most likely distribution consistent with given constraints is that for which a measure of uncertainty is maximal, then this result provides support for $H$ as an appropriate measure of uncertainty. (N.B. it does not support a claim that $H$ is the unique such, since any monotonic increasing function of $W$, not only $\frac{1}{MB} \log W$, may be chosen for maximization).

5.3.4 The Proof of the Pudding... (an empirical justification)

Some authors (Jaynes [10], Shore and Johnson [27]) note, in support of the MEP, the wide range of applications where it has been found to give good results in practice, including statistical mechanics and thermodynamics, statistics, traffic networks, spectral analysis and stock market analysis.

Jaynes [10] claims that the results of using the MEP have settled the matter of justification: “It appears, in retrospect, that the development of computer programs capable of dealing with dozens to thousands of simultaneous constraints was the key factor in establishing the power of [the MEP] in a way that transcended all
philosophical arguments”.

While practical applicability does support the validity of the MEP, no amount of empirical success can reduce the desire for sound, theoretical justification. It seems that the philosophical arguments can never be ‘transcended’ in the way Jaynes suggests.

5.4 Criticisms and Limitations

Jaynes [10] acknowledges that the MEP can give inappropriate or misleading results with noisy data, and acknowledges that it is not able to deal with untestable or ‘vague’ data, although such may be relevant and used in human reasoning. This need not be considered a criticism of the MEP, but a limitation of it.

Criticisms related to the above justifications of the MEP have been discussed in the relevant sections. Other criticisms and shortcomings of the MEP as a method of inference are discussed (and answered) by Jaynes [9] and Paris and Vencovská [22].

5.5 Summary

The Maximum Entropy Principle was introduced and some of its various justifications were presented and discussed, along with criticisms of these.

The information theoretic justification (that the MEP yields the distribution which is maximally non-committal regarding unavailable information) was found to be compelling under the assumptions:

- that the Principle of Maximal Non-commitment is necessary and sufficient to characterize inductive reasoning in situations where the MEP is applicable and

- that Shannon’s function $H$ is an optimal measure of uncertainty.

The arguments for and against these assumptions have been discussed in this and previous chapters.
The logical justification (that the MEP yields the distribution which follows by principles of common sense from the available information) was found to be convincing in as much as no outright flaw has been found in the literature or can be suggested. However, both the information-theoretic and logical justifications depend on the acceptance of ‘commonsense’ principles which dictate how inductive reasoning should be performed. The nature of common sense as it applies to uncertain reasoning defies precise definition, making it difficult to know when it has been adequately formalized. For this reason it seems unlikely that either of these approaches can ever be completely convincing.

The statistical justification (that the MEP yields the most likely distribution consistent with the given constraints) was found to be the most compelling since it does not depend on any assumption beyond those required to apply the MEP. With this justification, the fact that the MEP obeys the above commonsense principles could then be considered to justify the principles, rather than the other way around.

The empirical success of the MEP was found to be important but insufficient to justify its use.
Chapter 6

The Minimum Gain Inference Process

6.1 Introduction

The Maximum Entropy Inference Process, by the property of Open-mindedness (§5.3.2), does not assign probability zero to any event (atom) unless constrained to do so by the knowledge base under consideration. Paris [18] claims that this property is desirable, since to assign zero probability to an atom where not obliged to do so introduces a certainty or an assumption which is not justified by the available data. This argument implicitly grants special status to the assignment of probability 0 (and 1 by symmetry), by asserting that such effectively constitutes a certainty or new assumption, whereas assignment of probabilities in the open interval (0, 1) does not.

However, if the assignment of values 0 and 1 is given equal status to assignment of any other value in [0, 1], on the grounds that they are simply members of the range of possible values from which to select, then the above justification of the principle of Open-mindedness no longer applies. Indeed, from a different point of view it could be argued that forcing assignments always to be non-zero where consistent with the given information is in itself introducing unjustified assumptions: of the real-world existence of every scenario modelled by an atom of the language, which may in fact
CHAPTER 6. THE MINIMUM GAIN INFERENCE PROCESS

be unrealistic.

The principle of Occam’s razor may be stated as follows: if two models fit given
data equally well then the one which introduces fewest entities or assumptions is
preferable. Katz [12] offers the justification that if hypothesis H explains the same
evidence as hypothesis G but does so by postulating more entities than G, then, other
things being equal, the evidence has to bear greater weight in the case of H than in
the case of G, and hence the amount of support it gives H is less than it gives G. This
principle seems intuitively appealing, and seems to suggest the opposite approach
from that of open-mindedness: that all atoms should be assigned zero probability
unless constrained otherwise by the knowledge base.

The inference process investigated in this chapter adopts an approach between
these two extremes. It allows the assignment of zero probability to some or all atoms
where consistent with the given knowledge base, in order to minimize

$$G(\vec{x}) = \log N(\vec{x}) - H(\vec{x})$$

where $\vec{x}$ is the desired distribution for the atoms of the language used, $N(\vec{x})$ is the
number of non-zero entries in $\vec{x}$ and $H$ is Shannon’s information function (1.1) with
$K = 1$.

6.2 Derivation of $G$

The expression (6.1) is derived from Rényi’s [23], [24] expression

$$I_n(Q\|P) = \sum_{i=1}^{n} q_i \log \frac{q_i}{p_i}$$

for the ‘gain of information’ when a prior distribution $P = (p_1, \ldots, p_n) \in \Gamma_n$ repre-
senting some random system of events is updated in light of some new information
and replaced by $Q = (q_1, \ldots, q_n) \in \Gamma_n$.

When $P \in \Gamma_n$ replaces the uniform distribution $\mathcal{E}_n = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ the gain of
information is
\[ I_n(P\|\mathcal{E}_n) = \sum_{i=1}^{n} p_i \log np_i \]
\[ = \sum_{i=1}^{n} p_i \log n + \sum_{i=1}^{n} p_i \log p_i \]
\[ = \log n - H(P) \]

since \( \sum_{i=1}^{n} p_i = 1 \).

By maximizing \( H_n(P) \) the Maximum Entropy inference process has the effect of minimizing \( I_n(P\|\mathcal{E}_n) \). If this process is modified to allow assignment of zero probabilities, then \( n \) may be reduced accordingly to \( N \), the number of non-zero entries in \( P \). If \( I_N(P'\|\mathcal{E}_N) \) is then minimized, where \( P' \in \Gamma_N \) consists of precisely the non-zero values of \( P \), then the value obtained can be reduced further.

Thus, this modified process, which will be called the Minimum Gain Inference Process, is designed to yield the solution with the maximum entropy (greater than or equal to the entropy of the ME solution for the same \( K \)) by allowing for the possibility that not all ‘possible worlds’ may be possible! (In fact there is not always a unique such solution, which case is considered below).

### 6.3 Properties of \( G \)

Theorem 2 (Aczel et al. [1]) tells us that, since \( G (6.1) \) is not a non-negative linear combination of the Shannon (1.1) and Hartley (4.1) entropies (since the coefficient of \( \phi H \) is negative), then \( G \) does not have all of the properties of Additivity (§3.6.1), Subadditivity (§3.6.3), Expansibility (§3.4) and Symmetry (§3.3). In fact, it has all of these except Subadditivity, as the following theorem shows.

**Lemma 8.** \( G \) is additive, expansible and symmetric but is not subadditive.

**Proof.** Let \( P = (p_1, \ldots, p_m) \in \Gamma_m \), \( Q = (q_1, \ldots, q_n) \in \Gamma_n \),
\[
P \ast Q = (p_1q_1, \ldots, p_1q_n, \ldots, p_mq_1, \ldots, p_mq_n).
\] Then
\[ G(P \ast Q) = \log(N(P \ast Q)) - H(P \ast Q) \]
\[ = \log(N(P) \cdot N(Q)) - (H(P) + H(Q)) \quad \text{by the Additivity of } H \]
\[ = \log(N(P)) - H(P) + \log(N(Q)) - H(Q) \]
\[ = G(P) + G(Q) \]

Therefore \( G \) is additive (3.12).

\[ G(p_1, \ldots, p_n, 0) = \log(N(p_1, \ldots, p_n, 0)) - H(p_1, \ldots, p_n, 0) \]
\[ = \log(N(p_1, \ldots, p_n)) - H(p_1, \ldots, p_n) \quad \text{by the Expansibility of } H \]
\[ = G(p_1, \ldots, p_n) \]

Therefore \( G \) is expansible (3.2).

For \( \pi \) any permutation of \( \{1, 2, \ldots, n\}, \ n \in \mathbb{N}^+ \)

\[ G(p_{\pi(1)}, p_{\pi(2)}, \ldots, p_{\pi(n)}) \]
\[ = \log(N(p_{\pi(1)}, p_{\pi(2)}, \ldots, p_{\pi(n)})) - H(p_{\pi(1)}, p_{\pi(2)}, \ldots, p_{\pi(n)}) \]
\[ = \log(N(p_1, p_2, \ldots, p_n)) - H(p_1, p_2, \ldots, p_n) \quad \text{by the Symmetry of } H \]
\[ = G(p_1, p_2, \ldots, p_n) \]

Therefore \( G \) is symmetric (3.1).

Therefore, by theorem 2, \( G \) is not subadditive. \( \square \)

### 6.4 Definition of the Minimum Gain Inference Process

Much of the notation introduced in this section is taken from Paris [18]. Let \( L = \{v_1, v_2, \ldots, v_n\} \) be a finite propositional language and \( SL \) the set of all sentences of \( L \), formed by repeated application of the logical connectives \( \land, \lor \) and \( \neg \) to the propositional variables of \( L \). Then let \( At^L \) be the set of atoms of \( L \) where \( |At^L| = J = 2^n \).
Let $CL$ be the set of all finite, consistent sets of linear constraints
\[
\left\{ \sum_{j=1}^{s} a_{ji} w(\theta_j) = b_i \mid i = 1, \ldots, m \right\}
\]
where $w : SL \rightarrow [0, 1]$ is the probability function entailed by the desired distribution $\vec{x}$ for $At^L$, the $\theta_j \in SL$ and the $b_i, a_{ji} \in \mathbb{R}$.

Let $K \in CL$ be the given knowledge base, then let $V^L(K)$ be the solution space of $K$:
\[
V^L(K) = \{ \vec{x} \mid \vec{x}A_K = \vec{b}_K \} \subseteq \mathbb{D}^L
\]
where $\vec{x}A_K = \vec{b}_K$ is an expression of the constraints of $K$ in matrix form, and
\[
\mathbb{D}^L = \{ \vec{x} \in \mathbb{R}^J \mid x_i \geq 0 \quad i = 1, \ldots, J; \quad \sum_{i=1}^{J} x_i = 1 \}
\]
This definition ensures that $V^L(K)$ is a closed, convex region (Paris [18]).

Let
\[
X^L(K) = \{ \vec{x} \in V^L(K) \mid G(\vec{x}) \text{ is minimal} \} \subseteq V^L(K)
\]

**Lemma 9.** $X^L(K)$ is well-defined, finite and non-empty.

*Proof.* $V^L(K)$ is non-empty since $K \in CL$ is consistent. Let $V^L(K)$ be partitioned such that the members of each subset all have the same configuration of zero and non-zero terms, i.e. for any such subset $V_i$ with $\vec{x}, \vec{y} \in V_i$, $x_j = 0 \Leftrightarrow y_j = 0$ for $j = 1, \ldots, J$. At least one such subset is non-empty since $V^L(K)$ is non-empty.

This partition is of finite size $S = 2^J - 1$, since this is the total number of ways of arranging $m$ non-zero terms with $(J - m)$ zero terms for $m = 1, \ldots, J$.

Let $m_i$ be the number of non-zero terms in members of $V_i$ for $i = 1, \ldots, S$. Then $G_i(\vec{x}) = \log m_i - H(\vec{x})$ is continuous and has a unique minimum in any non-empty $V_i$, by the concavity of $H$ and the closedness of $V^L(K)$. Thus $X^L(K)$ is well-defined, since its members may be found by consideration of a finite set of continuous functions, and non-empty.

Furthermore, this means that at most one member $\vec{x}$ of $V_i$ belongs to $X^L(K)$, so $X^L(K)$ is finite. \qed
So \( X^L(K) = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_r \} \) for some \( r \in \mathbb{N} \) with \( r \geq 1 \).

We now define the Minimum Gain Inference Process (MGIP) as follows.

\[
MG^L(K) = \frac{1}{r} \sum_{i=1}^{r} \vec{x}_i
\]

where \( X^L(K) = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_r \} \). I.e. \( MG(K) \) selects the termwise average of the set \( \{\vec{x}\} \subseteq V^L(K) \) for which \( G(\vec{x}) \) is minimal. The fact that \( X^L(K) \subseteq V^L(K) \), which is a convex region, ensures that the average of any number of points from \( X^L(K) \) will belong to \( V^L(K) \) (although not necessarily to \( X^L(K) \) itself).

### 6.5 Behaviour of MG in a simple case

This section presents a study of the behaviour of the Minimum Gain inference process in the following simple case.

Suppose \( L = \{v_1, v_2\} \). Let the atoms of \( L, At^L \), be labelled in the usual way as follows.

\[
\begin{array}{c|c}
\alpha_1 & v_1 \land v_2 \\
\alpha_2 & v_1 \land \neg v_2 \\
\alpha_3 & \neg v_1 \land v_2 \\
\alpha_4 & \neg v_1 \land \neg v_2 \\
\end{array}
\]

Then let the desired distribution of belief values (modelled by a probability function \( w : SL \rightarrow [0, 1] \)) be represented by the vector

\[
\vec{w} = \langle w_1, w_2, w_3, w_4 \rangle
\]

where \( w_i = w(\alpha_i), w_i \geq 0 \) for \( i = 1, \ldots, 4 \) and \( \sum_{i=1}^{4} w_i = 1 \).

Suppose also that \( K_{a,b} = \{w(v_1) = a, w(v_2) = b\} \) for \( a, b \in [0, 1] \). If either \( a \) or \( b \) (or both) takes the value 0 or 1 then there is no freedom of choice in how to assign probabilities to each \( \alpha_i \in At^L \), as follows.

\[
\begin{array}{c|c}
a = 1 & \vec{w} = \langle b , 1-b , 0 , 0 \rangle \\
a = 0 & \vec{w} = \langle 0 , 0 , b , 1-b \rangle \\
b = 1 & \vec{w} = \langle a , 0 , 1-a , 0 \rangle \\
b = 0 & \vec{w} = \langle 0 , a , 0 , 1-a \rangle \\
\end{array}
\]
Therefore it is the case when both \( a, b \in (0, 1) \) which is of interest. In this case, the ME solution (which will not assign zero terms) is

\[
ME(K_{a,b}) = \langle ab, a(1-b), (1-a)b, (1-a)(1-b) \rangle \tag{6.2}
\]

Where an unforced zero is assigned to \( w_i \), this has the effect of adding the constraint \( w_i = 0 \) to the existing constraints

\[
w_1 + w_2 = a, \quad w_1 + w_3 = b, \quad w_1 + w_2 + w_3 + w_4 = 1
\]

This uniquely determines a solution in each case. For example, for the configuration \( \langle 0, +, +, + \rangle \) where the constraints \( w_1 = 0, w_i > 0 \) for \( i = 2, 3, 4 \) are effectively added, this yields

\[
w_2 = a, \quad w_3 = b, \quad w_4 = 1 - (a + b)
\]

and the unique solution \( \langle 0, a, b, 1 - (a + b) \rangle \) where \( a + b < 1 \). Similar calculations yield a unique solution for each of the following possible cases.

<table>
<thead>
<tr>
<th>configuration</th>
<th>solution</th>
<th>condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \langle 0, +, +, + \rangle )</td>
<td>( \langle 0, a, b, 1 - (a + b) \rangle )</td>
<td>( a + b &lt; 1 )</td>
</tr>
<tr>
<td>( \langle +, 0, +, + \rangle )</td>
<td>( \langle a, 0, b - a, 1 - b \rangle )</td>
<td>( a &lt; b )</td>
</tr>
<tr>
<td>( \langle +, +, 0, + \rangle )</td>
<td>( \langle b, a - b, 0, 1 - a \rangle )</td>
<td>( a &gt; b )</td>
</tr>
<tr>
<td>( \langle +, +, +, 0 \rangle )</td>
<td>( \langle (a + b) - 1, 1 - b, 1 - a, 0 \rangle )</td>
<td>( a + b &gt; 1 )</td>
</tr>
<tr>
<td>( \langle 0, 0, +, + \rangle )</td>
<td>n/a (a=0)</td>
<td></td>
</tr>
<tr>
<td>( \langle 0, +, 0, + \rangle )</td>
<td>n/a (b=0)</td>
<td></td>
</tr>
<tr>
<td>( \langle 0, +, +, 0 \rangle )</td>
<td>( \langle 0, a, b, 0 \rangle )</td>
<td>( a + b = 1 )</td>
</tr>
<tr>
<td>( \langle +, 0, 0, + \rangle )</td>
<td>( \langle a, 0, 0, 1 - a \rangle )</td>
<td>( a = b )</td>
</tr>
<tr>
<td>( \langle +, 0, +, 0 \rangle )</td>
<td>n/a (b=1)</td>
<td></td>
</tr>
<tr>
<td>( \langle +, +, 0, 0 \rangle )</td>
<td>n/a (a=1)</td>
<td></td>
</tr>
</tbody>
</table>

All configurations where three values are set to zero are clearly not applicable since each must have either \( a \) or \( b \) equal to 1, and the configuration with all zero entries is inconsistent with the requirement that they sum to 1.

Thus, it can be seen that for a given \( K_{a,b} \) the possible configurations (in terms of the positioning of zero and non-zero terms) of solutions containing zeros are determined by the comparative size of \( a \) and \( b \), and that of \( a + b \) and 1. Thus, for
any pair \(a, b \in (0, 1)\), there are three possible candidates \(W^L(K_{a,b}) = \{\vec{w}_1, \vec{w}_2, \vec{w}_3\}\) for membership of \(X^L(K_{a,b})\): two solutions from the table above, depending on the values of \(a\) and \(b\), as well as the ME solution (6.2). \((ME(K_{a,b})\) is, by its definition and that of \(MG(K_{a,b})\), the only candidate with four non-zero terms).

For which of these three \(G\) is minimal depends on the values of \(a\) and \(b\). For example, when \(a = \frac{1}{8}\), \(b = \frac{3}{4}\),

\[W^L(K_{a,b}) = \{\vec{w}_1, \vec{w}_2, \vec{w}_3\}\]

where

\[
\vec{w}_1 = ME(K_{a,b}) = \left\langle \frac{3}{32}, \frac{1}{32}, \frac{21}{32}, \frac{7}{32} \right\rangle
\]

\[
\vec{w}_2 = \left\langle \frac{1}{8}, 0, \frac{5}{8}, \frac{1}{4} \right\rangle
\]

\[
\vec{w}_3 = \left\langle 0, \frac{1}{8}, \frac{3}{4}, \frac{1}{8} \right\rangle
\]

and

\[G(\vec{w}_1) = 0.447\]

\[G(\vec{w}_2) = 0.198\]

\[G(\vec{w}_3) = 0.363\]

(The natural logarithm was used in these and all following calculations. The choice of logarithm base does not affect the inference process, by the definition of \(G(\vec{x})\) (6.1) provided that it is used consistently).

So

\[X^L(K_{a,b}) = \{\vec{w}_2\}\]

and

\[MG(K_{a,b}) = \vec{w}_2 = \left\langle \frac{1}{8}, 0, \frac{5}{8}, \frac{1}{4} \right\rangle\] (6.3)

Whereas, when \(a = \frac{1}{8}\), \(b = \frac{1}{2}\),

\[W^L(K_{a,b}) = \{\vec{x}_1, \vec{x}_2, \vec{x}_3\}\]

where

\[\vec{x}_1 = ME(K_{a,b}) = \left\langle \frac{1}{16}, \frac{1}{16}, \frac{7}{16}, \frac{7}{16} \right\rangle\]
\[
\vec{x}_2 = \left\langle \frac{1}{8}, 0, \frac{3}{8}, \frac{1}{2} \right\rangle
\]
\[
\vec{x}_3 = \left\langle 0, \frac{1}{8}, \frac{1}{2}, \frac{3}{8} \right\rangle
\]

and

\[
G(\vec{x}_1) = 0.316
\]
\[
G(\vec{x}_2) = 0.124
\]
\[
G(\vec{x}_3) = 0.124
\]

so

\[
X^L(K_{a,b}) = \{\vec{x}_2, \vec{x}_3\}
\]

and

\[
MG(K_{a,b}) = \frac{1}{2} \vec{x}_2 + \vec{x}_3 = \left\langle \frac{1}{16}, \frac{1}{16}, \frac{7}{16}, \frac{7}{16} \right\rangle = ME(K_{a,b})
\]

The above examples show that there exist pairs \((a, b) \in (0, 1)^2\) for which \(MG(K_{a,b}) = ME(K_{a,b})\) as well as pairs where this is not the case. An investigation to determine for which values of \(a\) and \(b\) this holds was undertaken, but unfortunately, due to time constraints, no general results were obtained. This would be a possible area for further study.

### 6.6 Properties of the Minimum Gain Inference Process

The MGIP can be shown to satisfy Paris and Vencovská’s principle (§5.3.2) of Equivalence, and not to satisfy those of Open-mindedness (clearly, by design) or Independence, as shown below. This author has been unable to show whether or not the MGIP satisfies Language Invariance (Paris [18]) or Paris and Vencovská’s other principles of uncertain reasoning. This would be a possible area for further investigation.
**Equivalence Principle** (Paris [18])

If $K_1, K_2 \in CL$ are equivalent in the sense that $V^L(K_1) = V^L(K_2)$ then $N(K_1) = N(K_2)$

The MGIP clearly satisfies Equivalence since

$$V^L(K_1) = V^L(K_2)$$
$$\Rightarrow X^L(K_1) = X^L(K_2)$$
$$\Rightarrow MG(K_1) = MG(K_2)$$

**Open-mindedness Principle** (Paris [18])

If $K \in CL$, $\theta \in SL$ and $K + w(\theta) \neq 0$ is consistent then $N(K)(\theta) \neq 0$.

The MGIP clearly does not satisfy Open-mindedness (by design), as the following counter-example shows. Let $K = \{ w(v_1) = \frac{1}{8}, w(v_2) = \frac{3}{4} \}$. Then by (6.3), $MG(v_1 \land \neg v_2) = 0$ although $K + \{ w(v_1 \land \neg v_2) \neq 0 \}$ is consistent.

**Independence Principle** (Paris [18])

In the special case of $L = \{ v_1, v_2, v_3 \}$ and

$$K + \{ w(v_1) = a, w(v_2|v_1) = b, w(v_3|v_1) = c \} \quad (a > 0),$$

$N^L(K)(v_2 \land v_3|v_1) = bc$.

The MGIP does not satisfy Independence, as the following counter-example shows. Let $a = 1, b = \frac{1}{8}, c = \frac{3}{4}$. Then by (6.3), $MG(K) = \langle \frac{1}{8}, 0, \frac{5}{8}, \frac{1}{4}, 0, 0, 0, 0 \rangle$ where the usual ordering of atoms is used. Therefore, $MG^L(K)(v_2 \land v_3|v_1) = \frac{1}{8} \neq bc$.

### 6.7 Discussion

A very limited investigation of the properties of the Minimum Gain Inference Process has been presented here, and considerable further investigation would be required to allow a meaningful evaluation of it. However, what has been discovered allows some initial discussion.
CHAPTER 6. THE MINIMUM GAIN INFERENCE PROCESS

\( G(6.1) \) is a measure of the difference, for a given distribution \( P \) with \( N \) non-zero entries, between the maximum possible measure of uncertainty for a distribution \( Q \in \Gamma_N \) and the actual measure of uncertainty of \( P \), using Shannon’s measure \( H(1.1) \). It has been shown to have the properties of Additivity (§3.6.1), Expansibility (§3.4 and Symmetry (§3.3) and to lack that of Subadditivity (§3.6.3).

It may seem at first that \( G \) may be interpreted as a measure of the ‘certainty’ of \( P \), since for a given \( P \in \Gamma \) it is simply an inversion and translation of \( H(P) \). However, \( G(P) = 0 \) for any distribution \( P \) representing a certain outcome \( P = (0, \ldots, 0, 1, 0, \ldots, 0) \), so this interpretation is unsatisfactory. \( G \) may better be regarded as a measure of non-uniformity disregarding zero entries, since \( G(Q) = 0 \) for any such ‘uniform’ distribution \( Q = (\frac{1}{N}, \ldots, \frac{1}{N}, 0, \ldots, 0) \), and by a result of Shannon’s [25] [p 21], will increase with any change away from uniformity, where the number of non-zero entries is fixed at \( N \).

The effect of the MGIP may therefore be described as the selection of the most uniform distribution consistent with the given data, from possibilities of any size from 1 to \( J \) (although for cases where there is no unique such and the average of all candidates is taken, no such neat intuitive interpretation can be suggested).

Considered in these terms, it seems difficult to justify the MGIP as a tool for uncertain reasoning. No intuitively convincing reason can be suggested why a more uniform distribution should necessarily be preferable to a less uniform one when it is not given that the distributions are directly comparable (represent the same set of possible events).

The motivation for introducing the MGIP was to explore the consequences of allowing an inference process to fail the Open-mindedness principle, for reasons discussed above (§6.1). Arguments for (§5.3.2) and against (§6.1) Open-mindedness have been presented, none of which has been found entirely convincing.

In evaluating whether Open-mindedness should or should not be required of an inference process, it is relevant to consider that failure of Open-mindedness may, clearly and undesirably, allow real-world possibilities to be discounted. Furthermore, while it may also, desirably, allow real-world impossibilities to be discounted, it does
not ensure that they will be where they cannot be identified as such from the given information. The MG method of assigning zero probabilities involves no mechanism to assess to what extent zero might or might not be a realistic assignment. Thus it is doubtful whether the ‘advantage’ of the latter property outweighs the disadvantage of the former.

On the other hand, an inference process which does obey the principle of Open-mindedness may unrealistically assign positive probability to an actual impossibility, but cannot rule out an actual possibility. This arrangement seems preferable to that described in the previous paragraph, so Open-mindedness seems to be desirable.

Another property of the MGIP whose desirability is questionable is its failure to satisfy the principle of Independence. Although the argument for this principle was not found to be completely convincing (§5.3.2), neither has any convincing argument against it been found.

It seems that further investigation of the MGIP may be useful in order to learn more about inference processes in general, through comparison of its properties with those of the MEP and other inference processes. However, in this author’s view, due to the problem of interpretation and the properties indicated above, the MGIP is not likely to prove useful as a method of uncertain reasoning in its own right.

6.8 Summary

The Minimum Gain Inference Process was introduced as a way of exploring the consequences of allowing the unforced assignment of zero probabilities, and an intuitive justification presented. It was defined mathematically, and its properties were explored and compared to those of the MEP. In particular, it was found that the MGIP obeys Paris and Vencovská’s principle of Equivalence, but not those of Independence or (clearly) Open-mindedness (as defined in Paris [18]). The behaviour of the MGIP was explored in a simple case, although no general results were obtained. A discussion of the interpretation and properties of the MGIP found that it is unlikely to prove useful as a method of uncertain reasoning.
Chapter 7

Conclusions

This study has essentially been concerned with an investigation of the mathematical theory of information, and with the various justifications which exist for Shannon’s measure of information $H$ (1.1) and for its use in the Maximum Entropy Principle. It seems to be the nature of Information Theory that universal agreement on its axioms cannot be achieved. The existence of alternative information measures and characterizations of these is a testament to this. Many of the justifications given for suggested axiomatic properties or principles rely on an appeal to intuition, which will provoke different responses in different people. Likewise, there is no obvious and incontrovertible method to deal with situations of uncertain reasoning; the MEP is one candidate.

Therefore, as stated in the introduction, it was not expected that this study would draw definitive conclusions regarding the optimality or otherwise of Shannon’s measure as a measure of information, or of the MEP as a method of uncertain reasoning, or of any particular justification of either of these. Rather, it was intended to present a balanced account of the literature, and to draw comparisons and reasoned conclusions, which are summarized below. While it is hoped that these have been convincingly justified in the text, it is anticipated that the reader may find alternative justifications more convincing than those preferred here, and from the same material may draw different conclusions from the author’s.

The nature of information, and the questions of how it might be measured, and
what the object of such a measure might be, have been discussed (chapter 2). Shannon’s axiom that the domain of a measure of information should be a set of probability distributions was found to be convincing based on arguments by Shannon himself [25], Feinstein [4], Khinchin [13] and Rényi [23], concerning the relationship of information to uncertainty, and the use of a probability distribution to model an information-carrying event.

The approach of Feinstein [4] and Rényi [23], [24]: to additionally treat individual probabilities and partial (non-exhaustive) probability distributions as objects of an information measure (§2.2, §4.3.2), was found to be justified due to its intuitive appeal. This approach yields an interpretation of the information content of an individual event as being correlated with the ‘element of surprise’ one might feel on learning of its occurrence, which was found to be intuitively agreeable. It also allows the interpretation of the information content of a distribution as a mean value of the information contents of its constituent probabilities, which was also found to be intuitively reasonable.

Various proposed properties of an information measure and some justifications of them were examined and compared (chapter 3), with some found to be more reasonable than others. In the comparison (chapter 4) of various axiomatizations of Shannon’s information function (1.1) based on various combinations of these properties, that of Aczél et al. [1] was found to be superior, due to its avoidance of the assumptions of linearity and of (explicit) continuity, necessary to all others considered.

Of the various justifications of the Maximum Entropy Principle considered (chapter 5), Jaynes’ [7] maximal multiplicity argument was found to be the most compelling, since it establishes that the MEP distribution is overwhelmingly the most likely in a given situation, without reliance on any intuitive assumption about how uncertain reasoning should be performed, or on any particular interpretation of $H$. It also provides a justification of $H$ as a measure of information under the assumption of the Principle of Maximal Non-commitment, or of the Principle of Maximal Non-commitment under the assumption of $H$ as the optimal measure of information.
The Minimum Gain Inference Process was introduced and its general properties and behaviour in a simple case were investigated. It was found unlikely to prove useful as a method of uncertain inference due to certain undesirable properties.

Various areas were identified as candidates for further study. These include a comparison of the branching (§3.5) and additivity (§3.6) properties defined for complete distributions with Rényi’s mean-value properties (§4.3.2, §4.3.3) defined for partial distributions; further study of Paris and Vencovská’s Maximal Multiplicity justification [18], [20] of the MEIP; and further investigation of the properties of the MG inference process in order to discover whether or not it satisfies Language Invariance [18] and those of Paris and Vencovská’s principles [18], [19], [21] which have not yet been determined (§6.6), and to explore its behaviour more widely.
Bibliography


