4open 2022, **5**, 1

https://doi.org/10.1051/fopen/2021004

4 Open

Logical Entropy
Guest Editor: Giovanni Manfredi

Available online at:
www.4open-sciences.org

RESEARCH ARTICLE OPEN ACCESS

Introduction to logical entropy and its relationship to Shannon entropy

David Ellerman*

Faculty of Social Sciences, University of Ljubljana, Ljubljana 1000, Slovenia

Received 23 August 2021, Accepted 5 October 2021

Abstract – We live in the information age. Claude Shannon, as the father of the information age, gave us a theory of communications that quantified an "amount of information," but, as he pointed out, "no concept of information itself was defined." Logical entropy provides that definition. Logical entropy is the natural measure of the notion of information based on distinctions, differences, distinguishability, and diversity. It is the (normalized) quantitative measure of the distinctions of a partition on a set-just as the Boole–Laplace logical probability is the normalized quantitative measure of the elements of a subset of a set. And partitions and subsets are mathematically dual concepts – so the logic of partitions is dual in that sense to the usual Boolean logic of subsets, and hence the name "logical entropy." The logical entropy of a partition has a simple interpretation as the probability that a distinction or dit (elements in different blocks) is obtained in two independent draws from the underlying set. The Shannon entropy is shown to *also* be based on this notion of information-as-distinctions; it is the average minimum number of binary partitions (bits) that need to be joined to make all the *same* distinctions of the given partition. Hence all the concepts of simple, joint, conditional, and mutual logical entropy can be transformed into the corresponding concepts of Shannon entropy by a uniform non-linear dit-bit transform. And finally logical entropy linearizes naturally to the corresponding quantum concept. The quantum logical entropy of an observable applied to a state is the probability that two different eigenvalues are obtained in two independent projective measurements of that observable on that state.

Keywords: Logical entropy, Shannon entropy, Partitions, MaxEntropy, Quantum logical entropy, Von Neumann entropy

Introduction

This paper is an introduction to the concept of logical entropy as the direct measure of the definition of information in terms of distinctions, differences, distinguishability, and diversity. The formula for logical entropy goes back to the early twentieth century, but the current development comes out of seeing the formula as the quantification of information in a partition as the normalized number of distinctions or dits (ordered pairs of elements in different blocks) of the partition. Just as the Laplace–Boole notion of probability, as the normalized number of elements in a subset, quantifies the logic of subsets, so logical entropy, as the normalized number of distinctions in a partition, quantifies the logic of partitions – and hence the adjective "logical." The logical entropy of a partition is, in fact, a probability measure – the probability of obtaining a distinction of the partition in two independent draws from the universe set, just as the logical Laplace–Boole probability of a subset (or event) is the one-draw probability of obtaining an element of the subset.

Far from displacing the usual notion of Shannon entropy; the point is to show that the Shannon entropy of a partition is a different quantification of the *same* notion of information-as-distinctions, i.e., the average minimum number of binary partitions (bits) that need to be joined together to make the *same* distinctions of a partition. In fact, there is a non-linear dit-to-bit transformation that transforms all the concepts of simple, joint, conditional and mutual logical entropy into the corresponding formulas for Shannon entropy, where the latter are especially suited for the theory of coding and communications.

Edwin Jaynes' MaxEntropy method is intended to generalize the Laplace indifference principle by determining the "best" probability distribution consistent with given constraints (e.g., that rule out the uniform distribution of the indifference principle) by maximizing Shannon entropy subject to those constraints. We show that maximizing logical entropy subject to the same constraints gives a different probability distribution. The logical entropy solution is the closest to the uniform distribution in terms of the usual notion of (Euclidean) distance while the Jaynes solution is the closest in terms of the Kullback–Leibler (KL) divergence from the uniform distribution. The notion of information-as-differences also connects

*Corresponding author: david@ellerman.org

to ordinary statistical theory since the metrical version of logical entropy is just twice the usual notion of variance (or equals the variance if one counts unordered pairs), and similarly for the notion of covariance.

There is a quasi-algorithmic method, linearization, that transforms set-based concepts into vector-space concepts. Applied to the set-based concepts of "classical" logical entropy, the linearization to Hilbert spaces generates the quantum versions of logical entropy. The quantum logical entropy of an observable applied to a quantum state is the probability of getting different eigenvalues in two independent (projective) measurements of the observable on that state.

Logical entropy

Partitions on a set

A *partition* $\pi = \{B_1, \ldots, B_m\}$ on a finite set $U = \{u_1, \ldots, u_n\}$ is a set of non-empty subsets $B_i \subseteq U$ called *blocks* that are disjoint and whose union is all of U. A *distinction* or *dit* of π is an ordered pair $(u_j, u_k) \in U \times U$ where u_j and u_k are in different blocks of π. The set of all distinctions of π is the *ditset* $\mathrm{dit}(\pi) \subseteq U \times U$. An ordered pair $(u_j, u_k) \in U \times U$ is an *indistinction* or indit of π if u_j and u_k are in the same block of π, and the set of all indits of π is the *inditset* $\mathrm{indit}(\pi) = \cup_{j=1}^{m}(B_j \times B_j)$. A binary relation $E \subseteq U \times U$ is an *equivalence relation* on U if it is reflexive (i.e., for all $u \in U$, $(u, u) \in E$), *symmetric* (i.e., for all $(u, u') \in E$, $(u', u) \in E$), and transitive (i.e., if $(u, u') \in E$ and $(u', u'') \in E$, then $(u', u'') \in E$). The inditset *indit* (π) of a partition on U is an equivalence relation on U. Given an equivalence relation E on U, two elements are said to be *equivalent*, $u \sim u'$, if $(u,u') \in E$. Let $[u]_E \subseteq U$ be the set of elements of U equivalent to $u \in U$, i.e., an *equivalence class* of E. The set of equivalence classes of E is a partition on U and the inditset of that partition is E. Hence the notion of an equivalence relation and an inditset of a partition are equivalent notions.

Since each ordered pair $(u_j, u_k) \in U \times U$ is either an dit of π or an indit of π but not both, the ditset $\mathrm{dit}(\pi) = U \times U - \mathrm{indit}(\pi)$ is the complement of the inditset in $U \times U = U^2$. As a binary relation $\mathrm{dit}(\pi) \subseteq U \times U$, the ditsets of a partition are called a *partition relation* or an *apartness relation*. Partition relations $P \subseteq U \times U$ can be characterized as being irreflexive (i.e., for any $u \in U$, $(u, u) \notin P$), symmetric, and anti-transitive (i.e., for any $(u_j, u_k) \in P$ and for any sequence $u_j = u_{j_0}, u_{j_1}, \ldots, u_{j_k}, u_{j_{k+1}} = u_k$ of elements of U, there is a pair $(u_{j_i}, u_{j_{i+1}}) \in P$). Every ditset of a partition is a partition relation and vice-versa.

Given another partition $\sigma = \{C_1, \ldots, C_k\}$ on U, the partition π *refines* σ, written $\sigma \precsim \pi$, if for every block $B \in \pi$, there is a block $C \in \sigma$ such that $B \subseteq C$. Intuitively, π is obtained from σ by splitting up some of the blocks of σ which creates more distinctions. Indeed, $\sigma \precsim \pi$ if and only if (iff) $\mathrm{dit}(\sigma) \subseteq \mathrm{dit}(\pi)$. The refinement relation on the partitions on U is a *partial order* in the sense that it is reflexive, anti-symmetric (i.e., if $\sigma \precsim \pi$ and $\pi \precsim \sigma$ then $\sigma = \pi$), and transitive. The partial order has a maximal or top partition and a minimal or bottom partition. The top is the *discrete partition* $\mathbf{1}_U = \{\{u\}\}_{(u \in U)}$ where all the blocks are singletons, and the bottom is the *indiscrete partition* or "*blob*" $\mathbf{0}_U = \{U\}$ with only one block U. Both the join (least upper bound) and meet (greatest lower bound) of two partitions always exist so the refinement partial order is a lattice $\mathbf{\Pi}(U)$.[1] Only the join operation is used here, but all the Boolean operations on subsets can be extended to partitions to form the logic of partitions [3, 4] that is the dual counterpart to the Boolean logic of subsets (which is usually presented in the special case of propositional logic). Given π and σ, the *join* $\pi \vee \sigma$ is the partition on U whose blocks are all the non-empty intersections $B \cap C$ for $B \in \pi$ and $C \in \sigma$.

One of the easiest ways to see the dual pairing of the concepts of a subset and a partition is to consider a function $f: X \to Y$ from a set X to a set Y. The *image* is the subset $f(X) = \{y \in Y : \exists_x \in X, f(x) = y\}$ of the codomain Y, and the *inverse-image* or *coimage* is the partition $\{f^{-1}(y)\}_{y \in f(X)}$ on the domain X (Fig. 1).[2]

Logical entropy: the quantification of distinctions

The set of all subsets of a set U, the powerset $\wp(U)$, also forms a lattice under the inclusion partial order with the top U, the bottom $\emptyset$, and the join and meet being set union and intersection respectively. Given the duality between subsets and partitions, it is natural to see what concepts carry over from subsets to partitions. In particular, the quantitative measure of a subset $S \subseteq U$ is its cardinality $|S|$, and the normalized cardinality of a subset S is the logical notion of probability $\Pr(S) = \frac{|S|}{|U|}$ developed by Boole and Laplace (where each point $u \in U$ is considered equiprobable). Gian-Carlo Rota in his Fubini Lectures [6] and in his lectures at MIT on probability theory [7] argued that information or entropy should be to partitions what probability was to subsets, i.e.,

$$\frac{\text{Probability}}{\text{Subsets}} \approx \frac{\text{Information}}{\text{Partitions}}. \tag{1}$$

[1] In some of the older literature, the partial order is written in the opposite way as "unrefinement," so that interchanges the top and bottom and the join and meet [1, 2].

[2] In category theory, the notion of a subset generalizes to the notion of a subobject or "part" and the "dual notion (obtained by reversing the arrows) of 'part' is the notion of partition" ([5], p. 85).

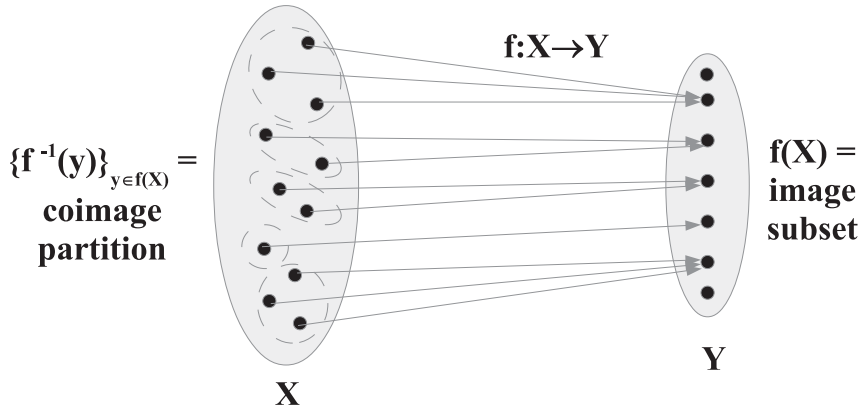

Figure 1. Image subset and inverse-image partition of a function $f: X \to Y$.

The quantitative notion attached to a subset is its number of elements $|S|$, so the question is; What is the quantitative notion associated with a partition? The duality between subsets and partitions can be analyzed back to its conceptual building blocks which are the dual notions of *elements* (*its*) of a subset and the *distinctions* (*dits*) of a partition [8]. Hence, the natural notion of information in a partition would, by this reasoning, be the normalized number of distinctions, and that is our definition of the *logical entropy of a partition* π;

$$\begin{aligned} h(\pi) &= \tfrac{|\mathrm{dit}(\pi)|}{|U\times U|} = \tfrac{|U\times U| - |\mathrm{indit}(\pi)|}{|U\times U|} \\ &= 1 - \tfrac{|\cup_i (B_i \times B_i)|}{|U\times U|} = 1 - \sum_{i=1}^{m} \left(\tfrac{|B_i|}{|U|}\right)^2 = 1 - \sum_i \Pr(B_i)^2, \end{aligned} \tag{2}$$

where $\Pr(B_i) = \frac{|B_i|}{|U|}$ is the probability of a random draw from U will give an element of B_i (with equiprobable points).

When there are point probabilities $p = (p_1, \ldots, p_n)$ for p_j as the probability of the outcome $u_j \in U$ with $\sum_{j=1}^{n} p_j = 1$, then $\Pr(B_i) = \sum \{p_j : u_j \in B_i\}$ in the formula for logical entropy. This also gives the definition of logical entropy for any probability distribution $p = (p_1, \ldots, p_n)$,

$$h(p) = 1 - \sum_{j=1}^{n} p_j^2. \tag{3}$$

Logical entropy always has an ultra-simple and logical interpretation. Logical information theory is built on the idea that information is about distinctions, differences, distinguishability, and diversity. The notion of difference requires *two* things in order to have a difference. Hence, given a partition $\pi = \{B_1, \ldots, B_m\}$ or a probability distribution $p = (p_1, \ldots, p_n)$, the obvious measure for idea of information as distinctions or difference is the probability that in *two* independent samples or draws from U or from the distribution p, one will obtain elements in different blocks of π, i.e., a distinction of π, or different outcomes p_j, p_k for $j \neq k$. And that is precisely the interpretation of logical entropy, the "probability of difference." The probability of obtaining elements from the same block of π is $\sum_i \Pr(B_i)^2$ so the probability of getting elements from different blocks is $h(\pi) = 1 - \sum_i \Pr(\mathrm{B}_i)^2$. And similarly for the logical entropy of a probability distribution $h(p) = 1 - \sum_j p_j^2$. Another way to express this result is the formula:

$$1 = 1^2 = (p_1 + \ldots + p_n)(p_1 + \ldots + p_n) = \sum_{j=1}^{n} p_i^2 + \sum_{j\neq k} p_j p_k \tag{4}$$

so that:

$$h(p) = 1 - \sum_{j=1}^{n} p_i^2 = \sum_{j=1}^{n} p_j(1 - p_j) = \sum_{j\neq k} p_j p_k = 2\sum_{j<k} p_j p_k \tag{5}$$

for $j, k = 1, \ldots, n$. Thus, to be more specific, logical entropy is the probability of getting an ordered pair of distinct indices p_j and p_k for $j \neq k$ – which is twice the probability of getting an unordered pair of different indices such as p_j and p_k for $j < k$.

There is a simple way to picture the logical entropy. Given partition $\pi = \{\{u_1, u_2\}, \{u_3\}, \{u_4\}\}$ with the corresponding point probabilities $p = (p_1, p_2, p_3, p_4)$. Since the sum of the probabilities is 1, the logical entropy $h(\pi)$ can be pictured in a 1×1 box, Figure 2, as the shaded area outside the boxed diagonal.

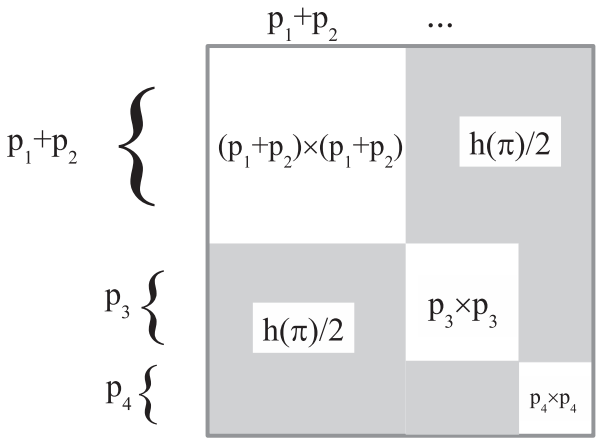

Figure 2. Logical entropy box diagram.

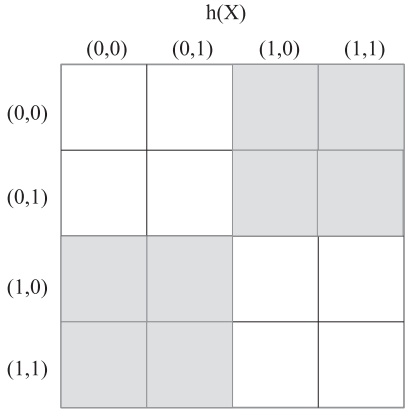

Figure 3. Box diagram for $h(X) = \sum \{p(x, y)p(x', y') : x \neq x'\} = \frac{8}{16} = \frac{1}{2}$ which can also be seen as a Venn diagram.

Logical entropy is also a measure, indeed, a probability measure, in the usual sense of measure theory ([9], p. 30) (although terminology differs) which includes being non-negative. A finitely additive set function (the values on disjoint sets add together) that can take negative values is usually called a "signed measure" ([9], p. 118) (or a "charge" [10]), and, as we will see, Shannon mutual information can be negative.

Partitions often arise as the inverse-images of random variables $X : U \to \mathbb{R}$. To use an example that we will have use of later, consider the throw of one fair die followed by the throw of a second fair die. All that is recorded is whether the face up on each die was even or odd, i.e., its parity (or *mod*(2) value). With even represented by 0 and odd by 1, then the space of possible outcomes for the throws of the dice is $U = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Let $X: U \to 2 = \{0,1\}$ represent the outcome of the first die, the X-die, and $Y: U \to 2$ the outcome of the second die, the Y-die. For instance, the point $(1, 0) \in U$ represents that the first die came up with odd parity 1 and the second die with even parity 0.

A measure on a finite set is determined by just an assignment of a non-negative number to each point in the set. The set on which logical entropy is a (probability) measure is $U \times U$ so it can again be represented in a box diagram with the equiprobable outcome pairs in U along each edge. Each square in Figure 3, representing a pair $((x, y), (x', y'))$ of pairs, has the probability weight of $\frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$ assigned to it. The inverse-image partition of the random variable X is

$$X^{-1} = \{X^{-1}(0), X^{-1}(1)\} = \{\{(0,0),(0,1)\},\{(1,0),(1,1)\}\}. \tag{6}$$

The ditset $\mathrm{dit}(X^{-1})$ is the set of pairs of pairs, i.e., points in $U \times U$, that differ in the first coordinate:

$$\mathrm{dit}(X^{-1}) = \{((0,0),(1,0)),((0,0),(1,1)),((0,1),(1,0)),((0,1),(1,1)),\ldots\}, \tag{7}$$

where the ellipsis ... represents the pairs of pairs with the reversed order. The shaded squares in Figure 3 box diagram are the ones included in the logical entropy $h(X)$ since they are the ones which differ in the first coordinate of the ordered

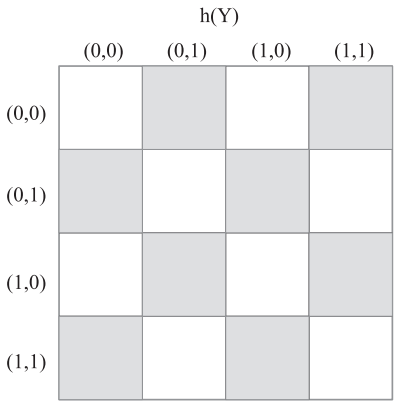

Figure 4. Box/Venn diagram for $h(Y) = \sum \{p(x,y)p(x',y') : y \neq y'\} = \frac{1}{2}$.

pairs of outcomes, i.e.,, the pairs where the first die's outcomes had different parities. Each outcome (x, y) has probability $p(x, y) = \frac{1}{4}$ and the only squares that count for the logical entropy of X are the ones for $((x, y), (x',y'))$ where $x \neq x'$.

The logical entropy of the random variable $Y: U \to 2$ is computed and represented in Figure 4 in the same manner except that the relevant pairs of pairs are those that differ in the second coordinate representing the parity of the second die.

The logical entropy $h(X)$ for X (the parity of the outcome for the first die) and $h(Y)$ for Y (the parity of outcome for the second die) is the probability that on two independent throws of the relevant die, one will obtain outcomes of different parity.

History of the logical entropy formula

The concept of information as a measure of differences goes back to 1641, the year before Isaac Newton was born, when the polymath John Wilkins (1614–1672) anonymously published one of the earliest books on cryptography, *Mercury or the Secret and Swift Messenger*. This book not only pointed out the fundamental role of differences but noted that any (finite) set of different things could be encoded by words in a binary code.

> For in the general we must note, That whatever is capable of a competent Difference, perceptible to any Sense, may be a sufficient Means whereby to express the Cogitations. It is more convenient, indeed, that these Differences should be of as great Variety as the Letters of the Alphabet; but it is sufficient if they be but twofold, because Two alone may, with somewhat more Labour and Time, be well enough contrived to express all the rest. ([11], Chap. XVII, p. 69)

Wilkins explains that a five letter binary code would be sufficient to code the letters of the alphabet since $2^5 = 32$:

> Thus any two Letters or Numbers, suppose *A.B.* being transposed through five Places, will yield Thirty Two Differences, and so consequently will superabundantly serve for the Four and twenty Letters... .([11], Chap. XVII, p. 69)

In James Gleick's 2011 book, *The Information: A History, A Theory, A Flood*, he noted that:

> Any difference meant a binary choice. Any binary choice began the expressing of cogitations. Here, in this arcane and anonymous treatise of 1641, the essential idea of information theory poked to the surface of human thought, saw its shadow, and disappeared again for [three] hundred years. ([12], p. 161)[3]

The idea that information is about differences was also expressed by the polymath, Gregory Bateson, who noted that (the transmission of) "[i]nformation consists of differences that make a difference." ([13], p. 99)

The formula that is a measure of differences, $h(p) = 1 - \sum_j p_j^2$ (or its complementary form $1 - h(p) = \sum_j p_j^2$), goes back at least to Corrado Gini (1884–1965) who published it as an *index of mutability* [14] in 1912 (not to be confused with Gini's better-known index of inequality). Some of the immediate following history of the formula was connected to cryptology as foreshadowed by Wilkins. William F. Friedman, an American cryptologist, devoted a 1922 book [15] to the "index of

[3] Gleick is referring to the old Pennsylvania Dutch superstition that on February 2 each year, if a groundhog emerges from its den and sees its shadow, then it will go back in for six more weeks.

coincidence" (i.e., $\sum p_i^2$). Solomon Kullback worked as an assistant to Friedman and wrote a book on cryptology which used the index [16].

During World War II, Alan M. Turing worked for a time in the Government Code and Cypher School at the Bletchley Park facility in England. Probably unaware of the earlier work, Turing used $\rho = \sum p_i^2$ in his cryptoanalysis work and called it the *repeat rate* since it is the probability of a repeat in a pair of independent draws from a population with those probabilities. Polish cryptographers had independently used the repeat rate in their work on the Enigma [17]. After WWII, Edward H. Simpson, a British statistician, proposed $\sum_{B\in\pi} p_B^2$ as a measure of species concentration (the opposite of diversity) where π is the partition of animals or plants according to species and where each animal or plant is considered as equiprobable. And Simpson gave the interpretation of this homogeneity measure as "the probability that two individuals chosen at random and independently from the population will be found to belong to the same group." ([18], p. 688) Hence $1 - \sum_{B\in\pi} p_B^2$ is the probability that a random ordered pair will belong to different species, i.e., will be distinguished by the species partition. The biodiversity literature [19] refers to the formula as " Simpson's index of diversity" or sometimes, the "Gini-Simpson diversity index." In the bioinformatics literature, Masatoshi Nei [20] introduced the logical entropy formula as a measure of gene diversity.

But the Simpson story has a twist. Simpson along with I.J. Good worked at Bletchley Park during WWII, and, according to Good, "E.H. Simpson and I both obtained the notion [the repeat rate] from Turing." ([21], p. 395) When Simpson published the index in 1948, he (again, according to Good) did not acknowledge Turing " fearing that to acknowledge him would be regarded as a breach of security." ([22], p. 562) Perhaps logical entropy should be called "Turing entropy" to compete with the "big names" attached to Shannon entropy and von Neumann entropy. But given its frequent discovery and rediscovery, Good also negated that idea.

> If $p_1, p_2, \ldots, p_n$ are the probabilities of mutually exclusive and exhaustive events, any statistician of this century who wanted a measure of homogeneity would have taken about two seconds to suggest $\sum p_i^2$, which I shall call ρ. ... Thus it is unjust to associate ρ with any one person. It would be better to use such names as "repeat rate" or "quadratic index of homogeneity" for ρ and perhaps "quadratic index of heterogeneity or diversity" for $1 - \rho$. ([22], pp. 561–2]

Thus the name "logical entropy" seems appropriate, particularly in view of Stigler's Law of Eponymy, i.e., "No scientific discovery is named after its original discoverer" [23], and since it is the quantitative measure associated with partitions in the logic of partitions just as finite probability is the quantitative measure associated with subsets in the usual Boolean logic of subsets.

In economics, Albert O. Hirschman ([24], p. 159) suggested in 1945 using $\sqrt{\sum p_i^2}$ as an index of trade concentration (where p_i is the relative share of trade in a certain commodity or with a certain partner). A few years afterwards, Orris Herfindahl [25] independently suggested using $\sum p_i^2$ as an index of industrial concentration (where p_i is the relative share of the ith firm in an industry). In the literature on industrial economics, the index $H = \sum p_i^2$ is variously called the Hirschman–Herfindahl index, the *HH* index, or just the H index of concentration.

Another way to look at logical entropy is that two elements from $U = \{u_1, \ldots, u_n\}$ are either identical or distinct. Gini [14] introduced $d_{ij} = 1 - \delta_{ij}$ (the complement of the Kronecker delta function) as the "distance" between the ith and jth elements where $d_{ij} = 1$ for $i \neq j$ and $d_{ii} = 0$. Then Gini's index of mutability, $h(p) = \sum_{i,j} d_{ij} p_i p_j$, is the average (logical) distance between a pair of independently drawn elements. But one might generalize by allowing other non-negative distances $d_{ij} = d_{ji}$ for $i \neq j$ (but always $d_{ii} = 0$) so that $Q = \sum_{i,j} d_{ij} p_i p_j$ would be the average distance between a pair of independently drawn elements from U. In 1982, C.R. (Calyampudi Radhakrishna) Rao introduced precisely this concept as *quadratic entropy* [26]. The logical entropy is also the quadratic special case of the Tsallis–Havrda–Charvat entropy [27, 28].

Časlav Brukner and Anton Zeilinger have also developed the logical entropy formula $1 - \sum_{i=1}^{n} p_i^2$ in their treatment of quantum information [29, 30] and have also used the normalized form of the (Euclidean) distance squared of a probability distribution from the uniform distribution, which is closely related to the logical entropy since: $\sum_{i=1}^{n} \left(p_i - \frac{1}{n}\right)^2 = \left(1 - \frac{1}{n}\right) - h(p)$.

Compound notions of logical entropy

We now consider a joint probability distribution $\{p(x, y)\}$ on the finite sample space $X \times Y$ (where to avoid trivialities, assume $|X|, |Y| \geq 2$), with the marginal distributions $\{p(x)\}$ and $\{p(y)\}$ where $p(x) = \sum_{y\in Y} p(x, y)$ and $p(y) = \sum_{x\in X} p(x, y)$. The setting is a pair of random variables X and Y where we also consider X as the set of possible values x of the r.v. X and similarly for the r.v. Y. Then the joint probability distribution is $p(x, y) = \Pr(X = x, Y = y)$, and the marginals are $p(x) = \Pr(X = x)$, and $p(y) = \Pr(Y = y)$. For notational simplicity, the entropies can be considered as functions of the random variables or of their probability distributions, e.g., $h(\{p(x)\}) = h(X)$ and $h(\{p(y)\}) = h(Y)$. Logical entropy is characterized in terms of the probability that in two independent draws (x, y) and (x', y') from the sample space, one will get different outcomes. Hence in this case,

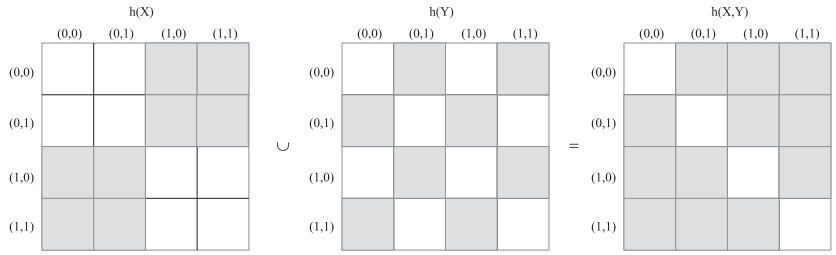

Figure 5. Union of Box/Venn diagrams for $h(X)$ and $h(Y)$ gives the box diagram for joint entropy $h(X, Y) = \frac{12}{16} = \frac{3}{4}$.

$$h(X) = \sum_{x,y} \{p(x,y)p(x',y') : x \neq x'\}, \tag{8}$$

$$h(Y) = \sum_{x,y} \{p(x,y)p(x',y') : y \neq y'\}. \tag{9}$$

Then the joint entropy $h(X, Y)$ is just the logical entropy $h(\{p(x, y)\}_{((x, y) \in X \times Y)})$ of the joint probability distribution which can also be characterized as:

$$h(X,Y) = \sum_{x,y} \{p(x,y)p(x',y') : x \neq x' \text{ or } y \neq y'\}. \tag{10}$$

In the previous even-odd dice example of throwing an X-die and a Y-die, each die had an outcome set of $\{0, 1\}$ so $X \times Y = \{(0, 0), (0, 1), (1, 0), (1, 1)\} = U$. The space on which the probabilities are assigned is $U \times U = (X \times Y) \times (X \times Y)$ so the probability assigned to each point $((x, y),(x', y'))$ is $p(x, y)p(x', y')$. The points in the space $(X \times Y)^2$ whose probabilities add up to give $h(X, Y)$ are just the union of the points for $h(X)$, i.e., where $x \neq x'$, and for $h(Y)$, i.e., where $y \neq y'$. Since each point in $(X \times Y)^2$ is represented by a square with probability $\frac{1}{16}$, the shaded squares for $h(X, Y)$ are just the union of the squares for $h(X)$ and $h(Y)$ as shown in Figure 5.

The usual interpretation carries over to the compound notions such as the joint entropy; in two independent throws of the pair of dice, the probability that one will get a different parity in the X-die *or* in the Y-die (or both) is $h(X, Y) = \frac{3}{4}$.

In a Venn diagram that is merely illustrative, the logical entropies would be represented as circles and the union of the circles would represent the joint entropy as in Figure 6.

Figure 6 also illustrates the "formulas" for the other compound logical entropies. The *conditional logical entropy*

$$h(X|Y) = \sum_{x,y} \{p(x,y)p(x',y') : x \neq x' \quad \text{and} \quad y = y'\} \tag{11}$$

represents the distinctions made by X (i.e., the cases where the two throws of X-die had different parities) after the distinctions made by Y are taken away (so $y = y'$), and vice-versa for $h(Y|X)$. And the *mutual logical information*

$$m(X,Y) = \sum_{x,y} \{p(x,y)p(x',y') : x \neq x' \quad \text{and} \quad y \neq y'\} \tag{12}$$

is the probability that in the two throws of the pair of dice, the pair of pairs of outcomes will have different parity in the X-die *and* in the Y-die – as one can easily see from the shaded squares for $m(X, Y)$ in Figure 7.

These specific box/Venn diagrams illustrate general relationships such as the two conditional entropies and mutual information all being disjoint and adding to the joint entropy. In general (not just for this example), the compound logical entropies stand in the relationships shown by the areas in the illustrative Figure 6:

$$h(X,Y) = h(X|Y) + h(Y|X) + m(X,Y), \tag{13}$$

$$h(X) = h(X|Y) + m(X,Y), \tag{14}$$

$$h(Y) = h(Y|X) + m(X,Y) \tag{15}$$

$$h(X,Y) = h(X) + h(Y) - m(X,Y). \tag{16}$$

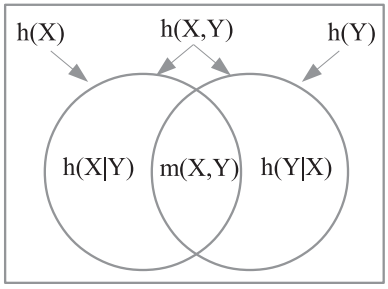

Figure 6. Illustrative Venn diagram for the compound logical entropies.

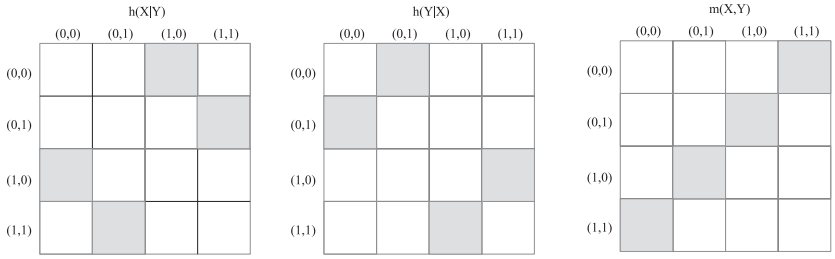

Figure 7. Box diagrams representing the two conditional logical entropies and the mutual logical information all with the value $\frac{1}{4}$.

Shannon entropy

The basic definitions

Both the logical and the Shannon entropies are defined for probability distributions regardless of whether the distribution is derived from the blocks of a partition $\Pr(B_i)$ or the values of a random variable $\Pr(X = x)$. Hence we can start the treatment of *Shannon entropy* [31, 32] defined on a probability distribution $p = (p_1, \ldots, p_n)$:

$$H(p) = -\sum_{i=1}^{n} p_i \log_2(p_i) = \sum_{i=1}^{n} p_i \log_2\left(\frac{1}{p_i}\right) \tag{17}$$

where for $p_i = 0$, $p_i \log_2\left(\frac{1}{p_i}\right)$ is defined to be 0. Henceforth, the logs are to base 2 unless otherwise specified. For a random variable X with $p(x) = \Pr(X = x)$, then:

$$H(X) = \sum_{x \in X} p(x) \log\left(\frac{1}{p(x)}\right). \tag{18}$$

Given a joint probability distribution $p(x, y)$ on $X \times Y$, the *joint Shannon entropy* is:

$$H(X, Y) = \sum_{(x,y) \in X \times Y} p(x, y) \log\left(\frac{1}{p(x, y)}\right). \tag{19}$$

The conditional Shannon entropy $H(X|Y)$ is defined as the average of the Shannon entropies for conditional probability distributions. Given the joint distribution $\{p(x, y)\}$ on $X \times Y$, then for a specific $y_0 \in Y$, then the conditional probability distribution is $p(x|y_0) = \frac{p(x,y_0)}{p(y_0)}$ which has the Shannon entropy: $H(X|y_0) = \sum_{x \in X} p(x|y_0) \log\left(\frac{1}{p(x|y_0)}\right)$. Then the *Shannon conditional entropy* is defined as the *average* of these entropies:

$$H(X|Y) = \sum_{y \in Y} p(y) \sum_{x} \frac{p(x, y)}{p(y)} \log\left(\frac{p(y)}{p(x, y)}\right) = \sum_{x,y} p(x, y) \log\left(\frac{p(y)}{p(x, y)}\right). \tag{20}$$

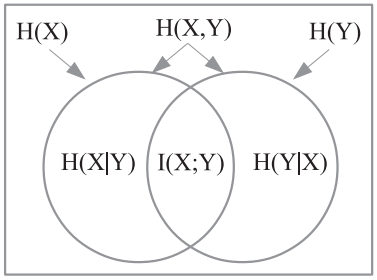

Figure 8. Venn diagram mnemonic for the compound Shannon entropies.

Since the Venn diagram for any measure like logical entropy satisfies a relationship like $h(X) + h(Y) - h(X,Y) = m(X,Y)$, Shannon defined the *mutual Shannon information* as:

$$I(X;Y) = \sum_{x \in X, y \in Y} p(x,y) \left[\log\left(\frac{1}{p(x)}\right) + \log\left(\frac{1}{p(y)}\right) - \log\left(\frac{1}{p(x,y)}\right) \right]. \tag{21}$$

Then it is perhaps no surprise that these compound Shannon entropies satisfy the Venn diagram relationship *as if* the Shannon entropy was defined as a measure on a set. Hence one finds in the textbooks on Shannon's theory of communications, a Venn diagram like Figure 8 to serve at least as a mnemonic about the interrelationships.

Shannon's communications theory and "information theory"

This paper presents a different version of "information theory" than the received version. There is no difference in the part of information theory where Shannon entropy actually does its work, namely the theory of coding and communication. Shannon himself did not name his original paper or book as "information theory" but rather as the "mathematical theory of communication" ([31, 32]). Thus the notion that the theory of communications (including coding theory) was an "information theory" was a creation of the science press, science popularizers, and textbook writers. Shannon even reacted against the "bandwagon" that inflated "information theory" far beyond the actual technical results of communications theory.

> Information theory has, in the last few years, become something of a scientific bandwagon. Starting as a technical tool for the communication engineer, it has received an extraordinary amount of publicity in the popular as well as the scientific press. In part, this has been due to connections with such fashionable fields as computing machines, cybernetics, and automation; and in part, to the novelty of its subject matter. As a consequence, it has perhaps been ballooned to an importance beyond its actual accomplishments. Our fellow scientists in many different fields, attracted by the fanfare and by the new avenues opened to scientific analysis, are using these ideas in their own problems. Applications are being made to biology, psychology, linguistics, fundamental physics, economics, the theory of organization, and many others. In short, information theory is currently partaking of a somewhat heady draught of general popularity. ([33], p. 462)

Shannon repeated the points in a 1961 interview with Myron Tribus.

> In 1961 Professor Shannon, in a private conversation, made it quite clear to me that he considered applications of his work to problems outside of communication theory to be suspect and he did not attach fundamental significance to them. ([34], p. 1)

Moreover, while Shannon noted that while his entropy formula indicates the "amount of information" (i.e., the average numbers of binary distinctions needed to distinguish all the "messages"), "no concept of information itself was defined" ([35], p. 458) in communications theory. Perhaps the most common idea about Shannon entropy is that it a measure of "amount of uncertainty." But there are many other interpretations.

> Other terms used to convey an intuitive feeling for entropy include randomness, disorganization, " mixed-up-ness" (Gibbs), missing information, in-complete knowledge, complexity, chaos, ignorance, and uncertainty. ([36], p. 9)

There is also the view that entropy and information were in fact opposites or complements; "Gain in entropy always means loss of information, and nothing more." ([37], p. 573) That view was later popularized by Leon Brillouin who claimed that:

> information must be considered as a negative term in the entropy of a system; in short, information is negentropy. ... Entropy measures the lack of information. ([38], p. xii)

However, there is no need for this conceptual chaos; the (simple) Shannon entropy is another way to quantify the notion of information-as-distinctions. That is, Shannon entropy is the minimum average number of binary partitions (bits) that need to be joined in order to make the distinctions that distinguish all the "messages." And simple logical entropy is the direct measure of distinctions.

Is Shannon entropy a "measure"?

Shannon entropy and a host of other entropy "formulas" (sans interpretation) are routinely called "measures" of information [39]. A prominent information theorist, Lorne Campbell, has noted in 1965 the analogies between Shannon entropy and measures (in the usual non-negative sense).

> Certain analogies between entropy and measure have been noted by various authors. These analogies provide a convenient mnemonic for the various relations between entropy, conditional entropy, joint entropy, and mutual information. It is interesting to speculate whether these analogies have a deeper foundation. It would seem to be quite significant if entropy did admit an interpretation as the measure of some set. ([40], p. 112)

We only need be concerned with the simplest case of a measure [9] on a finite set where for any finite set U, a *measure* μ is a function from the powerset of U (the subsets of U) to the reals μ: $\wp(\mathrm{U}) \to \mathbb{R}$ such that:

1. $\mu(\emptyset) = 0$,
2. for any $E \subseteq U$, $\mu(E) \geq 0$, and
3. for any disjoint subsets E_1 and E_2, $\mu(E_1 \cup E_2) = \mu(E_1) + \mu(E_2)$.

The whole measure is determined by the values on singletons and simply summed over larger finite subsets.

It would be desirable for Shannon entropy to be a measure in this technical sense so:

> that $H(\alpha)$ and $H(\beta)$ are measures of sets, that $H(\alpha, \beta)$ is the measure of their union, that $I(\alpha, \beta)$ is the measure of their intersection, and that $H(\alpha|\beta)$ is the measure of their difference. The possibility that $I(\alpha, \beta)$ is the entropy of the " intersection" of two partitions is particularly interesting. This " intersection," if it existed, would presumably contain the information common to the partitions α and β. ([40], p. 113)

Logical entropy satisfies all those desiderata.

There are some differences in the use of the word "measure." It would seem that the usual notion of a measure is always non-negative [9,41] and then there is an extended notion of a "signed measure" that can take on negative values. Other authors define a "measure" to allow negative values and then define a "positive measure" to have only non-negative values. The most general usage, adopted here, is that a measure is non-negative and the generalized notion to allow negative value is a "signed measure." This is important since logical entropy is defined as a measure, indeed a probability measure, while Shannon entropy is not defined as a measure on a set. Given any Venn diagram of Shannon entropies, then, as with any Venn diagram, an *ex post* measue or signed measure can always be trivially constructed. Both measures and signed measures can be represented as additive set functions ([42], Part 8, Chap. 1, Prob. 26) [43] ([44], Chap. 2) [45] that satisfy the inclusion-exclusion principle (or overcount-undercount relationships) that can be associated with Venn diagrams (if we allow negative areas).

For logical entropy, consider a set $U = \{u_1, \ldots, u_n\}$ with point probabilities $\{p_i\}_{i=1}^n$ and a random variable $X : U \to \mathbb{R}$ which induces a partition X^{-1} on U and similarly for $Y : U \to \mathbb{R}$. The set on which the logical entropy measure is defined is $U \times U$ and the value assigned to a point $(u_j, u_k) \in U \times U$ is $\mu(\{(u_j, u_k)\}) = p_j\, p_k$. The logical entropy associated with the random variable is:

$$h(X) = \mu\big(\mathrm{dit}\big(X^{-1}\big)\big) = \sum_{u_j, u_k \in U} \big\{p_j p_k : \big(u_j, u_k\big) \in \mathrm{dit}\big(X^{-1}\big)\big\}, \tag{22}$$

namely the sum of all the products $p_j\, p_k$ for which $X(u_j) \neq X(u_k)$. Thus it is interpreted as the probability that on two independent trials, the random variable X will give different values. That illustrates how logical entropy measures differences. If the values of X have no differences, i.e., if it is constant, then $X^{-1} = \mathbf{0}_U$ and $h(\mathbf{0}_U) = 0$. The more

refined the inverse-image partition X^{-1}, the higher the logical entropy. Then all the usual Venn diagram relationships hold such as

$$\begin{aligned} h(X) &= \sum \{p_j p_k : (u_j, u_k) \in \mathrm{dit}(X^{-1})\} \\ &= \sum_{u_j,u_k \in U} \{p_j p_k : (u_j, u_k) \in \mathrm{dit}(X^{-1}) \text{ and } (u_j, u_k) \notin \mathrm{dit}(Y^{-1})\} \\ &+ \sum_{u_j,u_k \in U} \{p_j p_k : (u_j, u_k) \in \mathrm{dit}(X^{-1}) \text{ and } (u_j, u_k) \in \mathrm{dit}(Y^{-1})\} \\ &= h(X|Y) + m(X, Y) \end{aligned} \tag{23}$$

and all of Campbell's desiderata are satisfied. For instance, the logical conditional entropy is the measure on the difference of the sets for $h(X)$ and $h(Y)$:

$$h(X|Y) = \sum_{u_j,u_k \in U} \{p_j p_k : (u_j, u_k) \in \mathrm{dit}(X^{-1}) - \mathrm{dit}(Y^{-1})\}. \tag{24}$$

To see why Shannon entropy is not in general a (non-negative) measure, consider the previous even-odd dice example of two random variables $X, Y : U \to 2$ for $U = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ where X was the parity of the first die thrown and Y the parity of a second die thrown. Each point $(x, y) \in U = X \times Y$ has probability $p(x, y) = \frac{1}{4}$ and marginal distributions have $p(x) = \frac{1}{2} = p(y)$. A two-variable joint distribution $p(x, y)$ has the *independence* property if $p(x, y) = p(x)p(y)$ for all $(x, y) \in U$. Hence the two r.v.s X and Y are independent. One of the original "selling points" of Shannon entropy was that for independent r.v.s, $H(X, Y) = H(X) + H(Y)$, i.e., independent r.v.s have "no information in common" so that $I(X, Y) = 0$. It might be noted that having an overlap of $H(X)$ and $H(Y)$ of 0 is not the same as the Venn diagrams for $H(X)$ and $H(Y)$ not overlapping.

Consider a third random variable $Z : U \to 2$ whose value is the parity of the sum $X + Y$ so $Z((0, 0)) = Z((1, 1)) = 0$ and $Z((0, 1)) = Z((1, 0)) = 1$. Then $\Pr(Z = 0) = \Pr(Z = 1) = \frac{1}{2} = p(z)$ and $p(x, z) = p(x)p(z)$ for all $x, z \in \{0, 1\} = 2$ so X and Z are also independent and similarly for Y and Z. Thus the three variables are pair-wise independent but they are not mutually independent for the simple reason that if you know the values of any two of them, you know the value of the third variable. Hence in the Venn diagram for the Shannon entropies of X, Y, and Z, each pair of areas must have zero overlap but the three areas must intersect in non-zero overlap. The only way this can happen is for the three-way overlap to be negative and the two-way overlaps be the sum of that negative triple overlap and the equal positive remaining two-way overlap so all the two-way overlaps are zero as shown in Figure 9.

Thus the intuitively satisfactory idea of the two-way overlaps for independent variables being zero ("no information in common") leads to the interpretive "problem" of three-way mutual information being possibly negative. Shannon dealt with this problem in the simplest possible way; he never defined mutual information for more than two variables. Or, as perhaps the most definitive monograph on information theory casually put it; "There isn't really a notion of mutual information common to three random variables." ([46], p. 49) But the three-way definition is automatically given by the usual inclusion-exclusion formulas that hold for measures and signed measures. For two variables, $H(X, Y) = H(X) + H(Y) - I(X; Y)$, and for three variables, it is:

$$H(X, Y, Z) = H(X) + H(Y) + H(Z) - I(X; Y) - I(X; Z) - I(Y; Z) + I(X; Y; Z), \tag{25}$$

where Shannon defined all the terms in the equation except the last one $I(X; Y; Z)$ which is thus determined. The Shannon entropy for each variable X, Y, and Z, is:

$$H(X) = H(Y) = H(Z) = p(0) \log (1/p(0)) + p(1) \log (1/p(1)) = \frac{1}{2} \log (2) + \frac{1}{2} \log (2) = 1. \tag{26}$$

And all the two-way overlaps have the values:

$$I(X; Y) = I(X; Z) = I(Y; Z) = 0. \tag{27}$$

The three-way joint entropy is the Shannon entropy of the probability distribution $p(x, y, z)$ which is easily computed in Table 1. Since the values of any two variables determine the third, the probabilities are either $\frac{1}{4}$ if the third value agrees with the values of the other two or 0 otherwise.

The sum of the last column gives the three-way joint Shannon entropy of $H(X, Y, Z) = 2$. Hence the inclusion-exclusion formula gives:

$$\begin{aligned} I(X; Y; Z) &= H(X, Y, Z) - H(X) - H(Y) - H(Z) + I(X; Y) + I(X; Z) + I(Y; Z) \\ &= 2 - 1 - 1 - 1 + 0 + 0 + 0 = -1. \end{aligned} \tag{28}$$

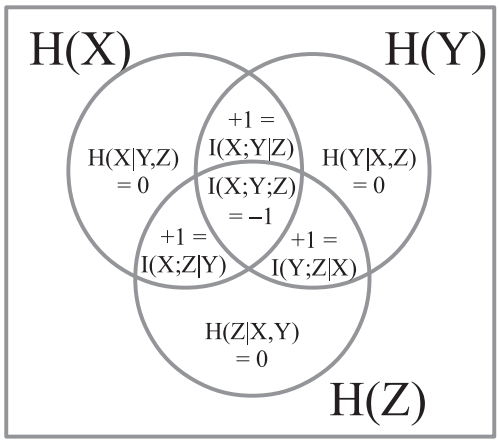

Figure 9. Venn diagram for three-way negative Shannon mutual information $I(X;\ Y;\ Z)$.

Table 1. Probability distribution $p(x,\ y,\ z)$ and computation of $H(X,\ Y,\ Z)$.

X	Y	Z	$p(x,\ y,\ z)$	$p(x,\ y,\ z)\log(1/p(x,\ y,\ z))$
0	0	0	1/4	$1/4 \times 2 = 1/2$
0	0	1	0	0
0	1	0	0	0
0	1	1	1/4	1/2
1	0	0	0	0
1	0	1	1/4	1/2
1	1	0	0	0
1	1	1	1/4	1/2

Thinking in term of underlying points, the three-way overlap has points that are common to $H(X)$, $H(Y)$, and $H(Z)$, so some of the points must have negative values. *Thus all three, $H(X)$, $H(Y)$, and $H(Z)$, cannot be the value of a (non-negative) measure on some set.* Moreover, the intuitive appeal of $I(X;\ Y) = 0$ as meaning "no information in common" for independent variables is lessened when it turns out to mean not disjoint or non-overlapping areas but that the positive information in each of the three two-way overlap of these independent random variables must be balanced by "negative information" in the three-way overlap, which, as Csiszar and Kröner remark, has "no natural intuitive meaning." ([47], p. 53)

Finally, we might consider how this example is treated by logical entropy. The r.v. Z has a logical entropy $h(Z)$ as the sum of the shaded $\frac{1}{16}$ squares in Figure 10.

The two-way mutual logical information, say for $m(X,\ Y)$, is given by the shaded squares that are in common, i.e., the two-way overlap of $h(X)$ and $h(Y)$, and the three-way mutual logical entropy $m(X,\ Y,\ Z)$ is given by the shaded squares in common to all three. But since logical entropy is a measure (in the usual non-negative sense), we can compute the three-way mutual information by using the undercount-overcount formula:

$$m(X,Y,Z) = h(X,Y,Z) - h(X) - h(Y) - h(Z) + m(X,Y) + m(X,Z) + m(Y,Z). \tag{29}$$

The three-way joint logical entropy includes all squares except the diagonal so its value is $\frac{12}{16} = \frac{3}{4}$. The single logical entropies are all $\frac{1}{2}$ and the two-way mutual informations are all $\frac{4}{16} = \frac{1}{4}$ so the formula yields:

$$m(X,Y,Z) = \frac{3}{4} - \frac{1}{2} - \frac{1}{2} - \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{6}{4} - \frac{3}{2} = 0. \tag{30}$$

The two-way logical mutual informations for independent variables are not zero since logical entropy is a probability distribution – so for independent variables, it is multiplicative, e.g., $m(X,\ Y) = h(X)h(Y)$. The calculation of the three-way mutual logical information can be intuitively checked by considering the three areas for $h(X)$, $h(Y)$, and $h(Z)$ in Figure 11.

It can then be checked by inspection that there is no shaded square common to all three diagrams so the three-way overlap is zero.

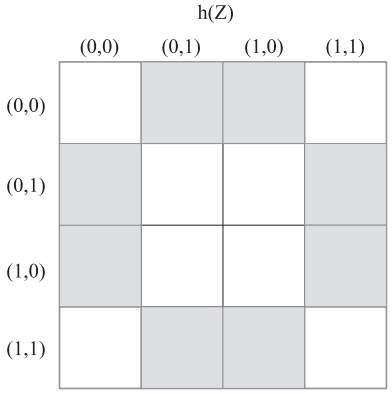

Figure 10. Venn diagram for logical entropy $h(Z) = \frac{8}{16} = \frac{1}{2}$.

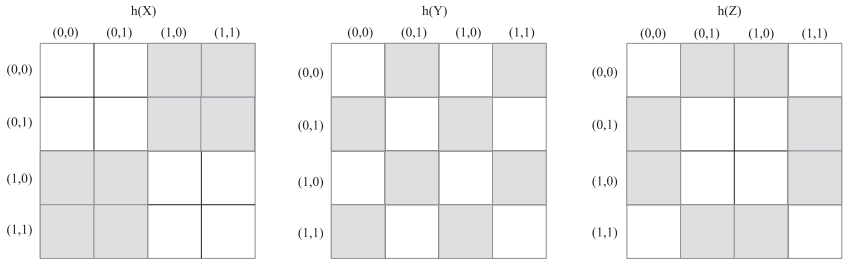

Figure 11. The box diagrams for $h(X)$, $h(Y)$, and $h(Z)$.

It is interesting to note that all *two*-way mutual logical informations such as $m(X, Y)$, $m(X, Z)$, and $m(Y, Z)$, are in general never zero when all point probabilities are positive. This is the result of the Common-Dits Theorem that any *two* non-empty ditsets have a non-empty intersection.[4]

Theorem 3.1. *Common Dits. Any two non-empty ditsets intersect, i.e., have some dits in common.*

Proof. A ditset $\text{dit}(\pi) = \emptyset$ iff $\pi = \mathbf{0}_U$, the indiscrete partition or blob. Consider any two non-empty ditsets $\text{dit}(\pi)$ and $\text{dit}(\sigma)$. Since π is not the blob $\mathbf{0}_U$, consider two elements u and u' distinguished by π but identified by σ; otherwise $(u, u') \in \text{dit}(\pi) \cap \text{dit}(\sigma)$ and we are finished. Since σ is also not the blob, there must be a third element u'' not in the same block of σ as u and u', as shown in Figure 12.

But since u and u' are in different blocks of π, the third element u'' must be distinguished from one or the other or both in π, e.g., distinguished from u' in both partitions as in Figure 12. Hence (u, u'') or (u', u'') must be distinguished by both partitions and thus must be in $\text{dit}(\pi) \cap \text{dit}(\sigma)$. □

The three non-trivial partitions on a three-element set show that there are no common dits to all three of them (as in the dice example), only to each pair of partitions.

The connection between the logical and Shannon entropies

One question lingers. If, as we have seen, Shannon entropy is not defined as a measure in the usual non-negative sense, then what accounts for the compound Shannon entropies satisfying the Venn diagram relationships? As one author surmised: "Shannon carefully contrived for this 'accident' to occur" ([49], p. 153), and Campbell asked "whether these analogies have a deeper foundation." ([40], p. 112) Since Shannon arranged or "contrived" for the compound entropies to satisfy the Venn diagram relationships for two random variables, they can be extended to any number of variables using the inclusion-exclusion formulas [50, 51]. As we have seen, mutual information can be negative for three or more variables.

[4] This is a restatement of the graph-theoretic result that the complement of any disconnected graph is connected ([48], p. 30). In terms of inditsets or equivalence relations E and E', if $E \cup E' = U \times U$, then $E = U \times U$ or $E' = U \times U$.

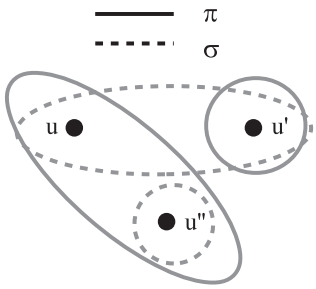

Figure 12. Solid circles = blocks of π, dashed circles = blocks of σ, and (u', u'') as a common dit to π and σ.

But there is an interesting twist to the story. Information theorists do not *define* Shannon entropy as a signed measure on a given set. But such a set can be trivially constructed *ex post* in the manner shown by Hu [43] and Yeung [52] but the underlying mathematical fact about additive set functions goes back at least to the 1925 first edition of Polya-Szego's book [42]. In the Venn diagram showing all possible overlaps of three "circles" for three random variables, there are $2^3 = 8$ atomic areas (2^n in the general finite case of n random variables), each of which can trivially be taken as a single element in a set. Then numbers (positive or negative) can be assigned arbitrarily to those points and then summed to get the values attached to the circles. Then all the Venn diagram relationships are automatically satisfied. Since these sets are constructed in terms of the independently defined Shannon entropies, the set and value assignments may change when more variables come into play. In the even–odd dice example, as long as only X and Y are considered, then all the compound Shannon entropies are non-negative and a set with a (non-negative) measure on it can be constructed to yield the values of $H(X)$ and $H(Y)$. But when the variable Z is brought into consideration, then the underlying set must be reconstructed to have a negative-valued point representing $I(X; Y; Z)$ so that the *signed* measure on that set will give the values of all the compound Shannon entropies.[5] This serves to underline the fact that Shannon entropy is not *defined* as a measure on a set in the first place.

In contrast, the logical entropy defines the set beforehand, namely $U \times U$, and the values assigned to the points is determined beforehand, namely $p_i\, p_j$ is assigned to (u_i, u_j), and then the simple and compound logical entropies are defined by collections of those points and their values. Nothing changes when new random variables are considered; it just means considering a different set of points. Thus it is not a simple matter of saying logical entropy is a (non-negative) measure and Shannon entropy is a signed measure. Logical entropy is defined as a probability measure on a set given beforehand, and the Shannon entropies are only a signed measure on a set *ex post* constructed for the purpose after all the numerical values are independently given in the Venn diagram formulas for a given set of random variables.

There is, however, a deeper connection between the two entropies since there is a transform from all the compound logical entropy formulas to the corresponding compound Shannon entropy formulas that preserves the Venn diagram relationships. To understand this transform, consider the canonical examples of 2^m equiprobable points in U. For any m such as $m = 3$, the binary partitions required to distinguish the $2^3 = 8$ "messages" (or leaves) can be pictured in the (upside–down) binary tree of Figure 13.

It was previously asserted that logical entropy and Shannon entropy are two different ways to quantify the definition of information-as-distinctions. Logical entropy is the direct (normalized) count of the distinctions or dits in a partition and Shannon entropy is the minimum average number of binary partitions that need to be joined together to make *the same distinctions*.

This connection can be easily demonstrated using Figure 13 example. Let $\pi = \{\{u_1\}, \ldots, \{u_8\}\}$ be the discrete partition on the equi-probable outcomes (messages or leaves in the tree) in $U = \{u_1, \ldots, u_8\}$. The number of distinctions $|\mathrm{dit}(\pi)|$ is $|U \times U - \Delta| = 64 - 8 = 56$ (where Δ is the diagonal of self-pairs $\{(u_1, u_1), \ldots, (u_8, u_8)\}$), and the logical entropy $h(\pi)$ of π is $\frac{|\mathrm{dit}(\pi)|}{|U \times U|} = \frac{56}{64} = \frac{7}{8} = 1 - \frac{1}{8}$, the probability that in two independent draws from U, different elements of U are obtained, i.e., the probability $1 - \frac{1}{8}$ that the second draw isn't the same as the first draw. Since the outcomes u_i are the leaves of the tree in Figure 13, one could image a marble rolling down from the root (like on a Galton board) and then going one way or the other with equal probability at each branching. The logical entropy is the probability that two such marbles will end up in different leaves.

We need to show that the Shannon entropy of π is the minimum number of binary partitions (corresponding to yes-or-no questions in the game of 20-questions) necessary to make all the same distinctions of π. Recall that given π and σ partitions

[5] The construction is easy; take the seven atomic areas inside the three circles in Figure 9 as containing one point having the value assigned to that atomic area. The eighth area outside the three circles can have an arbitrary value – at least until another random variable appears.

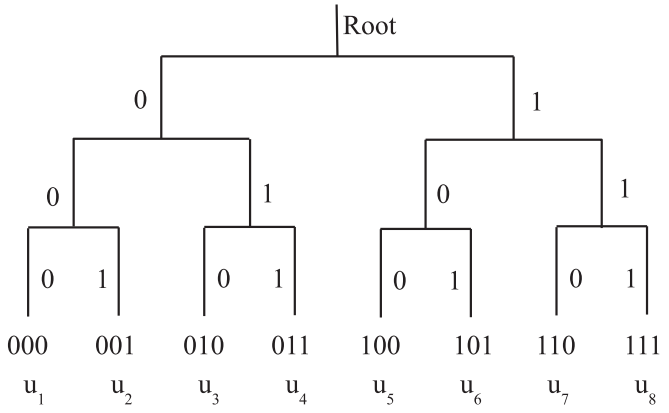

Figure 13. Three equiprobable binary partitions distinguish the $2^3 = 8$ leaves on the tree.

on U, the *join* $\pi \vee \sigma$ is the partition on U whose blocks are all the non-empty intersections $B \cap C$ for $B \in \pi$ and $C \in \sigma$. Since $\mathrm{dit}(\pi \vee \sigma) = \mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma)$, the join of partitions accumulates the distinctions made by each of the partitions. Gian-Carlo Rota formulated the problem as the Devil selecting a particular u_i or message and not revealing it to the questioner but having to answer any yes-or-no question truthfully. Or less colorfully, "To determine an object, we need to ensure that the responses to the sequence of questions uniquely identifies the object from the set of possible objects" ([46], p. 120). But the problems of 1) making all the distinctions, and 2) uniquely determining any given outcome or message, are equivalent. If the binary partitions or binary questioning does not distinguish u_i from u_j, then it would not determine the hidden message if it happened to be u_i or u_j – and if the questioning cannot determine the message if it was u_i or u_j, then the corresponding binary partitions do not distinguish u_i and u_j. This means that the usual Shannon interpretation about the minimum average number of binary questions necessary to uniquely determine the message is also the minimum average number of binary partitions necessary to make all the distinctions between messages.

This can be illustrated with the example of Figure 13. The first binary partition π^1 corresponds to the first branching point in Figure 13 and the first binary digit in the codes (reading from left to right in the code words):

$$\pi^1 = \{\{u_1, ..., u_4\}, \{u_5, ..., u_8\}\}, \tag{31}$$

where $u_1, \ldots, u_4$ have 0 as the first digit and $u_5, \ldots, u_8$ have 1 as the first digit. The binary partition π^1 corresponds to the yes-or-no question, "Is the first letter in the code for u_i a 0?". The partition has 16 distinctions from $\{u_1, \ldots, u_4\} \times \{u_5, \ldots, u_8\}$ and another 16 from the reverse ordering for a total of 32 distinctions.
The second binary partition π^2 in effect asks about the second digit in the codes for the u_i, and it is:

$$\pi^2 = \{\{u_1, u_2, u_5, u_6\}, \{u_3, u_4, u_7, u_8\}\} \tag{32}$$

so the join is:

$$\pi^1 \vee \pi^2 = \{\{u_1, u_2\}, \{u_3, u_4\}, \{u_5, u_6\}, \{u_7, u_8\}\}. \tag{33}$$

Comparing $\pi^1 \vee \pi^2$ to π^1, we see the splitting $\{u_1, \ldots, u_4\}$ into $\{u_1, u_2\}$ and $\{u_3, u_4\}$ so that creates the new distinctions $|\{u_1, u_2\} \times \{u_3, u_4\}| \times 2 = 8$ and similarly for $\{u_5, u_6\}$ and $\{u_7, u_8\}$, so π^2 makes $8 + 8 = 16$ new distinctions for $32 + 16 = 48$ distinctions. Equivalently, one could compute the distinctions of $\pi^1 \vee \pi^2$ from scratch to get the same total.

The third binary partition π^3 in effect asks about the third digit in the codes for the u_i, and it is:

$$\pi^3 = \{\{u_1, u_3, u_5, u_7\}, \{u_2, u_4, u_6, u_8\}\} \tag{34}$$

and the final join is:

$$\pi^1 \vee \pi^2 \vee \pi^3 = \{\{u_1\}, ..., \{u_8\}\} = \pi. \tag{35}$$

Since π^3 distinguishes each of the four pairs in $\pi^1 \vee \pi^2$, it introduces $|\{u_1\} \times \{u_2\}| \times 4 \times 2 = 8$ new distinctions for a total of $48 + 8 = 56$ distinctions. Thus the three partitions together make the same 56 distinctions, but Shannon entropy counts the number of those binary partitions (bits) necessary to make the distinctions instead of counting the distinctions or dits themselves. This illustrates that Shannon entropy $H(p) = \sum_{i=1}^{8} \frac{1}{8} \log_2\left(\frac{1}{1/8}\right) = \log_2\left(\frac{1}{1/8}\right) = 3$ is *also* quantifying distinctions in the

Table 2. The dit-bit transform from the compound logical entropies to the corresponding Shannon entropies.

	The Dit-Bit Transform: $1-p_i \rightsquigarrow \log\left(\frac{1}{p_i}\right)$
$h(p) =$	$\sum_i p_i\,(1-p_i)$
$H(p) =$	$\sum_i p_i \log(1/p_i)$
$h(X,Y) =$	$\sum_{x,y} p(x,\,y)\,[1-p(x,y)]$
$H(X,Y) =$	$\sum_{x,y} p(x,y)\log\left(\frac{1}{p(x,y)}\right)$
$m(X,Y) =$	$\sum_{x,y} p(x,y)[[1-p(x)]+[1-p(y)]-[1-p(x,y)]]$
$I(X,\mathrm{Y}) =$	$\sum_{x,y} p(x,y)\left[\log\left(\frac{1}{p(x)}\right)+\log\left(\frac{1}{p(y)}\right)-\log\left(\frac{1}{p(x,y)}\right)\right]$

sense of counting the minimum number of binary partitions, namely three in this case, needed to make the same distinctions. Hence Shannon entropy is a different quantification of the *same* notion of information, *information-as-distinctions* (of a partition). It is not just a quantification of the "amount of uncertainty" – whatever that may be.

Moreover, the example shows how to transform the dit-quantification of information-as-distinctions (logical entropy) into the bit-quantification of information-as-distinctions (Shannon entropy). In this canonical case (all $p_i=\frac{1}{2^n}$), $h(p)=1-p_i$ and $H(p)=\log_2\left(\frac{1}{p_i}\right)$ so the dit-count and bit-count are precisely related: $h(p)=1-\frac{1}{2^{H(p)}}$ and $H(p)=\log_2\left(\frac{1}{1-h(p)}\right)$. In general, the two entropies are the probability averages $\sum_i p_i(\cdots)$ of those canonical values $1-p_i$ and $\log_2\left(\frac{1}{p_i}\right)$. Hence the transform

$$1-p_i \rightsquigarrow \log_2\left(\frac{1}{p_i}\right) \tag{36}$$

transforms logical entropy into Shannon entropy in general:

$$h(p)=\sum_i p_i(1-p_i) \rightsquigarrow H(p)=\sum_i p_i \log_2\left(\frac{1}{p_i}\right). \qquad \text{The Dit}-\text{Bit Transform.} \tag{37}$$

Since the dit-bit transform works for the simple entropies, let us consider the conditional entropies where Shannon constructed $H(X|Y)$ as the average of the Shannon entropies for the conditional probability distributions for $y \in Y$,

$$H(X|Y)=\sum_y p(y)\sum_x \frac{p(x,y)}{p(y)}\log\left(\frac{p(y)}{p(x,y)}\right)=\sum_{x,y} p(x,y)\log\left(\frac{p(y)}{p(x,y)}\right). \tag{38}$$

First, we express the logical conditional entropy as a probability average:

$$\begin{aligned} h(X|Y)=h(X,Y)-h(Y)&=\sum_{x,y} p(x,y)(1-p(x,y))-\sum_y p(y)(1-p(y)) \\ &=\sum_{x,y} p(x,y)[(1-p(x,y))-(1-p(y))], \end{aligned} \tag{39}$$

and then we make the substitutions of the dit-bit transform: $1-p(x,y) \rightsquigarrow \log(1/p(x,y))$ and $1-p(y) \rightsquigarrow \log(1/p(y))$ to get:

$$\sum_{x,y} p(x,y)[\log(1/p(x,y))-\log(1/p(y))]=\sum_{x,y} p(x,y)\log\left(\frac{p(y)}{p(x,y)}\right)=H(X|Y). \tag{40}$$

The other dit-bit transforms go in the same manner at indicated in Table 2.

As one can see, the preservation of the Venn diagram relationships is built into the dit-bit transformation. For instance,

$$m(X,Y)=\sum_{x,y} p(x,y)[[1-p(x)]+[1-p(y)]-[1-p(x,y)]]=h(X)+h(Y)-h(X,Y) \tag{41}$$

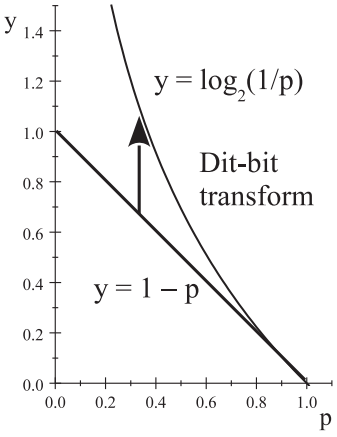

Figure 14. Dit-bit transform and inequality: $1 - p \leq \log_2(1/p)$ for $0 < p \leq 1$.

transforms to:

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y)\left[\log\left(\tfrac{1}{p(x)}\right) + \log\left(\tfrac{1}{p(y)}\right) - \log\left(\tfrac{1}{p(x,y)}\right)\right] \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{42}$$

so that Venn diagram relationships are preserved. The dit-bit transform thus provides the "deeper foundation" ([40], p. 112) sought more than a half-century ago by Lorne Campbell for the Shannon entropies satisfying the Venn diagram relationships in spite of not being defined as a measure on a set.

A basic inequality in Shannon's communications theory is that for positive x, $1 - x \leq \ln(1/x)$. Substituting logs to base 2 for natural logs, it is still true that $1 - p_i \leq \log_2(1/p_i)$ for $0 < p_i \leq 1$ as shown in Figure 14.

The dit-bit transform is just replacing the left-hand side with the right-hand side of the inequality so the transform is highly nonlinear – unlike converting units of measurement like feet and meters. Hence Shannon is correct when he terms his entropy as the "amount of information" ([35], p. 458) denominated in bits. The logical entropy $h(\pi)$ of a partition is a direct measure of the distinctions made by a partition and the Shannon entropy $H(\pi)$ of a partition is statistically the minimum average number of binary partitions that must be joined to make all the distinctions of the partition.

Boltzmann and Shannon entropies: A conceptual connection?

When Shannon showed his formula to John von Neumann, then von Neumann suggested calling it "entropy" for two reasons: there is a similar formula in Boltzmann's statistical mechanics and you can win more arguments using the name "entropy" since no one knows what it really is. (34], pp. 2–3) How does the Shannon formula $\sum_{i=1}^{m} p_i \ln(1/p_i)$ (using natural logs) arise in Boltzmann's statistical mechanics?

The context is n particles that can be in m different states (e.g., energy levels) with a *configuration* (or macrostate) being defined by having n_i particles in the ith state so $\sum_{i=1}^{m} n_i = n$. If all the m^n possible assignments ("microstates") of the n particles to the m states are equiprobable, then Boltzmann's idea was that the system would evolve to the macrostate that had the highest probability. Since the number of microstates for any configuration is the multinomial coefficient $\binom{n}{n_1, \ldots, n_m} = \frac{n!}{n_1! \ldots n_m!}$, the larger the multinomial coefficient, the larger the probability of that configuration. Hence to find the equilibrium configuration, the problem is to maximize the multinomial coefficient subject to the relevant constraints. In addition to $\sum_{i=1}^{m} n_i = n$, each of the m states would have an associated energy level ε_i and the total energy $\sum_{i=1}^{m} n_i \varepsilon_i$ should equal a constant value E. Where did the Shannon formula come from in Boltzmann's nineteenth-century statistical mechanics?

Since the natural log is a monotonic transformation, it is equivalent to maximize $\ln\left(\frac{n!}{n_1! \ldots n_m!}\right)$. Moreover, the log gives an additive quantity to be associated with the extensive quantity of entropy in thermodynamics. However, maximizing $\ln\left(\frac{n!}{n_1! \ldots n_m!}\right)$ subject to the constraints is not very analytically tractable due to the presence of the factorials $n!$ and $n_i!$. But there is the Stirling infinite series expression for $\ln(n!)$ and for large n, just the first few terms in the series will give

Table 3. Feasible integer solutions.

n_1	n_2	n_3	$n!/(n_1!n_2!n_3!)$
1	6	3	840
2	4	4	3150
3	2	5	2520
4	0	6	210

a "for all practical purposes" good approximation. In particular, the first two terms give the approximation: $n\ln(n) - n \approx \ln(n!)$. Using that numerical approximation, normalizing by dividing by n, and ignoring the physical Boltzmann's constant, yields a familiar expression as the approximation:

$$\begin{aligned} S &= \tfrac{1}{n}\ln\left(\tfrac{n!}{n_1!\ldots n_m!}\right) = \tfrac{1}{n}\left[\ln(n!) - \sum_{i=1}^{m}\ln(n_i!)\right] \\ &\approx \tfrac{1}{n}\left[n[\ln(n)-1] - \sum_{i=1}^{m} n_i[\ln(n_i)-1]\right] = \tfrac{1}{n}\left[n\ln(n) - \sum_i n_i\ln(n_i)\right] \\ &= \tfrac{1}{n}\left[\sum n_i\ln(n) - \sum n_i\ln(n_i)\right] = -\tfrac{1}{n}\sum_i n_i\ln\left(\tfrac{n_i}{n}\right) \\ &= -\sum_{i=1}^{m} p_i\ln(p_i) = \sum_{i=1}^{m} p_i\ln(1/p_i) = H_e(p) \quad \text{where } p_i = n_i/n. \end{aligned} \tag{43}$$

That is how the two-term Stirling approximation brings the Shannon formula into the statistical mechanics of Boltzmann (and Gibbs). It should be noted that the two probability distributions are quite different. For the exact maximal configuration $(n_1, \ldots, n_m)$, there are $\frac{n!}{n_1!\ldots n_m!}$ terms in the equiprobable distribution over the microstates. For the two-term Stirling approximate probability distribution, there are only m terms $(p_1, \ldots, p_m)$. It is the total quantity $\frac{1}{n}\ln\left(\frac{n!}{n_1!\ldots n_m!}\right)$ that is approximated by the Shannon formula $H_e(p)$ for large n.

And by taking more terms in the Stirling approximation, one information theorist notes that one would have an even better approximation ([53], p. 2), and a prominent physical chemist notes that $\ln(n!) \approx \sqrt{\ln(2\pi)} + \left(n+\frac{1}{2}\right)\ln(n) - n$ is a much better approximation ([54], p. 533). But neither uses those formulas since the purpose at hand is analytical tractability, not better approximations, and the Shannon formula leads to a very nice development in statistical mechanics – in particular to the beautiful partition function Z that connects statistical mechanics to thermodynamics. Unfortunately, the role of what became later known as the Shannon formula as a very convenient numerical approximation to Boltzmann entropy is often "forgotten" in the literature where one even sees expressions like "Shannon-Boltzmann entropy" ([46], p. 11) or "Boltzmann–Gibbs–Shannon entropy" (e.g., [36]). Perhaps nowhere else in mathematical physics has a *numerical* approximation been attributed such *conceptual* significance.

Another way to emphasize the conceptual difference is to consider a small n example where we can compute both entropies since the original Boltzmann problem is a tractable integer programming problem *not using any approximation*. Consider an example of $n = 10$ particles with three possible energy levels of $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3) = (1, 2, 3)$ and a total energy of $E = 22$. For n_i as the number of particles at energy level i, the energy constraint is $\sum_{i=1}^{3}\varepsilon_i n_i = E$ and of course $\sum_{i=1}^{3} n_i = n$.[6] There are only four non-negative integer solutions satisfying the two constraints (see Tab. 3).

The exact Boltzmann solution giving the maximum multinomial coefficient is $(n_1, n_2, n_3) = (2, 4, 4)$ and the (normalized) Boltzmann entropy is:

$$S = \frac{1}{10}\ln\left(\frac{10!}{2!4!4!}\right) = \frac{1}{10}\ln(3150) = \frac{1}{10}8.055 = 0.8055, \tag{44}$$

while maximizing the usual Shannon approximation gives the non-integer result $(n_1, n_2, n_3) = (2.3837, 3.2326, 4.3837)$ (to four decimal places) with the Shannon entropy of $H_e(p) = 1.0684$ (where $p_i = n_i/n$). The probability distribution in the Boltzmann case has 3150 equal terms with the value $\frac{1}{3150}$ while the probability distribution in the "Shannon case" has 3 terms, $\frac{1}{10}(2.3837, 3.2326, 4.3837)$. The maximization of the multinomial coefficient (or its normalized logarithm) and the Shannon expression are obviously different for low n. But for the enormous number of particles in a system of statistical mechanics, that numerical difference fades into insignificance – unless one forgets about it altogether and attaches *conceptual* significance to the Shannon formula in Boltzmannian statistical mechanics.

[6] The example was inspired by Eric Johnson's excellent treatment of Boltzmann's entropy [55].

MaxEntropy with which entropy for discrete distributions?

Edwin T. Jaynes [56] started a whole "MaxEntropy" subdiscipline in information theory by arguing that the classical indifference principle used to give an equiprobable probability distribution (in the lack of other knowledge) should be generalized to other more constrained contexts by choosing the probabilities that maximize the Shannon entropy subject to those constraints. His motivation was based, in significant part, on attaching conceptual significance to the maximizing of Shannon entropy in Boltzmannian statistical mechanics:

> the "method of the most probable distribution" dating back to Boltzmann ... which turns out in the end to be mathematically equivalent to maximum [Shannon] entropy. ([56], p. 441)

The question naturally arises: "What about maximizing logical entropy subject to the same constraints?" If there are no constraints, then maximizing both entropies yields the classical result of the equiprobable distribution, i.e., the indifference principle. But when there are constraints, then the two maximums yield different probability distributions.

Consider a function $X : U \to \mathbb{R}$ with values $X(u_i) = x_i$ for $i = 1, ..., n$ with unknown probabilities $p = (p_1, \ldots, p_n)$. A standard discrete MaxEntropy problem is to find the "best" probabilities so that the average value $\sum_{i=1}^{n} p_i x_i = m$ for some given value of m (which must be between the maximum and minimum values of the x_i).

Where "best" is defined by maximizing the Shannon entropy, the procedure is to maximize the Lagrangian:

$$\mathcal{L} = -\sum p_i \ln(p_i) + \lambda\left(1 - \sum p_i\right) + \tau\left(m - \sum p_i x_i\right) \tag{45}$$

so the first-order conditions are:

$$\partial\mathcal{L}/\partial p_i = -\ln(p_i) - p_i\frac{1}{p_i} - \lambda - x_i\tau = 0, \tag{46}$$

where it should be noted at the outset that the use of the log function $\ln(p_i)$ (and the term $1/p_i$ in the first-order conditions) assumes $p_i \neq 0$ for all i. Then exponentiating gives: $p_i = \mathrm{e}^{-(1+\lambda+\tau x_i)}$. Substituting into the constraints to determine the Lagrange multipliers:

$$1 = \sum p_i = \sum \mathrm{e}^{-(1+\lambda+\tau x_i)} = \mathrm{e}^{-(1+\lambda)} \sum \mathrm{e}^{-\tau x_i} \quad \text{so } \mathrm{e}^{1+\lambda} = \sum \mathrm{e}^{-\tau x_i} \tag{47}$$

which yields $p_{\mathrm{Max}H}$ in terms of the Lagrange multiplier τ as:

$$p_i = \mathrm{e}^{-\tau x_i} / \sum_{j=1}^{n} \mathrm{e}^{-\tau x_j}. \tag{48}$$

And $m = \sum x_i \mathrm{e}^{-(1+\lambda+\tau x_i)} = \sum x_i \mathrm{e}^{-\tau x_i} / \sum \mathrm{e}^{-\tau x_i}$. Rather than trying to solve directly for τ, it is best to let $w = \mathrm{e}^{-\tau}$ and then numerically solve for a real root w (aside from $w = 0$) of the equation:

$$\sum_i x_i w^{x_i} - m\left(\sum_i w^{x_i}\right) = 0. \tag{49}$$

Given such a real w, $\tau = -\ln(w)$ and then the $p_{\mathrm{Max}H} = (p_1, \ldots, p_n)$ for maximizing Shannon entropy $H(p)$ are determined by the above formula: $p_i = \mathrm{e}^{-\tau x_i} / \sum_{j=1}^{n} \mathrm{e}^{-\tau x_j}$. Moreover, it is clear from the formula that all the p_i are positive (and sum to 1).

Where "best" is defined by maximizing logical entropy, the procedure is to solve the quadratic programming problem of maximizing $h(p) = 1 - \sum_i p_i^2$ subject to the same constraints $\sum_i p_i x_i = m$ and $\sum_i p_i = 1$ plus the additional non-negativity constraints $0 \leq p_i$ for $i = 1, \ldots, n$. For a certain range of values of m, the non-negativity constraints will be automatically satisfied so one can approach that part of the problem using the Lagrangian approach.

$$\mathcal{L}(p_1, ..., p_n) = 1 - \sum_i p_i^2 - \lambda\left(1 - \sum p_i\right) + \tau\left(m - \sum p_i x_i\right) \tag{50}$$

so the first-order conditions are:

$$\partial\mathcal{L}/\partial p_i = -2p_i + \lambda - \tau \mathrm{x}_i = 0, \tag{51}$$

so

$$p_i = \frac{1}{2}(\lambda - \tau x_i). \tag{52}$$

Using the first constraint:

$$1 = \sum p_i = \frac{1}{2}\left(n\lambda - \tau \sum x_i\right) = \frac{n}{2}\lambda - \frac{1}{2}\tau \sum x_i, \tag{53}$$

and using the second constraint:

$$m = \sum p_i x_i = \sum x_i \frac{1}{2}(\lambda - \tau x_i) = \lambda \frac{1}{2}\sum x_i - \tau \frac{1}{2}\sum x_i^2 \tag{54}$$

so we have two linear equations that can be used to solve for the Lagrange multipliers λ and τ. Before going forward, it is useful to consider the mean and variance of the x_i's if they were equiprobable. Then $\mu = \frac{\sum x_i}{n}$ and $\mathrm{Var}(X) = E(X^2) - \mu^2 = \frac{\sum x_i^2}{n} - \mu^2 = \sigma^2$ so $n\mathrm{Var}(X) = \sum x_i^2 - n\mu^2$. Then the two equations are:

$$1 = \frac{n}{2}\lambda - \frac{n\mu}{2}\tau \quad \text{and} \quad m = \frac{n\mu}{2}\lambda - \frac{n}{2}\left[\mathrm{Var}(X) + \mu^2\right]\tau. \tag{55}$$

After a bit of algebra, one arrives at the informative formula for the p_i in $p_{\mathrm{Max}h} = (p_1, \ldots, p_n)$ that results from maximizing logical entropy subject to the same constraints:

$$p_i = \frac{1}{n} + \frac{(\mu - m)(\mu - x_i)}{n\mathrm{Var}(X)} = \frac{1}{n} + \frac{1}{n}\left(\frac{m-\mu}{\sigma}\right)\left(\frac{x_i - \mu}{\sigma}\right). \tag{56}$$

Since all the operations in the formula are rational (e.g., no square roots, not to mention transcendental functions), the probabilities are all rational if all the x_i are rational. One test of intuitiveness is: if x_i is equal to the equiprobable mean μ, then shouldn't that p_i equal the equiprobable value $\frac{1}{n}$ regardless of the other values? That is true aswe see from the formula for p_i. If any $x_i = \mu$ then that $p_i = \frac{1}{n}$ and if $m = \mu$, then all the $p_i = \frac{1}{n}$, the equiprobable solution. The condition for all the $p_i \geq 0$ is that $\frac{(\mu - m)(\mu - x_i)}{n\mathrm{Var}(X)} \geq -\frac{1}{n}$ or $(\mu - m)(\mu - x_i) \geq -\mathrm{Var}(X)$ for all i. If that condition is not satisfied for some p_i, then the non-negativity constraints must be enforced by using quadratic programming techniques instead of the Lagrangian technique used above.[7]

One of the best-known examples is Jaynes's Brandeis dice problem ([59], p. 47 or [60], p. 427). If a die was fair, then the average of the equiprobable outcomes is $\mu = 3.5$. But suppose that it is a given constraint that the average outcome is 4.5, then what is the "best" estimate of the probabilities for the six sides?

To maximize Shannon entropy, the x_i's are 1, 2, 3, 4, 5, 6 so the equation to be numerically solved is:

$$\sum_{i=1}^{6} i w^i - m\left(\sum_{i=1}^{6} w^i\right) = 0. \tag{57}$$

for $m = 4.5$. In addition to $w = 0$, the relevant real root to four decimal places is $w = 1.4493$. Then $\tau = -\ln(1.449\,3) = -0.371\,08$ and the Jaynes solution for the probabilities to four decimal places is:

$$p_{\mathrm{Max}H} = (0.0543, 0.0788, 0.1142, 0.1654, 0.2398, 0.3475). \tag{58}$$

To maximize logical entropy, $\mu = \frac{7}{2}$, $m = \frac{9}{2} = 4.5$, and $\mathrm{Var}(X) = \frac{35}{12}$, so the formula $p_i = \frac{1}{n} + \frac{(\mu - m)(\mu - x_i)}{n\mathrm{Var}(X)}$ can be used to solve for the *rational* maximum logical entropy solution:

$$\begin{aligned} p_{\mathrm{Max}h} &= \tfrac{1}{210}(5, 17, 29, 41, 53, 65) \\ &= (0.0238, 0.0810, 0.1381, 0.1952, 0.2524, 0.3095). \end{aligned} \tag{59}$$

In this case, $(\mu - m)(\mu - x_i) = \left(\frac{7}{2} - \frac{9}{2}\right)\left(\frac{7}{2} - x_i\right) = -\left(\frac{7}{2} - x_i\right) \geq -\mathrm{Var}(X) = -\frac{35}{12}$. The RHS is the smallest for $x_1 = 1$ where $-\left(\frac{7}{2} - \frac{2}{2}\right) = -\frac{30}{12} \geq -\frac{35}{12}$ so all the probabilities are positive. Equality holds when $\mu - m = -\frac{35}{12}\frac{2}{5} = -\frac{7}{6}$ or $m = \frac{7}{2} + \frac{7}{6} = \frac{28}{6} = 4\frac{2}{3}$. Hence for any $m > 4\frac{2}{3}$, p_1 and possibly other p_i will be 0 so quadratic programming must be used. A little calculation shows that $\frac{7}{3}$ is the lower bound so that for $m < \frac{7}{3}$, there will be some zero probabilities. For instance, for $m = 5$, the probabilities for logical entropy are: $p_{Maxh} = \frac{1}{10}(0, 0, 1, 2, 3, 4)$, while the Jaynes solution is $p_{\mathrm{Max}H} = (0.0205, 0.0385, 0.0723, 0.1357, 0.2548, 0.4781)$ to four decimal places.

It is interesting that the only alternative to maximizing Shannon entropy that Jaynes considers ([56], pp. 345–6) is minimizing $\sum_i p_i^2$ which is the same as maximizing logical entropy $h(p) = 1 - \sum_i p_i^2$. But then he criticizes it because some of the p_i may be negative if one uses only the Lagrangian method.

[7] Microsoft Excel with the Solver application is sufficient. For a thorough treatment, see [57] or [58].

> The formal solution for minimum $\sum_i p_i^2$ lacks the property of non-negativity. We might try to patch this up in an *ad hoc* way by replacing the negative values by zero and adjusting the other probabilities to keep the constraint satisfied. [56], p. 346]

It is unclear if Jaynes was aware of the field of quadratic programming which was well-developed in the 1960s ([61], p. 490) and which hardly proceeds in "an *ad hoc* way by replacing the negative values by zero and adjusting the other probabilities to keep the constraint satisfied."

Clearly the two solutions are different in general[8] and each one maximizes the corresponding type of entropy. How can one determine which probability distribution is "best"? One criterion that immediately suggests itself is the distribution $(p_1, \ldots, p_n)$ of numbers that is the most uniform in the sense of having the least variance $\mathrm{Var}(p)$ where each of the numbers p_i is considered equally probable. The minimum is $\mathrm{Var}(p) = 0$ for the uniform probability distribution which maximizes both entropies in the absence of constraints. In the two cases where $m = 4.5$ and $m = 5$, the logical entropy maximizing distribution $p_{\mathrm{Max}h}$ has the lower variance $\mathrm{Var}(p)$. But is that true in general?

At first, it seems rather intractable to prove in general which of the two distributions has least variance in the discrete case since the Jaynes solution involves finding the roots of a high-degree polynomial. But there is an easy and general proof that the logical entropy solution has a variance less than (or equal to) the Jaynes solution when both are maximized subject to the same constraints – and similarly for being closest to the uniform distribution in terms of the usual notion of Euclidean distance in $\mathbb{R}^n$.

Proposition 3.2. *$Var(p_{Maxh}) \leq Var(p_{MaxH})$.*

Proof. For any constraint set on the probability distributions $p = (p_1, \ldots, p_n)$, minimize the variance itself over all the feasible distributions (rather than maximize either of the two entropies – or any other entropy for that matter), and then show that the minimum variance distribution p_{MinVar} is the same as the maximum logical entropy distribution $p_{\mathrm{Max}h}$. The equality $p_{\mathrm{MinVar}} = p_{\mathrm{Max}h}$ is shown by computing the relationship between $\mathrm{Var}(p)$ and $h(p)$. Looking at $(p_1, \ldots, p_n)$ as just a set of equiprobable numbers with $\sum_i p_i = 1$, it has the variance:

$$\begin{aligned} \mathrm{Var}(p) &= \sum_{i=1}^{n} \tfrac{1}{n}\left(p_i - \tfrac{1}{n}\right)^2 = E(p^2) - E(p)^2 \\ &= \tfrac{1}{n}\sum_i p_i^2 - \left(\tfrac{1}{n}\sum_i p_i\right)^2 = \tfrac{1}{n}\sum_i p_i^2 - \left(\tfrac{1}{n}\right)^2 = \tfrac{1}{n}\left[\left(1 - \tfrac{1}{n}\right) - h(p)\right] \end{aligned} \tag{60}$$

since $\sum_{i=1}^n p_i^2 = 1 - h(p)$. Since $h(p)$ appears with a negative sign in the expression for $\mathrm{Var}(p)$, minimizing $\mathrm{Var}(p)$ is the same as maximizing $h(p)$ over the set of feasible probability distributions, so $p_{\mathrm{MinVar}} = p_{\mathrm{Max}h}$. □

Corollary 3.3. *p_{Maxh} minimizes the (Euclidean) distance to the uniform distribution $\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$.*

Proof. Minimizing Euclidean distance is the same as minimizing the distance squared and $\sum_{i=1}^n \left(p_i - \frac{1}{n}\right)^2 = \left(1 - \frac{1}{n}\right) - h(p)$. □

There is another specialized notion of "distance," namely the *Kullback–Leibler divergence* [62] $D(p||q) = \sum_{i=1}^n p_i \log\left(\frac{p_i}{q_i}\right)$ (p and q are probability distributions on the same index set with all $q_i > 0$) which is neither symmetrical nor satisfies the triangle inequality. But $D\left(p||\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)\right) = \log(n) - H(p)$ so that maximizing $H(p)$ subject to the constraints is equivalent to minimizing the Kullback–Leibler divergence of p from the uniform distribution.

The corresponding asymmetrical divergence formula for logical entropy, also for probability distributions p and q where $q_i > 0$ for all i, is the *directed logical divergence*:

$$d^*(p||q) := \sum_{i=1}^{n} \frac{1}{q_i}(q_i - p_i)^2 = \sum_i p_i\left(\frac{p_i}{q_i} - 1\right) = \sum_i \frac{p_i^2}{q_i} - 1 \geq 0 \quad \text{with equality iff } p = q. \tag{61}$$

Another way to prove non-negativity is to note that $(q_i - p_i)\left(1 - \frac{p_i}{q_i}\right) \geq 0$ since both terms are negative or both are non-negative, and $\sum_i (q_i - p_i)\left(1 - \frac{p_i}{q_i}\right) = \sum_i \frac{p_i^2}{q_i} - 1$. Since the KL divergence uses probability ratios inside the log term, we do the same for dit-bit transform so that: $1 - \frac{p_i}{q_i} \rightsquigarrow \log\left(\frac{1}{p_i/q_i}\right)$ and thus:

$$-d^*(p||q) = \sum_i p_i\left(1 - \frac{p_i}{q_i}\right) \rightsquigarrow \sum_i p_i \log\left(\frac{1}{p_i/q_i}\right) = \sum_i p_i \log\left(\frac{q_i}{p_i}\right) = -D(p||q) \tag{62}$$

[8] For $n = 2$, the two solutions are identical but diverge in general for $n \geq 3$.

so $d^*(p||q) \rightsquigarrow D(p||q)$, i.e., the KL divergence is the dit-bit transform of the directed logical divergence. What is the probability distribution closest to the uniform distribution using the directed logical divergence?

$$\begin{aligned} d^*\left(p||\left(\tfrac{1}{n},...,\tfrac{1}{n}\right)\right) &= \sum_i p_i(np_i - 1) = n\sum_i p_i^2 - 1 \\ &= n\left[\left(1 - \tfrac{1}{n}\right) - h(p)\right] = n\sum_{i=1}^{n}\left(p_i - \tfrac{1}{n}\right)^2 \end{aligned} \tag{63}$$

so it is the logical entropy solution that is the closest to the uniform distribution by the logical notion of directed divergence.

Metrical logical entropy = (twice) variance

The above results suggest a broader connection between the usual notion of the variance of a random variable and the logical entropy of "differences" when the differences have metrical significance. The logical entropy $h(X)$ of a random variable $X : U \to \mathbb{R}$ with n distinct values $(x_1, \ldots, x_n)$ with the probabilities $p = (p_1, \ldots, p_n)$ is computed as $h(X) = \sum_{i \neq j} p_i p_j$ which only takes notice of when values are the same or different. Logical entropy in that sense is a special case of C.R. Rao's notion of quadratic entropy $\sum_{i,j} d_{ij}\, p_i\, p_j$, where d_{ij} is a non-negative "distance function" such that $d_{ii} = 0$ and $d_{ij} = d_{ji}$ [26, 63], for the logical distance function $d_{ij} = 1 - \delta_{ij}$, the complement of the Kronecker delta. A natural *metrical* distance function is the Euclidean distance squared $d_{ij} = (x_i - x_j)^2$.

Proposition 3.4. $\sum_{j \neq i} p_i p_j \left(x_i - x_j\right)^2 = 2Var(X)$.

Proof. Firstly, since for $i = j$, $\left(x_i - x_j\right)^2 = 0$, we can sum over all i, j.

$$\begin{aligned} \sum_{j \neq i} p_i p_j \left(x_i - x_j\right)^2 &= \sum_{i,j} p_i p_j \left(x_i - x_j\right)^2 \\ &= \sum_{i,j} p_i p_j \left(x_i^2 - 2x_i x_j + x_j^2\right) = E\left(X^2\right) - 2E(X)^2 + E\left(X^2\right) = 2\mathrm{Var}(X) \end{aligned} \tag{64}$$

□

It was previously noted that when counting distinctions $(u_i, u_j) \in \mathrm{dit}(\pi)$, both (u_i, u_j) and (u_j, u_i) are included. If only the distinctions (u_i, u_j) for $i < j$ are counted, then one get half the number as is evident in the logical entropy box diagrams such as Figure 2.

Corollary 3.5. $\sum_{i<j} p_i p_j \left(x_i - x_j\right)^2 = Var(X)$.

Thus the variance of a metrical random variable X is the average distance squared between the values in an unordered pair of independent trials.

The result extends to covariances as well. Consider two real-valued random variables X with distinct values x_i for $i = 1, \ldots, n$ and Y with distinct values y_j for $j = 1, \ldots, m$ with the joint probability distribution $p\left(x_i, y_j\right) : X \times Y \to \mathbb{R}$. Two ordered draws from $X \times Y$ gives two ordered pairs: (x_i, y_j) and $\left(x_{i'}, y_{j'}\right)$. For this bivariate distribution, the generalization of $\sum_{j \neq i} p_i\, p_j\, (x_i - x_j)^2$ is:

$$\sum_{(i,j) \neq \left(i',j'\right)} \mathrm{p}(x_i, y_j) p\left(x_{i'}, y_{j'}\right)\left(x_i - x_{i'}\right)\left(y_j - y_{j'}\right). \tag{65}$$

Metrical logical entropy for bivariate distributions of metrical random variables which is no longer a special case of quadratic entropy since $\left(x_i - x_{i'}\right)\left(y_i - y_{i'}\right)$ can be negative. The (unordered) two-draw notion of metrical variation for a bivariate distribution reproduces the usual notion of covariance $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$.[9]

Proposition 3.6. $\sum_{(i,j) \neq \left(i',j'\right)} p(x_i, y_j) p\left(x_{i'}, y_{j'}\right)\left(x_i - x_{i'}\right)\left(y_j - y_{j'}\right) = 2\mathrm{Cov}(X, Y)$.

Proof. Since $\left(x_i - x_{i'}\right)\left(y_j - y_{j'}\right) = 0$ if $i = i'$ or $j = j'$, we can sum over all i, j. Abbreviating $p(x_i, y_j) = p_{ij}$, we have:

[9] These formulas, for the equiprobable case, were derived using the "difference method" by Zhang et al. [64] as new formulas for variance and covariance although the formulas may be much older "folk theorems."

$$\begin{aligned}
&\sum_{i,j,i',j'} p_{ij}p_{i'j'}\left(x_i - x_{i'}\right)\left(y_j - y_{j'}\right) \\
&= \sum_{i,j,i',j'} p_{ij}p_{i'j'}\left[x_iy_j - x_iy_{j'} - x_{i'}y_j + x_{i'}y_{j'}\right] \\
&= \sum_{i,j,i',j'} p_{ij}p_{i'j'}x_iy_j - \sum_{i,j,i',j'} p_{ij}p_{i'j'}x_iy_{j'} - \sum_{i,j,i',j'} p_{ij}p_{i'j'}x_{i'}y_j + \sum_{i,j,i',j'} p_{ij}p_{i'j'}x_{i'}y_{j'}.
\end{aligned} \tag{66}$$

Then using:

$$\sum_{i,j,i',j'} p_{ij}p_{i'j'}x_iy_j = \sum_{i,j} p_{ij}x_iy_j \sum_{i',j'} p_{i'j'} = \sum_{ij} p_{ij}x_iy_j = E(XY), \tag{67}$$

and

$$\begin{aligned}
\textstyle\sum_{i,j,i',j'} p_{ij}p_{i'j'}x_iy_{j'} &= \sum_{i,j'} x_iy_{j'} \sum_{i'} \sum_{j} p_{ij}p_{i'j'} = \sum_{i,j'} x_iy_{j'} \sum_{i'} p_ip_{i'j'} \\
&= \sum_{i,j'} p_ix_iy_{j'}p_{j'} = \left(\sum_i p_ix_i\right)\left(\sum_{j'} p_{j'}y_{j'}\right) = E(X)E(Y)
\end{aligned} \tag{68}$$

and similarly for the other cases, so we have:

$$\sum_{(i,j)\neq(i',j')} p(x_i,y_j)p\left(x_{i'},y_{j'}\right)\left(x_i - x_{i'}\right)\left(y_j - y_{j'}\right) = E(XY) - E(X)E(Y) - E(Y)E(X) + E(XY) = 2\mathrm{Cov}(X,Y). \tag{69}$$

□

The linear ordering on indices i and j can be extended to the linear lexicographic (or dictionary) ordering on ordered pairs of indices where $(i, j) < (i', j')$ if $i < i'$ or if $i = i'$, then $j < j'$. Then for each pair of distinct ordered pairs $(i, j) \neq (i', j')$, either $(i, j) < (i', j')$ or $(i', j') < (i, j)$ but not both, so $(i, j) < (i', j')$ picks out half the cases of $(i, j) \neq (i', j')$.

Corollary 3.7. $\sum_{(i,j)<(i',j')} p(x_i,y_i)p\left(x_{i'},y_{j'}\right)\left(x_i - x_{i'}\right)\left(y_j - y_{j'}\right) = Cov(X,Y)$.

In the switch from logical entropy to metrical logical entropy, the interpretation switches from being a two-draw probability (and thus always non-negative) to being a two-draw average metrical quantity which, like the covariance, might be positive or negative.

Thus logical entropy connects naturally with the notions of variance and covariance in statistics. Although beyond the scope of this paper, the metrical logical entropy for a discrete random variable $g(X)$, $\sum_{j\neq i} p_ip_j\left(g(x_i) - g\left(x_j\right)\right)^2 = 2\mathrm{Var}(g(X))$, shows how to generalize to the logical entropy of a continuous random variable $g(X)$ where X has the probability density $f(x)$:

$$h(g(X)) = \int f(x)\int f\left(x'\right)\left(g(x) - g\left(x'\right)\right)^2 \mathrm{d}x'\,\mathrm{d}x = 2\mathrm{Var}(g(X)). \tag{70}$$

The interpretation of $h(g(X))$ is the average (Euclidean) distance squared between the values of $g(X)$ on two independent trials – which is twice the variance.

Quantum logical entropy

Logical entropy via density matrices

The transition from "classical" (i.e., non-quantum) logical entropy to quantum logical entropy is facilitated by reformulating logical entropy using density matrices over the real numbers. A stepping stone in that reformulation is the notion of an incidence matrix of a binary relation. For a finite $U = \{u_1, ..., u_n\}$, a binary relation R on U is a subset $R \subseteq U \times U$. The $n \times n$ incidence matrix $\mathrm{In}(R)$ is defined by:

$$\mathrm{In}(R)_{ij} = \begin{cases} 1 & \text{if}\left(u_i,u_j\right) \in R \\ 0 & \text{if}\left(u_i,u_j\right) \notin R. \end{cases} \tag{71}$$

Then the incidence matrix associated with a partition $\pi = \{B_1, \ldots, B_m\}$ is In(indit(π)), the incidence matrix of the partition's inditset, i.e., the associated equivalence relation. And then for equiprobable points in U, the density matrix $\rho(\pi)$ associated with π is the incidence matrix In(indit(π)) rescaled to be of trace 1 (trace = sum of diagonal elements):

$$\rho(\pi) = \frac{1}{n}\mathrm{In}(\mathrm{indit}(\pi)). \tag{72}$$

Each off-diagonal element has two associated diagonal elements in its row and column. If an off-diagonal element in In(indit (π)) or $\rho(\pi)$ is non-zero, then the corresponding diagonal elements are for elements u_i, $u_j \in B_k$ for some block $B_k \in \pi$.

For U with point probabilities $p = (p_1,...,p_n)$, the density matrix $\rho(\pi)$ can be constructed block by block. For a block $B_i \in \pi$, let $|B_i\rangle$ be the column vector with the jth entry being $\sqrt{\frac{p_j}{\Pr(B_i)}}$ if $u_j \in B_i$ and otherwise 0. Then the $\rho(B_i)$ is the $n \times n$ matrix formed by the product of column vector $|B_i\rangle$ times its row vector transpose $|B_i\rangle^{\mathrm{t}}$, and the *density matrix* ρ (π) is the probability-weighted sum:

$$\rho(\pi) = \sum_{B_i \in \pi} \Pr(B_i)\rho(B_i). \tag{73}$$

Then each jk entry in $\rho(\pi)$ is:

$$\rho(\pi)_{jk} = \begin{cases} \sqrt{p_j p_k} & \text{if } (u_j, u_k) \in \mathrm{indit}(\pi) \\ 0 & \text{otherwise.} \end{cases} \tag{74}$$

These values are the square roots of the unshaded squares in the logical entropy box diagrams, e.g., Figures 2–5.
For instance, if $\pi = \{\{u_1, u_3\},\{u_2, u_4\}\}$, then:

$$\rho(\pi) = \begin{bmatrix} p_1 & 0 & \sqrt{p_1 p_3} & 0 \\ 0 & p_2 & 0 & \sqrt{p_2 p_4} \\ \sqrt{p_3 p_1} & 0 & p_3 & 0 \\ 0 & \sqrt{p_4 p_2} & 0 & p_4 \end{bmatrix}, \tag{75}$$

where the non-zero off-diagonal elements indicate which elements are in the same block of the partition. With a suitable interchange of rows and columns, the matrix would become block-diagonal – where the entries squared correspond to the values of the unshaded squares in the box diagram for logical entropy. The density matrix is symmetric, has trace 1, and all non-negative elements.

The most important calculation for our purposes is the trace of the square $\rho(\pi)^2$ of a density matrix. Consider a diagonal element $(\rho(\pi)^2)_{jj}$ where $u_j \in B_i$ which is the product of the jth row times the jth column of $\rho(\pi)$:

$$\left(\rho(\pi)^2\right)_{jj} = \sum_{u_k \in B_i} \sqrt{p_j p_k}\sqrt{p_k p_j} = p_j \sum_{u_k \in B_i} p_k = p_j \Pr(B_i). \tag{76}$$

Then summing over all those diagonal elements for $u_j \in B_i$ gives $\sum_{u_j} p_j \Pr(B_i) = \Pr(B_i)^2$. These block probabilities squared were the values assigned to the unshaded blocks in the box diagrams for logical entropy. Finally summing over all the diagonal elements yields the basic result about the trace of $\rho(\pi)^2$:

$$\mathrm{tr}\left[\rho(\pi)^2\right] = \sum_{B_i \in \pi} \Pr(B_i)^2. \tag{77}$$

This result immediately yields the translation of the logical entropy $h(\pi)$ into the density matrix formalism:

$$h(\pi) = 1 - \mathrm{tr}\left[\rho(\pi)^2\right], \tag{78}$$

i.e., the sum of the shaded squares in the box diagrams for logical entropy.
We will define the tensor product of matrices by considering the example of a 2×2 matrix $\mathbf{A}$ times a 3×3 matrix $\mathbf{B}$:

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{bmatrix}$$

$$= \begin{matrix} (1,1) \\ (1,2) \\ (1,3) \\ (2,1) \\ (2,2) \\ (2,3) \end{matrix} \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{11}b_{13} & a_{12}b_{11} & a_{12}b_{12} & a_{12}b_{13} \\ a_{11}b_{21} & a_{11}b_{22} & a_{11}b_{23} & a_{12}b_{21} & a_{12}b_{22} & a_{12}b_{23} \\ a_{11}b_{31} & a_{11}b_{32} & a_{11}b_{33} & a_{12}b_{31} & a_{12}b_{32} & a_{12}b_{33} \\ a_{21}b_{11} & a_{21}b_{12} & a_{21}b_{13} & a_{22}b_{11} & a_{22}b_{12} & a_{22}b_{13} \\ a_{21}b_{21} & a_{21}b_{22} & a_{21}b_{23} & a_{22}b_{21} & a_{22}b_{22} & a_{22}b_{23} \\ a_{21}b_{31} & a_{21}b_{32} & a_{21}b_{33} & a_{22}b_{31} & a_{22}b_{32} & a_{22}b_{33} \end{bmatrix}. \tag{79}$$

In particular, it might be noted that all the diagonal elements have the form $a_{ii}\, b_{jj}$ but their (row, column) designators are $(i, j)(i, j)$. Thus $a_{11}\, b_{33}$ is the diagonal element in the (1, 3) row and the (1, 3) column, i.e., a diagonal element of the tensor product $\mathbf{A}\otimes\mathbf{B}$.

One can take the tensor product of an $n \times n$ density matrix $\rho(\pi)$ (where $\pi = f^{-1}$ for some $f : U \to \mathbb{R}$) with itself to obtain a $n^2 \times n^2$ matrix whose diagonal elements are $(\rho(\pi)\otimes\rho(\pi))_{(i,j)(i,j)} = p_i p_j$. Let $P_{\mathrm{dit}}(\pi)$ be the $n^2 \times n^2$ diagonal (projection) matrix with diagonal elements $\left(P_{\mathrm{dit}(\pi)}\right)_{(i,j)(i,j)} = \chi_{\mathrm{dit}(\pi)}\left(u_i, u_j\right)$. Then the matrix product $P_{\mathrm{dit}(\pi)}\rho(\pi)\otimes\rho(\pi)$ will have the non-zero diagonal elements $p_i\, p_j$ for $\left(u_i, u_j\right) \in \mathrm{dit}(\pi)$, and thus:

$$h\left(f^{-1}\right) = h(\pi) = \sum_{\left(u_i,u_j\right)\in\mathrm{dit}(\pi)} p_i p_j = \mathrm{tr}\left[P_{\mathrm{dit}(\pi)}\rho(\pi)\otimes\rho(\pi)\right]. \tag{80}$$

That formula will carry over to the quantum case.

In general, a density matrix ρ is said to represent a *pure state* if $\mathrm{tr}[\rho^2] = 1$, and otherwise a *mixed state*. For partitions, the only pure state density matrix is $\rho(\mathbf{0}_U)$, the density matrix of the indiscrete partition $\mathbf{0}_U = \{U\}$ on U which has zero logical entropy.

Given another partition $\sigma = \{C_1, ..., C_{m'}\}$ on U, the join partition $\pi \vee \sigma$ is the partition whose blocks are all the non-empty intersections $B_i \cap C_j$ for $B_i \in \pi$ and $C_j \in \sigma$. Then $\mathrm{dit}(\pi \vee \sigma) = \mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma)$ so that $\mathrm{indit}(\pi \vee \sigma) = \mathrm{indit}(\pi) \cap \mathrm{indit}(\sigma)$. The logical entropy $h(\pi \vee \sigma)$, also the joint logical entropy $h(\pi,\ \sigma)$, is: $h(\pi \vee \sigma) = 1 - \sum_{B_i\in\pi, C_j\in\sigma}\Pr\left(B_i \cap C_j\right)^2$. This has an elegant formulation in the density matrix formalism which implies the earlier result since $\pi \vee \pi = \pi$.

Lemma 4.1. $h(\pi \vee \sigma) = 1 - tr[\rho(\pi)\rho(\sigma)]$.

Proof. The kth diagonal entry in $\rho(\pi)\rho(\sigma)$ is the scalar product $\sum_j \rho(\pi)_{kj}\rho(\sigma)_{jk}$ with $\rho(\pi)_{kj} = \sqrt{p_k p_j}$ if, $\left(u_j, u_k\right) \in \mathrm{indit}(\pi)$ and otherwise 0, and similarly for $\rho(\sigma)_{jk}$. Hence the only non-zero terms in that sum are for $\left(u_k, u_j\right) \in \mathrm{indit}(\pi) \cap \mathrm{indit}(\sigma) = \mathrm{indit}(\pi \vee \sigma)$. Hence

$$\mathrm{tr}[\rho(\pi)\rho(\sigma)] = \sum_{\left(u_j,u_k\right)\in\mathrm{indit}(\pi\vee\sigma)} p_j p_k = 1 - \sum_{\left(u_j,u_k\right)\in\mathrm{dit}(\pi\vee\sigma)} p_j p_k = 1 - h(\pi \vee \sigma) \tag{81}$$

so

$$h(\pi \vee \sigma) = 1 - \mathrm{tr}[\rho(\pi)\rho(\sigma)]. \tag{82}$$

□

In coding theory, the difference-based notion of distance between two 0,1 n-vectors is the *Hamming distance* ([65], p. 66) which is just the number of places where the corresponding entries in the two vectors are different. If we think of the 0,1 n-vectors as characteristic functions of subsets S and T of an n-element set, then the Hamming distance is the cardinality of the symmetric difference: $|S - T| + |T - S| = |S \cup T| - |S \cap T|$. This motivates the definition of the *logical distance* (or *Hamming distance*) between two partitions as: $h(\pi|\sigma) + h(\sigma|\pi) = h(\pi \vee \sigma) - m(\pi, \sigma)$, the product probability measure on the dits that in one partition but not the other. But there is the *Hilbert-Schmidt distance measure*, $\mathrm{tr}\left[(\rho - \tau)^2\right]$ between density matrices ρ and τ which does not mention logical entropy at all (see [66]). Taking the two density matrices as $\rho(\pi)$ and $\rho(\sigma)$, we have the following result that the logical (Hamming) distance between partitions is the Hilbert-Schmidt distance between the partitions.

Proposition 4.2. *tr*$\left[(\rho(\pi)-\rho(\sigma))^2\right] = h(\pi|\sigma)+h(\sigma|\pi)$.

Proof. $\mathrm{tr}\left[(\rho(\pi)-\rho(\sigma))^2\right] = \mathrm{tr}\left[\rho(\pi)^2\right] - \mathrm{tr}[\rho(\pi)\rho(\sigma)] - \mathrm{tr}[\rho(\sigma)\rho(\pi)] + \mathrm{tr}\left[\rho(\sigma)^2\right]$ so:

$$\begin{aligned}\mathrm{tr}\left[(\rho(\pi)-\rho(\sigma))^2\right] &= 2[1-\mathrm{tr}[\rho(\pi)\rho(\sigma)]] - \left(1-\mathrm{tr}\left[\rho(\pi)^2\right] - \left(1-\mathrm{tr}\left[\rho(\sigma)^2\right]\right)\right)\\ &= 2h(\pi\vee\sigma) - h(\pi) - h(\sigma) = h(\sigma|\pi)+h(\pi|\sigma).\end{aligned} \tag{83}$$

□

One point of developing these results in this classical case, is that the same theorems hold, *mutatis mutandis*, for quantum logical entropy [67].

Another set of classical results about logical entropy that extend to the quantum case are concerned with the quantum notion of (projective) measurement which is described by the Lüders mixture operation ([68], p. 279)]. For the partition $\sigma = \{C_1, ..., C_{m'}\}$ on U, let P_C be the diagonal $n \times n$ projection matrix whose diagonal entries are just the characteristic function $\chi_C(u_i)$ for $C \in \sigma$. Then the *Lüders mixture operation of performing a "σ-measurement"* on $\rho(\pi)$ is defined as: $\sum_{C\in\sigma} P_C\rho(\pi)P_C$.

Theorem 4.3. Luders mixture operation = partition join operation. $\sum_{C\in\sigma} P_C\rho(\pi)P_C = \rho(\pi\vee\sigma)$.

Proof. A non-zero entry in $\rho(\pi)$ has the form $\rho(\pi)_{jk} = \sqrt{p_jp_k}$ iff there is some block $B \in \pi$ such that $(u_j, u_k) \in B \times B$, i.e., if $u_j, u_k \in B$. The matrix operation $P_C\,\rho(\pi)$ will preserve the entry $\sqrt{p_jp_k}$ if $u_j \in C$, otherwise the entry is zeroed. And if the entry was preserved, then the further matrix operation $(P_C\rho(\pi))P_C$ will preserve the entry $\sqrt{p_jp_k}$ if $u_k \in C$, otherwise it is zeroed. Hence the entries $\sqrt{p_jp_k}$ in $\rho(\pi)$ that are preserved in $P_C\rho(\pi)P_C$ are the entries where both $u_j, u_k \in B$ for some $B \in \pi$ and $u_j, u_k \in C$. These are the entries in $\rho(\pi\vee\sigma)$ corresponding to the non-empty blocks $B \cap C$ of $\pi\vee\sigma$ for some $B \in \pi$, so summing over $C \in \sigma$ gives the result: $\sum_{C\in\sigma} P_C\rho(\pi)P_C = \rho(\pi\vee\sigma)$. □

Thus projective quantum measurement is modeled classically by the distinction-creating partition join. Hence the logical information *created* by the σ-measurement of $\rho(\pi)$ is $h(\sigma\vee\pi) - h(\pi) = h(\sigma|\pi)$. Moreover, this increase in logical entropy can be computed from the changes in the entries in the density matrices. A non-zero off-diagonal entry in a density matrix $\rho(\pi)$ indicates that the u_j and u_k for the corresponding diagonal elements must "cohere" together in the same block of π. If such a non-zero off-diagonal element of $\rho(\pi)$ was zeroed in the transition to the density matrix $\rho(\pi\vee\sigma)$ of the σ-measurement result, then it means that u_j and u_k were "decohered" by σ, i.e., were in different blocks of σ.

Corollary 4.4. *The sum of all the squares $p_j\,p_k$ of all the non-zero off-diagonal entries $\sqrt{p_jp_j}$ of $\rho(\pi)$ that were zeroed in the Lüders mixture operation that transforms $\rho(\pi)$ into $\sum_{C\in\sigma} P_C\rho(\pi)P_C = \rho(\pi\vee\sigma)$ is $h(\pi\vee\sigma) - h(\pi) = h(\sigma|\pi)$.*

Proof. All the entries $\sqrt{p_jp_j}$ that got zeroed were for ordered pair (u_j, u_k) that were indits of π but not indits of $\pi\vee\sigma$, i.e., $(u_j, u_k) \in \mathrm{indit}(\pi) \cap \mathrm{indit}(\pi\vee\sigma)^c = \mathrm{dit}(\pi)^c \cap \mathrm{dit}(\pi\vee\sigma) = \mathrm{dit}(\pi\vee\sigma) - \mathrm{dit}(\pi)$. The sum of products $p_j\,p_k$ for those pairs $(u_j,\ u_k)$ is just the product probability measure on that set $\mathrm{dit}(\pi\vee\sigma) - \mathrm{dit}(\pi)$ which is $h(\pi\vee\sigma|\pi)$. And since $\mathrm{dit}(\pi) \subseteq \mathrm{dit}(\pi\vee\sigma)$, the measure on $\mathrm{dit}(\pi\vee\sigma) - \mathrm{dit}(\pi)$ is $h(\pi\vee\sigma|\pi) = h(\pi\vee\sigma) - h(\pi) = h(\sigma|\pi)$. □

It might be noted that nothing about logical entropy was used in the definition of the Lüders mixture operation that describes the "σ-measurement of $\rho(\pi)$." Yet the logical information created by the σ-measurement of $\rho(\pi)$ is $h(\sigma|\pi)$, the logical information that is in σ over and above the information in π. And that logical entropy $h(\sigma|\pi)$ can be computed directly from the terms in the density matrix $\rho(\pi)$ that were zeroed in the Lüders operation.

Now we are ready to make the transition to quantum logical information theory where all the corresponding results will hold.

Linearizing "classical" to quantum logical entropy

One of the developers of quantum information theory, Charles Bennett, said that information was fundamentally about distinguishability.

> [Information] is the notion of distinguishability abstracted away from what we are distinguishing, or from the carrier of information. ... And we ought to develop a theory of information which generalizes the theory of distinguishability to include these quantum properties... ([69], pp. 155–157)

Given a normalized vector $|\psi\rangle$ in an n-dimensional Hilbert space V, a pure state density matrix is formed as $\rho(\psi) = |\psi\rangle\langle\psi| = |\psi\rangle(|\psi\rangle)^\dagger$ (where $()^\dagger$ is the conjugate-transpose) and mixed state density matrix is a probability mixture $\rho = \sum_i p_i\rho(\psi_i)$ of pure state density matrices. Any such density matrix always has a spectral decomposition into the form

$\rho = \sum p_i \rho(\psi_i)$ where the different vectors ψ_i and $\psi_{i'}$ are orthogonal. The general definition of the *quantum logical entropy of a density matrix* is: $h(\rho) = 1 - \mathrm{tr}[\rho^2]$, where if ρ is a pure state if and only if $\mathrm{tr}[\rho^2] = 1$ so $h(\rho) = 0$, and $\mathrm{tr}[\rho^2] < 1$ for mixed states so $1 > h(p) > 0$ for mixed states. The formula $h(\rho) = 1 - \mathrm{tr}[\rho^2]$ is hardly new. Indeed, $\mathrm{tr}[\rho^2]$ is usually called the *purity* of the density matrix so the complement $1 - \mathrm{tr}[\rho^2]$ has been called the " mixedness" ([70], p. 5) or " impurity" of the state ρ. The seminal paper of Manfredi and Feix [71] approaches the same formula $1 - \mathrm{tr}[\rho^2]$ (which they denote as S_2) from the advanced viewpoint of Wigner functions, and they present strong arguments for this notion of quantum entropy.

While that definition is an easy generalization of the classical one formulated using density matrices, our goal is to develop quantum logical entropy in a manner that brings out the analogy with classical logical entropy and relates it closely to quantum measurement.

There is a method (or what Gian-Carlo Rota would call a "yoga") to *linearize* set concepts to vector-space concepts:

> *The Yoga of Linearization*:
> Apply the set concept to a basis-set of a vector space (i.e., treat the basis set as a set universe U) and whatever is generated is the corresponding vector-space concept.

For instance, there is the classical Boolean logic of subsets, and a subset of a basis set generates a subspace so the Boolean logic of subsets linearizes to vector spaces as the logic of subspaces, and specializing to Hilbert spaces yields the usual quantum logic of subspaces [72].

In view of the category-theoretic duality of subsets of a set and partitions on a set (e.g., the image subset of the codomain and the inverse-image partition on the domain of a set function), there is a dual "classical" logic of partitions on a set ([3, 4]). To linearize the set concept of a partition to vector spaces, one considers a set partition on a basis set of a vector space and then sees what it generates. Each block generates a subspace and the set of subspaces corresponding to the blocks form a direct-sum decomposition (DSD) of the vector space. A *direct-sum decomposition* of a vector space V is a set $\{V_i\}_{i\in I}$ of subspaces such that $V_i \cap \sum_{i'\neq i} V_i = \{0\}$ (where $\sum_{i'\neq i} V_i$ is the subspace generated by those $V_{i'}$, and $\{0\}$ is the zero subspace), and which span the space V and is written $V = \oplus_{i\in I} V_i$. Then every non-zero vector $v \in V$ is a unique sum of vectors from the subspaces $\{V_i\}$. That is the vector-space version of characterizing a partition π on a set U as a collection of subsets B_i (blocks) of U such that every non-empty subset $S \subseteq U$ is uniquely expressed as a union of subsets of the blocks, i.e., $S \cap B_i$ for $B_i \in \pi$. Hence the logic of partitions linearizes to the dual logic of DSDs of a vector space which specialized to Hilbert spaces yields the quantum logic of DSDs [74] dual to the usual Birkhoff–von-Neumann quantum logic of subspaces.

Another basic set concept is the notion of a numerical attribute $f : U \to \mathbb{K}$ that evaluates the points of U in a field $\mathbb{K}$. Taking U to be a basis set of a vector space V over the field $\mathbb{K}$, the corresponding vector-space notion that can be seen as generated is the notion of a *diagonalizable linear operator* $F : V \to V$ defined by $Fu = f(u)u$ linearly extended to V. The values of f linearize to the eigenvalues of F, the constant sets of f linearize to the eigenvectors of F, and the set of constant sets for a specific value linearizes to the eigenspace of eigenvectors for that eigenvalue. For instance, if we let "rS" stand for assigning the value r to each element of the subset $S \subseteq U$, then the set version of the eigenvector equation $Fv = \lambda v$ is $f(S) = rS$.

The Cartesian product of two basis sets of two vector spaces (same base field) generates the tensor product of the two vector spaces – so the linearization of the direct or Cartesian product of sets is not the direct product (as might be suggested by category theory) but the tensor product of vector spaces. And the cardinality of sets linearizes to the dimension of vector spaces and so forth as illustrated in Table 4. Those examples show how the set-based classical logical information theory will linearize to vector spaces and particularly to Hilbert spaces for the logical version of quantum information theory.[10]

Let $F : V \to V$ be a self-adjoint (or Hermitian) operator (observable) on a n-dimensional Hilbert space V with the real eigenvalues $\phi_1,\ldots,\phi_I$, and let $U=\{u_1,\ldots,u_n\}$ be an orthonormal (ON) basis of eigenvectors of F. The quantum version of a "dit" is a "qudit." A *qudit* is *defined by the DSD of eigenspaces of an observable*, just as classically, a distinction or dit is defined by the partition $\{f^{-1}(r)\}_{r\in f(U)}$ determined a numerical attribute $f : U \to \mathbb{R}$. Then, there is a set partition $\pi = \{B_i\}_{i=1,\ldots,I}$ on the ON basis U so that B_i is a basis for the eigenspace of the eigenvalue ϕ_i and $|B_i|$ is the " multiplicity" (dimension of the eigenspace) of the eigenvalue ϕ_i for $i = 1,\ldots,I$. The eigenspaces V_i generated by the blocks B_i for the eigenvalues ϕ_i form a direct-sum decomposition of V. Note that the real-valued numerical attribute or eigenvalue function $f : U \to \mathbb{R}$ that takes each eigenvector in $u_j \in B_i \subseteq U$ to its eigenvalue ϕ_i so that $f^{-1}(\phi_i) = B_i$ contains all the information in the self-adjoint operator $F : V \to V$ since F can be reconstructed by defining it on the basis U as $Fu_j = f(u_j)u_j$. The important information-theoretic aspect of the eigenvalues is not their numerical value but when they are the same or different, and that information is there in the eigenspaces $\{V_i\}_{i\in I}$ of the direct-sum decomposition.[11]

[10] Since set-concepts can be formulated in vector spaces over $\mathbb{Z}_2$, that means there is a pedagogical or "toy" model of quantum mechanics over $\mathbb{Z}_2$, i.e., over sets [73].

[11] That is why the quantum logic of DSDs [74] is essentially the quantum logic of observables – in much the same sense that the logic of partitions on U is essentially the logic of numerical attributes $f : U \to \mathbb{R}$ on U.

Table 4. Linearization of set concepts to corresponding vector-space concepts.

Set concept	Vector-space concept
Universe set U	Basis set of a space V
Cardinality of a set U	Dimension of a space V
Subset of a set U	Subspace of a space V
Partition of a set U	Direct-sum decomposition of a space V
Numerical attribute $f: U \to \mathbb{K}$	Diagonalizable linear op. $F: V \to V$
Value r in image $f(U)$ of f	Eigenvalue λ_i of F
Constant set S of f	Eigenvector v of F
Set of constant r-sets $\wp(f^{-1}(r))$	Eigenspace V_i of λ_i
Direct product of sets	Tensor product of spaces
Elements $(u_k, u_{k'})$ of $U \times U$	Basis vectors $u_k \otimes u_{k'}$ of $V \otimes V$

Classically, a *dit of the partition* $\{f^{-1}(\phi_i)\}_{i\in I}$ on U, defined by $f: U \to \mathbb{R}$, is a pair $(u_k, u_{k'}) \in U \times U$ of points in distinct blocks of the partition, i.e., $f(u_k) \neq f(u_{k'})$. Hence, a *qudit of* F is a pair $(u_k, u_{k'})$ (interpreted as $u_k \otimes u_{k'} \in V \otimes V$) of vectors in the eigenbasis distinguished by F, i.e., $f(u_k) \neq f(u_{k'})$ for the eigenvalue function $f: U \to \mathbb{R}$. Let $G: V \to V$ be another self-adjoint operator on V, which commutes with F so that we may then assume that U is an orthonormal basis of simultaneous eigenvectors of F and G ([75], p. 177). The assumption that F and G commute plays the role of considering partitions $\pi = f^{-1}$ for $f: U \to \mathbb{R}$ and $\sigma = g^{-1}$ for $g: U \to \mathbb{R}$ being defined on the same universe U. Let $\{\gamma_j\}_{j\in J}$ be the set of eigenvalues of G, and let $g: U \to \mathbb{R}$ be the eigenvalue function so a pair $(u_k, u_{k'})$ is a *qudit of* G if $g(u_k) \neq g(u_{k'})$, i.e., if the two eigenvectors have distinct eigenvalues of G.

As Kolmogorov suggested:

> Information theory must precede probability theory, and not be based on it. By the very essence of this discipline, the foundations of information theory have a finite combinatorial character. ([76], p. 39)

In classical logical information theory, information is defined prior to probabilities by certain subsets (e.g., ditsets and differences between and intersections of ditsets) or, in the quantum case, quantum information is defined by certain subspaces prior to the introduction of any probabilities (unlike the case with Shannon or von Neumann entropies). Since the transition from classical to quantum logical information theory is straightforward, it will be first presented in a table (which does not involve any probabilities), where the qudits $(u_k, u_{k'})$ are interpreted as $u_k \otimes u_{k'}$. The *qudit space*, the vector-space analogue of the ditset, associated with F (the vector-space analogue of $f: U \to \mathbb{R}$) is the subspace $[\text{qudit}(F)] \subseteq V \otimes V$ generated by the qudits $u_k \otimes u_{k'}$ of F.

If $F = \lambda I$ is a scalar multiple of the identity I (the vector-space analogue of a constant function $f: U \to \mathbb{R}$), then it has no qudits, so its qudit subspace $[\text{qudit}(\lambda I)]$ is the zero subspace (the analogue of the empty ditset of the indiscrete partition). The Common Dits Theorem says that any two non-empty ditsets have a non-zero intersection. In the quantum case, this means any two non-zero qudit spaces $[\text{qudit}(F)]$ and $[\text{qudit}(G)]$ for commuting F and G have a non-zero intersection, i.e., have a non-zero mutual information space. That is, for commuting F and G, there are always two simultaneous eigenvectors u_k and $u_{k'}$ that have different eigenvalues both by F and by G.

The observables do not provide the point probabilities in a measurement; the probabilities come from the pure (normalized) state ψ being measured. Let $|\psi\rangle = \sum_{j=1}^{n} \langle u_j|\psi\rangle |u_j\rangle = \sum_{j=1}^{n} \alpha_j |u_j\rangle$ be the resolution of $|\psi\rangle$ in terms of the orthonormal basis $U = \{u_1, \ldots, u_n\}$ of simultaneous eigenvectors for F and G. Then, $p_j = \alpha_j \alpha_j^*$ (α_j^* is the complex conjugate of α_j) for $j = 1, \ldots, n$ are the point probabilities on U, and the pure state density matrix $\rho(\psi) = |\psi\rangle\langle\psi|$ (where $\langle\psi| = |\psi\rangle^\dagger$ is the conjugate-transpose) has the entries: $\rho_{jk}(\psi) = \alpha_j \alpha_k^*$, so the diagonal entries $\rho_{jj}(\psi) = \alpha_j \alpha_j^* = p_j$ are the point probabilities. Then we have Table 5 giving the remaining parallel development with the probabilities provided by the pure state ψ where we write $\rho(\psi)^\dagger \rho(\psi)$ as $\rho(\psi)^2$.

The definition of quantum logical entropy

$$h(F: \psi) = tr\left[P_{[\text{qudit}(F)]} \rho(\psi) \otimes \rho(\psi)\right] \tag{84}$$

is just the quantum version of the formulation of the classical logical entropy

$$h(f^{-1}) = h(\pi) = \sum_{(u_i, u_j) \in \text{dit}(\pi)} p_i p_j = tr\left[P_{\text{dit}(\pi)} \rho(\pi) \otimes \rho(\pi)\right] \tag{85}$$

Table 5. Ditsets and qudit subspaces without probabilities.

Classical logical information	Quantum logical information
$f, g : U \to \mathbb{R}$	Commuting self-adjoint ops. F, G
$U = \{u_1, ..., u_n\}$	ON basis simultaneous eigenvectors F, G
Values $\{\phi_i\}_{i\in I}$ of f	Eigenvalues $\{\phi_i\}_{i\in I}$ of F
Values $\{\gamma_j\}_{j\in J}$ of g	Eigenvalues $\{\gamma_j\}_{j\in J}$ of G
Partition $\{f^{-1}(\phi_i)\}_{i\in I}$	Eigenspace DSD of F
Partition $\{g^{-1}(\gamma_j)\}_{j\in J}$	Eigenspace DSD of G
dits of $\pi : (u_k, u_{k'}) \in U^2$, $f(u_k) \neq f(u_{k'})$	Qudits of F: $u_k \otimes u_{k'} \in V \otimes V$, $f(u_k) \neq f(u_{k'})$
dits of $\sigma : (u_k, u_{k'}) \in U^2$, $g(u_k) \neq g(u_{k'})$	Qudits of G: $u_k \otimes u_{k'} \in V \otimes V$, $g(u_k) \neq g(u_{k'})$
$\mathrm{dit}(\pi) \subseteq U \times U$	[Qudit(F)] = subspace gen. by qudits of F
$\mathrm{dit}(\sigma) \subseteq U \times U$	[Qudit(G)] = subspace gen. by qudits of G
$\mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma) \subseteq \mathrm{U} \times \mathrm{U}$	[Qudit(F) ∪ Qudit(G)] ⊆ $V \otimes V$
$\mathrm{dit}(\pi) - \mathrm{dit}(\sigma) \subseteq U \times U$	[Qudit(F) − Qudit(G)] ⊆ $V \otimes V$
$\mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma) \subseteq U \times U$	[Qudit(F) ∩ Qudit(G)] ⊆ $V \otimes V$

for $f : U \to \mathbb{R}$ with the point probabilities $(p_1,..., p_n)$ on U and thus $p \times p$ on $U \times U$. The tensor product $\rho(\psi) \otimes \rho(\psi)$ is an $n^2 \times n^2$ matrix with the diagonal entries $(\rho(\psi) \otimes \rho(\psi))_{(j,k),(j,k)} = \rho(\psi)_{jj}\rho(\psi)_{kk} = p_j p_k$ where $p_j = \alpha_j \alpha_j^*$ for $|\psi\rangle = \sum_{j=1}^n \langle u_j|\psi\rangle |u_j\rangle = \sum_{j=1}^n \alpha_j |u_j\rangle$ where $U = \{u_1,..., u_n\}$ is an ON basis of eigenvectors of the observable F. The $n^2 \times n^2$ diagonal projection matrix $P_{[\mathrm{qudit}(F)]}$ has a diagonal element $\left(P_{[\mathrm{qudit}(F)]}\right)_{(j,k)(j,k)} = 1$ if $u_j \otimes u_k \in [\mathrm{qudit}(F)]$. i.e., if the eigenvectors u_j and u_k have different eigenvalues, and 0 otherwise. Hence the product $P_{[\mathrm{qudit}(F)]}\rho(\psi) \otimes \rho(\psi)$ will pick out the products $p_j\, p_k$ for $u_j \otimes u_k \in [\mathrm{qudit}(F)]$ and the trace will sum them. Hence we have the result that:

$$\begin{aligned} h(F : \psi) &= \mathrm{tr}\left[\mathrm{P}_{[\mathrm{qudit}(F)]}\rho(\psi) \otimes \rho(\psi)\right] \\ &= \sum_{j,k} \left\{p_j p_k : u_j \otimes u_k \in [\mathrm{qudit}(F)]\right\} = \sum_{j,k} \left\{p_j p_k : f\left(u_j\right) \neq f(u_k)\right\} \end{aligned} \tag{86}$$

where $f : U \to \mathbb{R}$ is the eigenvalue function taking each eigenvector to its eigenvalue.

With those preliminaries, the definitions in Table 6 might be better motivated and the statements clearer.

Some basic results about quantum logical entropy

A self-adjoint operator F on V, i.e., an observable, alone defines the eigenvalue partition $f^{-1} = \{f^{-1}(\phi_i)\}_{i\in I}$ on a basis U of ON eigenvectors for F. But the points have no probabilities associated with them. The probabilities are supplied by a normalized vector $|\psi\rangle \in V$ as $p_i = \alpha_i \alpha_i^*$ for $\alpha_i = \langle u_i|\psi\rangle$. Then we have a completely classical situation, a set partition f^{-1} on a set U with point probabilities provided by $|\psi\rangle$ which will be denoted $\pi(F : \psi)$. Hence that partition will have a (classical) logical entropy $h(\pi(F : \psi))$. Since the blocks in that $\pi(F : \psi)$ partition on U are the sets of basis vectors each for a certain eigenvalue, the probabilities for those blocks are $\sum_j \left\{p_j : f\left(u_j\right) = \phi_i\right\} = \Pr(f^{-1}(\phi_i)) = \Pr(B_i)$ for $i = 1,..., I$. Hence we have:

$$\begin{aligned} h(\pi(F : \psi)) &= 1 - \sum_{i\in I} \Pr(f^{-1}(\phi_i))^2 = 1 - \sum_{i\in I} \left(\sum_j \left\{p_j : f\left(u_j\right) = \phi_i\right\}\right)^2 \\ &= 1 - \sum_i \left(\sum_{f\left(u_j\right)=\phi_i} p_j^2 + \sum_{j\neq k} \left\{p_j p_k : f\left(u_j\right) = \phi_i = f(u_k)\right\}\right) \\ &= \sum_{j,k} \left\{p_j p_k : f\left(u_j\right) \neq f(u_k)\right\} = h(F : \psi) = tr\left[P_{[\mathrm{qudit}(F)]}\rho(\psi) \otimes \rho(\psi)\right] \end{aligned} \tag{87}$$

And there is another way to arrive at this logical entropy, namely perform the F-measurement on the pure state density matrix $\rho(\psi)$. The results of the F-measurement is given by the Lüders mixture operation ([68], p. 279) on the density matrix $\rho(\psi)$. The block $B_i \in \pi(F : \psi)$ generates the eigenspace V_i corresponding to the eigenvalue ϕ_i so P_{V_i} is the projection matrix to that eigenspace for $i = 1, \ldots, I$. Then the Lüders mixture operation, representing the F-measurement of ψ, gives the mixed state density matrix:

$$\hat{\rho}(\psi) = \sum_{i\in I} P_{V_i}\rho(\psi)P_{V_i}. \tag{88}$$

Table 6. Probabilities applied to ditsets and qudit spaces.

"Classical" Logical Entropy	Quantum Logical Entropy
Pure state density matrix, e.g., $\rho(\mathbf{0}_U)$	Pure state density matrix $\rho(\psi)$
$U = \{u_1, \ldots, u_n\}$	ON basis simultaneous eigenvectors F, G
$p \times p$ *on* $U \times U$	$\rho(\psi)\otimes\rho(\psi)$ on $V \otimes V$
$h(\mathbf{0}_U) = 1 - \mathrm{tr}[\rho(\mathbf{0}_U)^2] = 0$	$h(\rho(\psi)) = 1 - \mathrm{tr}[\rho(\psi)^2] = 0$
$h(\pi) = p \times p(\mathrm{dit}(\pi))$	$h(F{:}\psi) = \mathrm{tr}[P_{[\mathrm{qudit}(F)]}\ \rho(\psi) \otimes \rho(\psi)]$
$h(\pi, \sigma) = p \times p(\mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma))$	$h(F,G{:}\psi) = \mathrm{tr}[P_{[\mathrm{qudit}(F)\ \cup\ \mathrm{qudit}(G)]}\ \rho(\psi)\otimes\rho(\psi)]$
$h(\pi\|\sigma) = p \times p(\mathrm{dit}(\pi) - \mathrm{dit}(\sigma))$	$h(F\|G{:}\psi) = \mathrm{tr}[P_{[\mathrm{qudit}(F)\ -\ \mathrm{qudit}(G)]}\ \rho(\psi) \otimes \rho(\psi)]$
$m(\pi, \sigma) = p \times p(\mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma))$	$m(F,G{:}\psi) = \mathrm{tr}[P_{[\mathrm{qudit}(F)\ \cap\ \mathrm{qudit}(G)]}\ \rho(\psi) \otimes \rho(\psi)]$
$h(\pi) = h(\pi\|\sigma) + m(\pi,\sigma)$	$h(F{:}\psi) = h(F\|G{:}\psi) + m(F,G{:}\psi)$
$h(\pi)$ = 2-draw prob. diff. f-values	$h(F{:}\psi)$ = 2-meas. prob. diff. F-eigenvalues
$\rho(\pi) = \sum_i P_{B_i}\rho(\mathbf{0}_U)P_{B_i}$	$\hat{\rho}(\psi) = \sum_i P_{V_i}\rho(\psi)P_{V_i}$
$h(\pi) = 1 - \mathrm{tr}[\rho(\pi)^2]$	$h(F:\psi) = 1 - tr\left[\hat{\rho}(\psi)^2\right]$
$h(\pi)$ = sum sq. zeroed $\rho(\mathbf{0}_U) \rightsquigarrow \rho(\pi)$	$h(F{:}\psi)$ = sum ab. sq. zeroed $\rho(\psi) \rightsquigarrow \hat{\rho}(\psi)$

To show that $h(\hat{\rho}(\psi)) = 1 - \mathrm{tr}\left[\hat{\rho}(\psi)^2\right] = h(\pi(F:\psi))$ for $\hat{\rho}(\psi) = \sum_{i=1}^{I} P_{V_i}\rho(\psi)P_{V_i}$, we need to compute $tr\left[\hat{\rho}(\psi)^2\right]$. An off-diagonal element in $\rho_{jk}(\psi) = \alpha_j\alpha_k^*$ of $\rho(\psi)$ survives (i.e., is not zeroed and has the same value) the Lüders operation if and only if $f(u_j) = f(u_k)$. Hence, the j-th diagonal element of $\hat{\rho}(\psi)^2$ is:

$$\sum_{k=1}^{n}\left\{\alpha_j^*\alpha_k\alpha_j\alpha_k^* : \phi(u_j) = \phi(u_k)\right\} = \sum_{k=1}^{n}\left\{p_jp_k : f(u_j) = f(u_k)\right\} = p_j\Pr(B_i) \tag{89}$$

where $u_j \in B_i$. Then, grouping the j-th diagonal elements for $u_j \in B_i$ gives $\sum_{u_j} p_j\Pr(B_i) = \Pr(B_i)^2$. Hence, the whole trace is: $\mathrm{tr}\left[\hat{\rho}(\psi)^2\right] = \sum_{i=1}^{I}\Pr(B_i)^2$, and thus:

$$h(\hat{\rho}(\psi)) = 1 - tr\left[\hat{\rho}(\psi)^2\right] = 1 - \sum_{i=1}^{I}\Pr(B_i)^2 = h(F:\psi). \tag{90}$$

This completes the proof of the following theorem which shows the different ways to characterize $h(F{:}\,\psi)$.

Theorem 4.5. $h(F:\psi) = h(\pi(F:\psi)) = h(\hat{\rho}(\psi))$.

Like the classical join operation on partitions, *quantum measurement creates distinctions*, i.e., turns coherences into "decoherences,"[12] which, classically, is the operation of distinguishing elements by classifying them according to some attribute like classifying the faces of a die by their parity. The fundamental theorem about quantum logical entropy and projective measurement, in the density matrix version, shows how the quantum logical entropy created (starting with $h(\rho(\psi)) = 0$ for the pure state ψ) by the measurement can be computed directly from the coherences of $\rho(\psi)$ that are decohered in $\hat{\rho}(\psi)$.

Theorem 4.6. *Fundamental theorem about quantum measurement and logical entropy.*

The increase in quantum logical entropy, $h(\hat{\rho}(\psi)) = h(F:\psi)$ due to the F-measurement of the pure state ψ is the sum of the absolute squares of the non-zero off-diagonal terms (coherences) in $\rho(\psi)$ (represented in an ON basis of F-eigenvectors) that are zeroed ("decohered") in the post-measurement Lüders mixture density matrix $\hat{\rho}(\psi) = \sum_{i=1}^{I} P_{V_i}\rho(\psi)P_{V_i}$.

Proof. $h(\hat{\rho}(\psi)) - h(\rho(\psi)) = \left(1 - \mathrm{tr}\left[\hat{\rho}(\psi)^2\right]\right) - \left(1 - \mathrm{tr}\left[\rho(\psi)^2\right]\right) = \sum_{j,k}\left(\left|\rho_{jk}(\psi)\right|^2 - \left|\hat{\rho}(\psi)\right|^2\right)$ since $tr[\rho^2] = \sum_{i,j}\left|\rho_{ij}\right|^2$ is the sum of the absolute squares of all the elements of ρ ([78], p. 77). If u_j and u_k are a qudit of F, then and only then are the corresponding off-diagonal terms zeroed by the Lüders mixture operation $\sum_{i=1}^{I} P_{V_i}\rho(\psi)P_{V_i}$ to obtain $\hat{\rho}(\psi)$ from $\rho(\psi)$. □

Example: For a simple quantum example, consider a system with two spin-observable σ eigenstates $|\uparrow\rangle$ and $|\downarrow\rangle$ (like electron spin up or down along the z-axis) where the given normalized superposition state is $|\psi\rangle = \alpha_\uparrow|\uparrow\rangle + \alpha_\downarrow|\downarrow\rangle = \begin{bmatrix}\alpha_\uparrow\\ \alpha_\downarrow\end{bmatrix}$ so the density matrix is $\rho(\psi) = \begin{bmatrix} p_\uparrow & \alpha_\uparrow\alpha_\downarrow^* \\ \alpha_\downarrow\alpha_\uparrow^* & p_\downarrow\end{bmatrix}$ where $p_\uparrow = \alpha_\uparrow\alpha_\uparrow^*$ and $p_\downarrow = \alpha_\downarrow\alpha_\downarrow^*$. Using the Lüders mixture operation, the measurement of that spin-observable σ goes from the pure state $\rho(\psi)$ to

[12] This notion of "decoherence" is used in an older sense, not the recent sense given by the work of Zurek [77] and others.

$$P_\uparrow \rho(\psi) P_\uparrow + P_\downarrow \rho(\psi) P_\downarrow = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} p_\uparrow & \alpha_\uparrow \alpha_\downarrow^* \\ \alpha_\downarrow \alpha_\uparrow^* & p_\downarrow \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p_\uparrow & \alpha_\uparrow \alpha_\downarrow^* \\ \alpha_\downarrow \alpha_\uparrow^* & p_\downarrow \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} p_\uparrow & 0 \\ 0 & p_\downarrow \end{bmatrix} = \hat{\rho}(\psi) \quad (91)$$

The logical entropy of any pure state such as $\rho(\psi)$ is 0. The logical entropy of $\hat{\rho}(\psi)$ is $h(\hat{\rho}(\psi)) = 1 - tr\left[\hat{\rho}(\psi)^2\right] = 1 - p_\uparrow^2 - p_\downarrow^2$. The entries that were zeroed in the Lüders mixture operation were the two off-diagonal elements $\alpha_\uparrow \alpha_\downarrow^*$ and $\alpha_\downarrow \alpha_\uparrow^*$ so the sum of their absolute squares is $2\alpha_\uparrow \alpha_\downarrow^* \alpha_\downarrow \alpha_\uparrow^* = 2p_\uparrow p_\downarrow$ which equals $1 - p_\uparrow^2 - p_\downarrow^2$ since $1 = \left(p_\uparrow + p_\downarrow\right)^2 = p_\uparrow^2 + p_\downarrow^2 + 2p_\uparrow p_\downarrow$.

Conclusions

The underlying thesis is that information is defined in terms of distinctions, differences, distinguishability, and diversity – or, with similar uses of the di-prefix (which means "two"), discriminations, divisions, or differentiations. Yet those are all vague concepts so this notion of information-as-distinctions is made precise using the basic mathematical concept that represents differences and non-differences (or equivalences), namely partitions (including the inverse-image partitions of random variables). The elements in the same block of a partition are similar or equivalent (block = equivalence class), and the ordered pairs of elements in *different* blocks are the distinctions or dits. Hence logical entropy measures information-as-distinctions by the probability measure on distinctions, so the logical entropy of a partition is the probability that a distinction of the partition is obtained in two independent draws from the underling universe set of elements. This notion of information-as-distinctions then encompasses the Shannon notion of entropy as the average minimum number of binary partitions (bits) that have to be joined to make the same distinctions of the partition. Moreover, there is the dit-bit transform that derives all of Shannon's definitions of entropy, joint entropy, conditional entropy, and mutual information from the corresponding definitions of logical entropy that are based on logical entropy being defined as a (probability) measure in the sense of measure theory. A few applications were discussed; distinguishing the Boltzmann and Shannon entropies, developing the MaxEntropy method with logical entropy, and showing how the metrical notion of logical entropy gives the notion of variance in statistical theory.

There is a method, linearization, to lift set-based concepts to the corresponding vector-space concepts, and that provides the method to develop the corresponding quantum notions from the "classical" or non-quantum notions of logical entropy. There are two equivalent formulations of quantum mechanics; one using wave functions and the other using density matrices ([79], p. 102). But only one of those formulations maps naturally to the mathematics of partitions, namely the density matrix formulation.

At the beginning of our presentation, density matrices were foreshadowed by the box diagrams representing logical entropy. The box diagrams led to the incidence matrices for indit(π), or the complementary ones for dit(π), and then point probabilities are introduced into the matrices so that when normalized by their trace, the matrices are density matrices. In that manner, a reformulation of the classical logical entropy framework is first presented using density matrices over the real numbers to foreshadow the later quantum results over the complex numbers. Every density matrix over the complex numbers has a spectral decomposition into a probability mixture of orthogonal pure states which correspond classically to the disjoint blocks and block probabilities of a partition.

The fundamental theorem for logical entropy and measurement shows there is a simple, direct and quantitative connection between density matrices and logical entropy. The theorem directly connects the changes in the density matrix due to a projective measurement (sum of absolute squares of zeroed off-diagonal terms) with the increase in logical entropy due to the F-measurement $h(F:\psi) = h(\hat{\rho}(\psi))$ (where $h(\rho(\psi)) = 0$ for the pure state ψ). Moreover, the quantum logical entropy has a simple " two-draw probability" interpretation, i.e., $h(F:\psi) = h(\hat{\rho}(\psi))$ is the probability that two independent F-measurements of ψ will yield distinct F-eigenvalues, i.e., will yield a qudit of F. In contrast, the von Neumann entropy has no such simple interpretation, and there seems to be no such intuitive connection between pre- and post-measurement density matrices and von Neumann entropy, although von Neumann entropy also increases in a projective measurement ([79], Thm. 11.9, p. 515).

This direct quantitative connection between state discrimination and quantum logical entropy reinforces the judgment of Boaz Tamir and Eliahu Cohen [66, 80] that quantum logical entropy is a natural and informative entropy concept for quantum mechanics.

> We find this framework of partitions and distinction most suitable (at least conceptually) for describing the problems of quantum state discrimination, quantum cryptography and in general, for discussing quantum channel capacity. In these problems, we are basically interested in a distance measure between such sets of states, and this is exactly the kind of knowledge provided by logical entropy (reference to [81]). ([80], p. 1])

In summary, the basic idea of information as distinctions, differences, distinguishability, and diversity is naturally quantified at the "classical" level in terms of logical entropy and then naturally linearized to the quantum notion of logical entropy.[13]

Conflicts of interest

The author declares no conflict of interest.

Funding

This research did not receive any specific funding.

References

1. Birkhoff G (1948), Lattice theory, American Mathematical Society, New York.
2. Grätzer G (2003), General Lattice Theory, 2nd edn., Birkhäuser Verlag, Boston.
3. Ellerman D (2010), The logic of partitions: introduction to the dual of the logic of subsets. Rev Symb Log 3, 287–350. https://doi.org/10.1017/S1755020310000018.
4. Ellerman D (2014), An introduction to partition logic. Log J IGPL 22, 94–125. https://doi.org/10.1093/jigpal/jzt036.
5. Lawvere FW, Rosebrugh R (2003), Sets for mathematics, Cambridge University Press, Cambridge, MA.
6. Rota G-C (2001), Twelve problems in probability no one likes to bring up, in: H. Crapo, D. Senato (Eds.), Algebraic combinatorics and computer science: a tribute to Gian-Carlo Rota, Springer, Milano, pp. 57–93.
7. Rota G-C (1998) Probability Vol. I & II: The guidi notes, MIT Copy Services, Cambridge, MA.
8. Ellerman D (2021), The logical theory of canonical maps: the elements and distinctions analysis of the morphisms, duality, canonicity and universal constructions in Set. https://ArXiv.org, https://arxiv.org/abs/2104.08583.
9. Halmos PR (1974), Measure theory, Springer-Verlag, New York.
10. Rao KPSB, Rao MB (1983), Theory of charges: a study of finitely additive measures, Academic Press, London.
11. Wilkins J (1707), Mercury or the secret and swift messenger, London. Original in 1641.
12. Gleick J (2011), The information: a history, a theory, a flood, Pantheon, New York.
13. Bateson G (1979), Mind and nature: a necessary unity, Dutton, New York.
14. Gini C (1912), Variabilità e mutabilità, Tipografia di Paolo Cuppini, Bologna.
15. Friedman WF (1922), The index of coincidence and its applications in cryptography, Riverbank Laboratories, Geneva IL.
16. Kullback S (1976), Statistical methods in cryptanalysis, Aegean Park Press, Walnut Creek, CA.
17. Rejewski M (1981), How Polish mathematicians deciphered the enigma. IEEE Ann Hist Comput 3, 213–234.
18. Simpson EH (1949), Measurement of diversity. Nature 163, 688.
19. Ricotta C, Szeidl L (2006), Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. Theor Popul Biol 70, 237–243. https://doi.org/10.1016/j.tpb.2006.06.003.
20. Nei M (1973), Analysis of Gene Diversity in subdivided populations. Proc Nat Acad Sci USA 70, 3321–3323.
21. Good IJ (1979), A.M. Turing's statistical work in World War II. Biometrika 66, 393–396.
22. Good IJ (1982), Comment (on Patil and Taillie: diversity as a concept and its measurement). J Am Stat Assoc 77, 561–563.
23. Stigler SM (1999), Statistics on the table, Harvard University Press, Cambridge.
24. Hirschman AO (1945), National power and the structure of foreign trade, University of California Press, Berkeley.
25. Herfindahl OC (1950), Concentration in the US Steel Industry, Unpublished Doctoral Dissertation, Columbia University.
26. Rao CR (1982), Diversity and dissimilarity coefficients: a unified approach. Theor Popul Biol 21, 24–43.
27. Havrda J, Charvat F (1967), Quantification methods of classification processes: concept of structural α-entropy. Kybernetika (Prague) 3, 30–35.
28. Tsallis C (1988), Possible generalization for Boltzmann-Gibbs statistics. J Stat Phys 52, 479–487.
29. Brukner Č, Zeilinger A (2000), Operationally invariant information in quantum measurements. https://ArXiv.org, https://arxiv.org/abs/quant-ph/0005084, 19 May 2000.
30. Brukner C, Zeilinger A (2003), Information and fundamental elements of the structure of quantum theory, in: L. Castell, O. Ischebeck (Eds.), Time, quantum and information, Springer-Verlag, Berlin, pp. 323–354.
31. Shannon CE (1948), A mathematical theory of communication. Bell Syst Tech J 27, 379–423, 623–656.
32. Shannon CE, Weaver W (1964), The mathematical theory of communication, University of Illinois Press, Urbana.
33. Shannon CE (1993), The Bandwagon, in: N.J.A. Sloane, A.D. Wyner (Eds.), Claude E. Shannon: Collected Papers, IEEE Press, Piscataway, NJ, p. 462.
34. Tribus M (1978), Thirty years of information theory, in: R.D. Levine, M. Tribus (Eds.), The maximum entropy formalism, MIT, Cambridge, MA, pp. 1–14.
35. Shannon CE (1993), Some topics in information theory, in: N.J.A. Sloane, A.D. Wyner (Eds.), Claude E. Shannon: Collected Papers, IEEE Press, Piscataway, NJ, pp. 458–459.
36. Ramshaw JD (2018), The Statistical Foundations of Entropy, World Scientific Publishing, Singapore.
37. Lewis GN (1930), The Symmetry of Time in Physics. Science 71, 569–577.
38. Brillouin L (1962), Science and Information Theory, Academic Press, New York.

[13] For further developments beyond the scope of this paper see [67, 82, 83], and the other papers in this issue.

39. Aczel J, Daroczy Z (1975), On Measures of Information and Their Characterization, Academic Press, New York.
40. Campbell LL (1965), Entropy as a Measure. IEEE Trans Inform Theory IT-11, 112–114.
41. Doob JL (1994), Measure Theory, Springer Science+Business Media, New York.
42. Polya G, Szego G (1998), Problems and Theorems in Analysis, Vol. II, Springer-Verlag, Berlin.
43. Hu KT (1962), On the amount of information. Probability Theory and Its Applications 7, 439–447. https://doi.org/10.1137/1107041.
44. Ryser HJ (1963), Combinatorial Mathematics, Mathematical Association of America, Washington DC.
45. Takacs L (1967), On the method of inclusion and exclusion. J Am Stat Assoc 62, 102–113. https://doi.org/10.1080/01621459.1967.10482891.
46. Cover T, Thomas J (2006), Elements of information theory, 2nd edn., John Wiley and Sons, Hoboken, NJ.
47. Csiszar I, Körner J (1981), Information theory: coding theorems for discrete memoryless systems, Academic Press, New York.
48. Wilson RJ (1972), Introduction to graph theory, Longman, London.
49. Rozeboom WW (1968), The theory of abstract partials: an introduction. Psychometrika 33, 133–167.
50. McGill WJ (1954), Multivariate information transmission. Trans IRE Prof Group Inform Theory 4, 93–111. https://doi.org/10.1109/TIT.1954.1057469.
51. Fano RM (1961), Transmission of Information, MIT Press, Cambridge, MA.
52. Yeung RW (1991), A new outlook on Shannon's information measures. IEEE Trans on Information Theory 37, 466–474. https://doi.org/10.1109/18.79902.
53. MacKay DJC (2003), Information theory, inference, and learning algorithms, Cambridge University Press, Cambridge, UK.
54. Atkins P, de Paula J, Keeler J (2018), Atkins' physical chemistry, 11th edn., Oxford University Press, Oxford UK.
55. Johnson E (2018), Anxiety and the equation: Understanding Boltzmann's entropy, MIT Press, Cambridge, MA.
56. Jaynes ET (2003), Probability theory: The logic of science, Cambridge University Press, Cambridge, UK.
57. Kaplan W (1999), Maxima and minima with applications: practical optimization and duality, John Wiley & Sons, New York.
58. Best MJ (2017), Quadratic programming with computer programs, CRC Press, Boca Raton FL.
59. Jaynes ET (1978), Where do we stand on maximum entropy? in: R.D. Levine, M. Tribus (Eds.), The Maximum Entropy Formalism, MIT, Cambridge, MA, pp. 15–118.
60. Papoulis A (1990), Probability and statistics, Prentice-Hall, Englewood Cliffs, NJ.
61. Dantzig GB (1963), Linear programming and extensions, Princeton University Press, Princeton.
62. Kullback S, Leibler RA (1951), On information and sufficiency. Ann Math Stat 22, 79–86. https://doi.org/10.1214/aoms/1177729694.
63. Rao CR (2010), Quadratic entropy and analysis of diversity. Sankhyā Indian J Stat 72-A, 70–80.
64. Zhang Y, Wu H, Cheng L (2012), Some new deformation formulas about variance and covariance, in: Proceedings of 2012 International Conference on Modelling, Identification and Control (ICMIC2012), pp. 987–992.
65. McEliece RJ (1977), The theory of information and coding: a mathematical framework for communication (Encyclopedia of Mathematics and its Applications, Vol. 3). Addison-Wesley, Reading, MA.
66. Tamir B, Cohen E (2015), A Holevo-type bound for a Hilbert Schmidt distance measure. J Quantum Inf Sci 5, 127–133. https://doi.org/10.4236/jqis.2015.54015.
67. Ellerman D (2018), Logical entropy: introduction to classical and quantum logical information theory. Entropy 20, 679. https://doi.org/10.3390/e20090679.
68. Auletta G, Fortunato M, Parisi G (2009), Quantum mechanics, Cambridge University Press, Cambridge, UK.
69. Bennett CH (2003), Quantum information: qubits and quantum error correction. Int J Theor Phys 42, 153–176. https://doi.org/10.1023/A:1024439131297.
70. Jaeger G (2007), Quantum information: an overview, Springer Science+Business Media, New York.
71. Manfredi G, Feix MR (2000), Entropy and Wigner Functions. Phys Rev E 62, 4665–4674. https://doi.org/10.1103/PhysRevE.62.4665.
72. Birkhoff G, Von Neumann J (1936), The logic of quantum mechanics. Ann Math 37, 823–843.
73. Ellerman D (2017), Quantum mechanics over sets: a pedagogical model with non-commutative finite probability theory as its quantum probability calculus. Synthese 194, 4863–4896. https://doi.org/10.1007/s11229-016-1175-0.
74. Ellerman D (2018), The quantum logic of direct-sum decompositions: the dual to the quantum logic of subspaces. Logic J IGPL 26, 1–13. https://doi.org/10.1093/jigpal/jzx026.
75. Hoffman K, Kunze R (1961), Linear algebra, Prentice-Hall, Englewood Cliffs, NJ.
76. Kolmogorov AN (1983), Combinatorial foundations of information theory and the calculus of probabilities. Russian Math Surv 38, 29–40.
77. Zurek WH (2003), Decoherence, einselection, and the quantum origins of the classical. Rev Modern Phys 75, 715–775.
78. Fano U (1957), Description of states in quantum mechanics by density matrix and operator techniques. Rev Mod Phys 29, 74–93.
79. Nielsen M, Chuang I (2000), Quantum computation and quantum information, Cambridge University Press, Cambridge.
80. Tamir B, Cohen E (2014), Logical entropy for quantum states. https://ArXiv.org. http://de.arxiv.org/abs/1412.0616v2
81. Ellerman D (2009), Counting distinctions: on the conceptual foundations of Shannon's Information Theory. Synthese 168, 119–149. https://doi.org/10.1007/s11229-008-9333-7.
82. Ellerman D (2021), New foundations for information theory: logical entropy and Shannon entropy, Springer Nature, Cham, Switzerland.
83. Tamir B, Piava IL, Schwartzman-Nowik Z, Cohen E (2021), Quantum logical entropy: fundamentals and general properties. https://ArXiv.org. https://arxiv.org/abs/2108.02726.

Cite this article as: Ellerman D 2022. Introduction to logical entropy and its relationship to Shannon entropy. 4open, **5**, 1.