Counting Distinctions: On the Conceptual Foundations of Shannon's Information Theory

David Ellerman*
Department of Philosophy
University of California at Riverside

Published as: Counting Distinctions: On the Conceptual Foundations of Shannon's Information Theory. *Synthese*. 168 (1, May 2009): 119-149.

Contents

Abstract

Categorical logic has shown that modern logic is essentially the logic of subsets (or "subobjects"). Partitions are dual to subsets so there is a dual logic of partitions where a "distinction" [an ordered pair of distinct elements (u, u') from the universe U] is dual to an "element". An element being in a subset is analogous to a partition π on U making a distinction, i.e., if u and u'

*This paper is dedicated to the memory of Gian-Carlo Rota—mathematician, philosopher, mentor, and friend.

were in different blocks of π. Subset logic leads to finite probability theory by taking the (Laplacian) probability as the normalized size of each subset-event of a finite universe. The analogous step in the logic of partitions is to assign to a partition the number of distinctions made by a partition normalized by the total number of ordered pairs $|U|^2$ from the finite universe. That yields a notion of "logical entropy" for partitions and a "logical information theory." The logical theory directly counts the (normalized) number of distinctions in a partition while Shannon's theory gives the average number of binary partitions needed to make those same distinctions. Thus the logical theory is seen as providing a conceptual underpinning for Shannon's theory based on the logical notion of "distinctions."

1 Towards a Logic of Partitions

Propositional logic may be modeled as the logic of subsets as was indeed originally proposed by Boole. The connectives are given subset interpretations and a *tautology* is a formula that regardless of what subsets of a universe set U are substituted for the variables, the formula evaluates to U. Then it is a theorem (rather than a definition) that it suffices to take $U = 1$ which has only two subsets symbolized as $0, 1$ or F, T, i.e., a formula is a tautology if and only it is a truth-table tautology. Thus "propositional" logic is seen as the *logic of subsets* of a set. Largely due to the efforts of William Lawvere, the modern treatment of logic was reformulated and vastly generalized using category theory in what is now called *categorical logic*. Subsets were generalized to subobjects or "parts" (equivalence classes of monomorphisms) so that logic has become the logic of subobjects.[1]

There is a duality between *subsets* of a set and *partitions*[2] on a set. "The dual notion (obtained by reversing the arrows) of 'part' is the notion of *partition*." [23, p. 85] In category theory, this emerges as the reverse-the-arrows duality between monomorphisms (monos), e.g., injective set functions, and epimorphisms (epis), e.g., surjective set functions, and between subobjects and quotient objects. If modern logic is formulated as the logic of subsets, or more generally, subobjects or "parts", then the question naturally arises of a dual logic that might play the analogous role for partitions and their generalizations.

Quite aside from category theory duality, it has long been noted in combinatorial mathematics, e.g., in Gian-Carlo Rota's work in combinatorial theory and probability theory [3], that there is a type of duality between subsets of a set and partitions on a set. Just as subsets of a set are partially ordered by inclusion, so partitions on a set are partially ordered by refinement.[3] Moreover, both partial orderings are in fact lattices (i.e., have meets and joins) with a top element $\widehat{1}$ and a bottom element $\widehat{0}$. In the lattice of all subsets $\mathcal{P}(U)$ (the power set) of a set U, the meet and join are, of course, intersection and union while the top element is the universe U and the bottom element is the null set $\emptyset$. In the lattice of all partitions $\Pi(U)$ on a non-empty set U, there are also meet and join operations (defined later) while the bottom element is the indiscrete partition (the "blob") where all of U is one block and the top element is the discrete partition where each element of U is a singleton block.[4]

This paper is part of a research programme to develop the general dual logic of partitions. The principal novelty in this paper is an analogy between the usual semantics for subset logic and a

[1] See [23] Appendix A for a good treatment.

[2] A *partition* π on a set U is usually defined as a mutually exclusive and jointly exhaustive set $\{B\}_{B\in\pi}$ of subsets or "blocks" $B \subseteq U$. Every equivalence relation on a set U determines a partition on U (with the equivalence classes as the blocks) and vice-versa. For our purposes, it is useful to think of partitions as binary relations defined as the complement to an equivalence relation in the set of ordered pairs $U \times U$. Intuitively, they have complementary functions in the sense that equivalence relations identify while partitions distinguish elements of U.

[3] A partition π more refined than a partition σ, written $\sigma \preceq \pi$, if each block of π is contained in some block of σ. Much of the older literature (e.g., [5, Example 6, p. 2]) writes this relationship the other way around but, for reasons that will become clear, we are adopting a newer way of writing refinement (e.g., [14]) so that the more refined partition is higher in the refinement ordering.

[4] Rota and his students have developed a logic for a special type of equivalence relation (which is rather ubiquitous in mathematics) using join and meet as the only connectives.[7]

suggested semantics for partition logic; the themes of the paper unfold from that starting point. Starting with the analogy between a subset of a set and a partition on the set, the analogue to the notion of an *element* of a subset is the notion of a *distinction* of a partition which is simply an ordered pair $(u, u') \in U \times U$ in distinct blocks of the partition.[5] The logic of subsets leads to finite probability theory where events are subsets S of a finite sample space U and which assigns probabilities $\text{Prob}(S)$ to subsets (e.g., the Laplacian equiprobable distribution where $\text{Prob}(S) = |S| / |U|$). Following the suggested analogies, the logic of partitions similarly leads to a "logical" information theory where the numerical value naturally assigned to a partition can be seen as the *logical information content* or *logical entropy* $h(\pi)$ of the partition. It is initially defined in a Laplacian manner as the number of distinctions that a partition makes normalized by the number of ordered pairs of the universe set U. The probability interpretation of $h(\pi)$ is the probability that a random pair from $U \times U$ is distinguished by π, just as $\text{Prob}(S)$ is the probability that a random choice from U is an element of S. This logical entropy is precisely related to Shannon's entropy measure [32] so the development of logical information theory can be seen as providing a new conceptual basis for information theory at the basic level of logic using "distinctions" as the conceptual atoms.

Historically and conceptually, probability theory started with the simple logical operations on subsets (e.g., union, intersection, and complementation) and assigned a numerical measure to subsets of a finite set of outcomes (number of favorable outcomes divided by the total number of outcomes). Then probability theory "took off" from these simple beginnings to become a major branch of pure and applied mathematics.

The research programme for partition logic that underlies this paper sees Shannon's information theory as "taking off" from the simple notions of partition logic in analogy with the conceptual development of probability theory that starts with simple notions of subset logic. But historically, Shannon's information theory appeared "as a bolt out of the blue" in a rather sophisticated and axiomatic form. Moreover, partition logic is still in its infancy today, not to mention the over half a century ago when Shannon's theory was published.[6] But starting with the suggested semantics for partition logic (i.e., the subset-to-partition and element-to-distinction analogies), we develop the partition analogue ("counting distinctions") of the beginnings of finite probability theory ("counting outcomes"), and then we show how it is related to the already-developed information theory of Shannon. It is in that sense that the developments in the paper provide a logical or conceptual foundation ("foundation" in the sense of a basic conceptual starting point) for information theory.[7]

The following table sets out some of the analogies in a concise form (where the diagonal in $U \times U$ is $\Delta_U = \{(u, u) \,|\, u \in U\}$).

[5] Intuitively we might think of an element of a set as an "it." We will argue that a distinction or "dit" is the corresponding logical atom of information. In economics, there is a basic distinction between rivalrous goods (where more for one means less for another) such a material things ("its") in contrast to non-rivalrous goods (where what one person acquires does not take away from another) such as ideas, knowledge, and information ("bits" or "dits"). In that spirit, an element of a set represents a material thing, an "it," while the dual notion of a distinction or "dit" represents the immaterial notion of two "its" being distinct. The distinction between u and u' is the fact that $u \neq u'$, not a new "thing" or "it." But for mathematical purposes we may represent a distinction by a pair of distinct elements such as the ordered pair (u, u') which is a higher level "it," i.e., an element in the Cartesian product of a set with itself (see next section).

[6] For instance, the conceptual beginnings of probability theory in subset logic is shown by the role of Boolean algebras in probability theory, but what is the corresponding algebra for partition logic?

[7] Perhaps an analogy will be helpful. It is *as if* the axioms for probability theory had first emerged full-blown from Kolmogorov [21] and then one realized belatedly that the discipline could be seen as growing out of the starting point of operations on subsets of a finite space of outcomes where the logic was the logic of subsets.

Table of Analogies	**Subsets**	**Partitions**
"Atoms"	Elements	Distinctions
All atoms	Universe U (all $u \in U$) $= \widehat{1}$	Discrete partition $\widehat{1}$ (all dits)
No atoms	Null set $\emptyset$ (no $u \in U$) $= \widehat{0}$	Indiscrete partition $\widehat{0}$ (no dits)
Model of proposition or event	Subset $S \subseteq U$	Partition π on U
Model of individual or outcome	Element u in U	Distinction (u, u') in $U \times U - \Delta_U$
Prop. holds or event occurs	Element u in subset S	Partition π distinguishes (u, u')
Lattice of propositions/events	Lattice of all subsets $\mathcal{P}(U)$	Lattice of all partitions $\Pi(U)$
Counting measure (U finite)	# elements in S	# dits (as ordered pairs) in π
Normalized count (U finite)	$\text{Prob}(S) = \frac{\text{\# elements in } S}{\lvert U \rvert}$	$h(\pi) = \frac{\text{\#distinctions in } \pi}{\lvert U \times U \rvert}$
Prob. Interpretation (U finite)	$\text{Prob}(S)$ = probability that random element u is in S	$h(\pi)$ = probability random pair (u, u') is distinguished by π

These analogies show one set of reasons why the lattice of partitions $\Pi(U)$ should be written with the discrete partition as the top element and the indiscrete partition (blob) as the bottom element of the lattice—in spite of the usual convention of writing the "refinement" ordering the other way around as what Gian-Carlo Rota called the "unrefinement ordering."

With this motivation, we turn to the development of this conceptual basis for information theory.

2 Logical Information Theory

2.1 The Closure Space $U \times U$

Claude Shannon's classic 1948 articles [32] developed a statistical theory of communications that is ordinarily called "information theory." Shannon built upon the work of Ralph Hartley [15] twenty years earlier. After Shannon's information theory was presented axiomatically, there was a spate of new definitions of "entropy" with various axiomatic properties but without concrete (never mind logical) interpretations [20]. Here we take the approach of starting with a notion that arises naturally in the logic of partitions, dual to the usual logic of subsets. The notion of a distinction or "dit" is taken as the logical atom of information and a "logical information theory" is developed based on that interpretation. When the universe set U is finite, then we have a numerical notion of "information" or "entropy" $h(\pi)$ of a partition π in the number of distinctions normalized by the number of ordered pairs. This logical "counting distinctions" notion of information or entropy can then be related to Shannon's measure of information or entropy.

The basic conceptual unit in logical information theory is the distinction or *dit* (from "DIsTinction" but motivated by "bit"). A pair (u, u') of distinct elements of U are distinguished by π, i.e., form a dit of π, if u and u' are in different blocks of π.[8] A pair (u, u') are identified by π and form an *indit* (from INDIsTinction or "identification") of the partition if they are contained in the same block of π. A partition on U can be characterized by either its dits or indits (just as a subset S of U can be characterized by the elements added to the null set to arrive at S or by the elements of U thrown out to arrive at S). When a partition π is thought of as determining an equivalence relation, then the equivalence relation, as a set of ordered pairs contained in $U \times U = U^2$, is the *indit set* $\text{indit}(\pi)$ of indits of the partition. But from the view point of logical information theory, the focus is on the distinctions, so the partition π qua binary relation is given by the complementary *dit set* $\text{dit}(\pi)$ of dits where $\text{dit}(\pi) = (U \times U) - \text{indit}(\pi) = \text{indit}(\pi)^c$. Rather than think of the partition as resulting from identifications made to the elements of U (i.e., distinctions excluded from the discrete partition), we think of it as being formed by making distinctions starting with the blob. This is

[8] One might also develop the theory using unordered pairs $\{u, u'\}$ but the later development of the theory using probabilistic methods is much facilitated by using ordered pairs (u, u'). Thus for $u \neq u'$, (u, u') and (u', u) count as two distinctions. This means that the count of distinctions in a partition must be normalized by $|U \times U|$. Note that $U \times U$ includes the diagonal self-pairs (u, u) which can never be distinctions.

analogous to a subset S being thought of as the set of elements that must be added to the null set to obtain S rather than the complementary approach to S by giving the elements excluded from U to arrive at S. From this viewpoint, the natural ordering $\sigma \preceq \pi$ of partitions would be given by the inclusion ordering of dit-sets $\operatorname{dit}(\sigma) \subseteq \operatorname{dit}(\pi)$ and that is exactly the new way of writing the refinement relation that we are using, i.e.,

$$\sigma \preceq \pi \text{ iff } \operatorname{dit}(\sigma) \subseteq \operatorname{dit}(\pi).$$

There is a natural ("built-in") closure operation on $U \times U$ so that the equivalence relations on U are given (as binary relations) by the closed sets. A subset $C \subseteq U^2$ is *closed* if it contains the diagonal $\{(u,u) \mid u \in U\}$, if $(u,u') \in C$ implies $(u',u) \in C$, and if (u,u') and (u',u'') are in C, then (u,u'') is in C. Thus the closed sets of U^2 are the reflexive, symmetric, and transitive relations, i.e., the equivalence relations on U. The intersection of closed sets is closed and the intersection of all closed sets containing a subset $S \subseteq U^2$ is the *closure* $\overline{S}$ of S.

It should be carefully noted that the closure operation on the closure space U^2 is not a *topological* closure operation in the sense that the union of two closed set is not necessarily closed. In spite of the closure operation not being topological, we may still refer to the complements of closed sets as being *open* sets, i.e., the dit sets of partitions on U. As usual, the *interior* $\operatorname{int}(S)$ of any subset S is defined as the complement of the closure of its complement: $\operatorname{int}(S) = \left(\overline{S^c}\right)^c$.

The open sets of $U \times U$ ordered by inclusion form a lattice isomorphic to the lattice $\Pi(U)$ of partitions on U. The closed sets of $U \times U$ ordered by inclusion form a lattice isomorphic to $\Pi(U)^{op}$, the opposite of the lattice of partitions on U (formed by turning around the partial order). The motivation for writing the refinement relation in the old way was probably that equivalence relations were thought of as binary relations $\operatorname{indit}(\pi) \subseteq U \times U$, so the ordering of equivalence relations was written to reflect the inclusion ordering between indit-sets. But since a partition and an equivalence relation were then taken as essentially the "same thing," i.e., a set $\{B\}_{B \in \pi}$ of mutually exclusive and jointly exhaustive subsets ("blocks" or "equivalence classes") of U, that way of writing the ordering carried over to partitions. But we identify a partition π *as a binary relation* with its dit-set $\operatorname{dit}(\pi) = U \times U - \operatorname{indit}(\pi)$ so our refinement ordering is the inclusion ordering between dit-sets (the opposite of the inclusion ordering of indit-sets).[9]

Given two partitions π and σ on U, the open set corresponding to the *join* $\pi \vee \sigma$ of the partitions is the partition whose dit-set is the union of their dit-sets:[10]

$$\operatorname{dit}(\pi \vee \sigma) = \operatorname{dit}(\pi) \cup \operatorname{dit}(\sigma).$$

The open set corresponding to the *meet* $\pi \wedge \sigma$ of partitions is the interior of the intersection of their dit-sets:[11]

$$\operatorname{dit}(\pi \wedge \sigma) = \operatorname{int}(\operatorname{dit}(\pi) \cap \operatorname{dit}(\sigma)).$$

The open set corresponding to the bottom or blob $\widehat{0}$ is the null set $\emptyset \subseteq U \times U$ (no distinctions) and the open set corresponding to the discrete partition or top $\widehat{1}$ is the complement of the diagonal, $U \times U - \Delta_U$ (all distinctions).

[9] One way to establish the duality between elements of subsets and distinctions in a partition is to start with the refinement relation as the partial order in the lattice of partitions $\Pi(U)$ analogous to the inclusion partial order in the lattice of subsets $\mathcal{P}(U)$. Then the mapping $\pi \longmapsto \operatorname{dit}(\pi)$ represents the lattice of partitions as the lattice of open subsets of the closure space $U \times U$ with inclusion as the partial order. Then the analogue of the elements in the subsets of $\mathcal{P}(U)$ would be the elements in the subsets $\operatorname{dit}(\pi)$ representing the partitions, namely, the distinctions.

[10] Note that this union of dit sets gives the dit set of the "meet" in the old reversed way of writing the refinement ordering.

[11] Note that this is the "join" in the old reversed way of writing the refinement ordering. This operation defined by the interior operator of the non-topological closure operation leads to "anomolous" results such as the non-distributivity of the partition lattice—in contrast to the distributivity of the lattice of open sets of a topological space.

2.2 Some Set Structure Theorems

Before restricting ourselves to finite U to use the counting measure $|\mathrm{dit}\,(\pi)|$, there are a few structure theorems that are independent of cardinality. If the "atom" of information is the dit then the atomic information in a partition π "is" its dit set, $\mathrm{dit}(\pi)$. The information common to two partitions π and σ, their *mutual information set*, would naturally be the intersection of their dit sets (which is not necessarily the dit set of a partition):

$$\mathrm{Mut}(\pi,\sigma) = \mathrm{dit}\,(\pi) \cap \mathrm{dit}\,(\sigma).$$

Shannon deliberately defined his measure of information so that it would be "additive" in the sense that the measure of information in two independent probability distributions would be the sum of the information measures of the two separate distributions and there would be zero mutual information between the independent distributions. But this is not true at the logical level with information defined as distinctions. There is *always* mutual information between two non-blob partitions—even though the interior of $\mathrm{Mut}\,(\pi,\sigma)$ might be empty, i.e., $\mathrm{int}\,(\mathrm{Mut}(\pi,\sigma)) = \mathrm{int}\,(\mathrm{dit}\,(\pi) \cap \mathrm{dit}\,(\sigma)) = \mathrm{dit}\,(\pi \wedge \sigma)$ might be empty so that $\pi \wedge \sigma = \widehat{0}$.

Proposition 1 *Given two partitions π and σ on U with $\pi \neq \widehat{0} \neq \sigma$, $\mathrm{Mut}\,(\pi,\sigma) \neq \emptyset$.*[12]

Since π is not the blob, consider two elements u and u' distinguished by π but identified by σ [otherwise $(u,u') \in \mathrm{Mut}(\pi,\sigma)$]. 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'. 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. Hence (u,u'') or (u',u'') must be distinguished by both partitions and thus must be in their mutual information set $\mathrm{Mut}\,(\pi,\sigma)$.■ (= end of proof marker)

The closed and open subsets of U^2 can be characterized using the usual notions of blocks of a partition. Given a partition π on U as a set of blocks $\pi = \{B\}_{B\in\pi}$, let $B \times B'$ be the Cartesian product of B and B'. Then

$$\begin{aligned}\mathrm{indit}\,(\pi) &= \bigcup_{B\in\pi} B \times B \\ \mathrm{dit}\,(\pi) &= \bigcup_{\substack{B\neq B' \\ B,B'\in\pi}} B \times B' = U \times U - \mathrm{indit}\,(\pi) = \mathrm{indit}\,(\pi)^c.\end{aligned}$$

The mutual information set can also be characterized in this manner.

Proposition 2 *Given partitions π and σ with blocks $\{B\}_{B\in\pi}$ and $\{C\}_{C\in\sigma}$, then*

$$\mathrm{Mut}\,(\pi,\sigma) = \bigcup_{B\in\pi, C\in\sigma} (B - (B\cap C)) \times (C - (B\cap C)) = \bigcup_{B\in\pi, C\in\sigma} (B - C) \times (C - B).$$

The union (which is a disjoint union) will include the pairs (u,u') where for some $B \in \pi$ and $C \in \sigma$, $u \in B - (B \cap C)$ and $u' \in C - (B \cap C)$. Since u' is in C but not in the intersection $B \cap C$, it must be in a different block of π than B so $(u,u') \in \mathrm{dit}\,(\pi)$. Symmetrically, $(u,u') \in \mathrm{dit}\,(\sigma)$ so $(u,u') \in \mathrm{Mut}\,(\pi,\sigma) = \mathrm{dit}\,(\pi) \cap \mathrm{dit}\,(\sigma)$. Conversely if $(u,u') \in \mathrm{Mut}\,(\pi,\sigma)$ then take the B containing u and the C containing u'. Since (u,u') is distinguished by both partitions, $u \notin C$ and $u' \notin B$ so that $(u,u') \in (B - (B \cap C)) \times (C - (B \cap C))$.■

[12] The contrapositive of this proposition is interesting. Given two equivalence relations $E_1, E_2 \subseteq U^2$, if every pair of elements $u, u' \in U$ is identified by one or the other of the relations, i.e., $E_1 \cup E_2 = U^2$, then either $E_1 = U^2$ or $E_2 = U^2$.

2.3 Logical Information Theory on Finite Sets

For a finite set U, the (normalized) "counting distinctions" measure of information can be defined and compared to Shannon's measure for finite probability distributions. Since the information set of a partition π on U is its set of distinctions $\operatorname{dit}(\pi)$, the un-normalized numerical measure of the information of a partition is simply the count of that set, $|\operatorname{dit}(\pi)|$ ("dit count"). But to account for the total number of ordered pairs of elements from U, we normalize by $|U \times U| = |U|^2$ to obtain the *logical information content* or *logical entropy* of a partition π as its normalized dit count:

$$\boxed{h(\pi) = \frac{|\operatorname{dit}(\pi)|}{|U \times U|}}.$$

Probability theory started with the finite case where there was a finite set U of possibilities (the finite sample space) and an event was a subset $S \subseteq U$. Under the Laplacian assumption that each outcome was equiprobable, the probability of the event S was the similar normalized counting measure of the set:

$$\operatorname{Prob}(S) = \frac{|S|}{|U|}.$$

This is the probability that any randomly chosen element of U is an element of the subset S. In view of the dual relationship between being in a subset and being distinguished by a partition, the analogous concept would be the probability that an ordered pair (u, u') of elements of U chosen independently (i.e., with replacement[13]) would be distinguished by a partition π, and that is precisely the logical entropy $h(\pi) = |\operatorname{dit}(\pi)| / |U \times U|$ (since each pair randomly chosen from $U \times U$ is equiprobable).

$$\boxed{\text{Probabilistic interpretation: } h(\pi) = \text{probability a random pair is distinguished by } \pi}.$$

In finite probability theory, when a point is sampled from the sample space U, we say the event S *occurs* if the point u was an element in $S \subseteq U$. When a random pair (u, u') is sampled from the sample space $U \times U$, we say the partition π *distinguishes*[14] if the pair is distinguished by the partition, i.e., if $(u, u') \in \operatorname{dit}(\pi) \subseteq U \times U$. Then just as we take $\operatorname{Prob}(S)$ as the probability that the event S occurs, so the logical entropy $h(\pi)$ is the probability that the partition π distinguishes.

Since $\operatorname{dit}(\pi \vee \sigma) = \operatorname{dit}(\pi) \cup \operatorname{dit}(\sigma)$,

$$\boxed{\text{probability that } \pi \vee \sigma \text{ distinguishes} = h(\pi \vee \sigma) = \text{probability that } \pi \text{ or } \sigma \text{ distinguishes}}.$$

The probability that a randomly chosen pair would be distinguished by π *and* σ would be given by the relative cardinality of the mutual information set which is called the *mutual information* of the partitions:

$$\boxed{\text{Mutual logical information: } m(\pi, \sigma) = \frac{|\operatorname{Mut}(\pi, \sigma)|}{|U|^2} = \text{probability that } \pi \text{ and } \sigma \text{ distinguishes}}.$$

Since the cardinality of intersections of sets can be analyzed using the inclusion-exclusion principle, we have:

$$|\operatorname{Mut}(\pi, \sigma)| = |\operatorname{dit}(\pi) \cap \operatorname{dit}(\sigma)| = |\operatorname{dit}(\pi)| + |\operatorname{dit}(\sigma)| - |\operatorname{dit}(\pi) \cup \operatorname{dit}(\sigma)|.$$

Normalizing, the probability that a random pair is distinguished by both partitions is given by the modular law:

[13] Drawing with replacement would allow diagonal pairs (u, u) to be drawn and requires $|U \times U|$ as the normalizing factor.

[14] Equivalent terminology would be "differentiates" or "discriminates."

$$m(\pi,\sigma)=\frac{|\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)|}{|U|^2}=\frac{|\mathrm{dit}(\pi)|}{|U|^2}+\frac{|\mathrm{dit}(\sigma)|}{|U|^2}-\frac{|\mathrm{dit}(\pi)\cup\mathrm{dit}(\sigma)|}{|U|^2}=h(\pi)+h(\sigma)-h(\pi\vee\sigma).$$

This can be extended by the inclusion-exclusion principle to any number of partitions. The mutual information set $\mathrm{Mut}(\pi,\sigma)$ is not the dit-set of a partition but its interior is the dit-set of the meet so the logical entropies of the join and meet satisfy the:

$$\text{Submodular inequality: } h(\pi\wedge\sigma)+h(\pi\vee\sigma)\leq h(\pi)+h(\sigma).$$

2.4 Using General Finite Probability Distributions

Since the logical entropy of a partition on a finite set can be given a simple probabilistic interpretation, it is not surprising that many methods of probability theory can be harnessed to develop the theory. The theory for the finite case can be developed at two different levels of generality, using the specific Laplacian equiprobability distribution on the finite set U or using an arbitrary finite probability distribution. Correctly formulated, all the formulas concerning logical entropy and the related concepts will work for the general case, but our purpose is not mathematical generality. Our purpose is to give the basic motivating example of logical entropy based on "counting distinctions" and to show its relationship to Shannon's notion of entropy, thereby clarifying the logical foundations of the latter concept.

Every probability distribution on a finite set U gives a probability p_B for each block B in a partition π but for the Laplacian distribution, it is just the relative cardinality of the block: $p_B=\frac{|B|}{|U|}$ for blocks $B\in\pi$. Since there are no empty blocks, $p_B>0$ and $\sum_{B\in\pi}p_B=1$. Since the dit set of a partition is $\mathrm{dit}(\pi)=\bigcup_{B\neq B'}B\times B'$, its size is $|\mathrm{dit}(\pi)|=\sum_{B\neq B'}|B|\,|B'|=\sum_{B\in\pi}|B|\,|U-B|$. Thus the logical information or entropy in a partition as the normalized size of the dit set can be developed as follows:

$$\boxed{h(\pi)=\frac{\sum_{B\neq B'}|B|\left|B'\right|}{|U|\times|U|}=\sum_{B\neq B'}p_Bp_{B'}=\sum_{B\in\pi}p_B(1-p_B)=1-\sum_{B\in\pi}p_B^2.}$$

Having defined and interpreted logical entropy in terms of the distinctions of a set partition, we may, if desired, "kick away the ladder" and define the logical entropy of any finite probability distribution $p=\{p_1,...,p_n\}$ as:

$$\boxed{h(p)=\textstyle\sum_{i=1}^n p_i(1-p_i)=1-\sum_{i=1}^n p_i^2}.$$

The probabilistic interpretation is that $h(p)$ is the probability that two independent draws (from the sample space of n points with these probabilities) will give distinct points.[15]

2.5 A Brief History of the Logical Entropy Formula: $h(p)=\sum_i p_i(1-p_i)$

The logical entropy formula $h(p)=\sum_i p_i(1-p_i)=1-\sum_i p_i^2$ was motivated as the normalized count of the distinctions made by a partition, $|\mathrm{dit}(\pi)|/|U|^2$, when the probabilities are the block probabilities $p_B=\frac{|B|}{|U|}$ of a partition on a set U (under a Laplacian assumption). The complementary measure $1-h(p)=\sum_i p_i^2$ would be motivated as the normalized count of the identifications made by a partition, $|\mathrm{indit}(\pi)|/|U|^2$, thought of as an equivalence relation. Thus $1-\sum_i p_i^2$, motivated by distinctions, is a measure of heterogeneity or diversity, while the complementary measure $\sum_i p_i^2$, motivated by identifications, is a measure of homogeneity or concentration. Historically, the formula

[15] Note that we can always rephrase in terms of partitions by taking $h(p)$ as the entropy $h\left(\widehat{1}\right)$ of discrete partition on $U=\{u_1,...,u_n\}$ with the p_i's as the probabilities of the singleton blocks $\{u_i\}$ of the discrete partition.

can be found in either form depending on the particular context. The p_i's might be relative shares such as the relative share of organisms of the i^{th} species in some population of organisms, and then the interpretation of p_i as a probability arises by considering the random choice of an organism from the population.

According to I. J. Good, the formula has a certain naturalness: "If $p_1, ..., p_t$ are the probabilities of t mutually exclusive and exhaustive events, any statistician of this century who wanted a measure of homogeneity would have take about two seconds to suggest $\sum p_i^2$ which I shall call ρ." [13, p. 561] As noted by Bhargava and Uppuluri [4], the formula $1-\sum p_i^2$ was used by Gini in 1912 ([10] reprinted in [11, p. 369]) as a measure of "mutability" or diversity. But another development of the formula (in the complementary form) in the early twentieth century was in cryptography. The American cryptologist, William F. Friedman, devoted a 1922 book ([8]) to the "index of coincidence" (i.e., $\sum p_i^2$). Solomon Kullback (see the Kullback-Leibler divergence treated later) worked as an assistant to Friedman and wrote a book on cryptology which used the index. [22]

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 (i.e., the identification probability $1 - h(p)$). Polish cryptoanalysts had independently used the repeat rate in their work on the Enigma [27].

After the war, 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."[33, 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. In the biodiversity literature [31], the formula is known as "Simpson's index of diversity" or sometimes, the "Gini-Simpson diversity index." However, Simpson along with I. J. Good worked at Bletchley during WWII, and, according to Good, "E. H. Simpson and I both obtained the notion [the repeat rate] from Turing." [12, 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." [13, p. 562]

In 1945, Albert O. Hirschman ([18, p. 159] and [19]) suggested 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 later, Orris Herfindahl [17] independently suggested using $\sum p_i^2$ as an index of industrial concentration (where p_i is the relative share of the i^{th} firm in an industry). In the industrial economics literature, the index $H = \sum p_i^2$ is variously called the Hirschman-Herfindahl index, the HH index, or just the H index of concentration. If all the relative shares were equal (i.e., $p_i = 1/n$), then the identification or repeat probability is just the probability of drawing any element, i.e., $H = 1/n$, so $\frac{1}{H} = n$ is the number of equal elements. This led to the "numbers equivalent" interpretation of the reciprocal of the H index [2]. In general, given an event with probability p_0, the "numbers-equivalent" interpretation of the event is that it is 'as if' an element was drawn out of a set of $\frac{1}{p_0}$ equiprobable elements (it is 'as if' since $1/p_0$ need not be an integer). This numbers-equivalent idea is related to the "block-count" notion of entropy defined later.

In view of the frequent and independent discovery and rediscovery of the formula $\rho = \sum p_i^2$ or its complement $1 - \sum p_i^2$ by Gini, Friedman, Turing, Hirschman, Herfindahl, and no doubt others, I. J. Good wisely advises that "it is unjust to associate ρ with any one person." [13, p. 562][16]

After Shannon's axiomatic introduction of his entropy [32], there was a proliferation of axiomatic entropies with a variable parameter.[17] The formula $1 - \sum p_i^2$ for logical entropy appeared as a

[16] The name "logical entropy" for $1-\sum p_i^2$ not only denotes the basic status of the formula, it avoids "Stigler's Law of Eponymy": "No scientific discovery is named after its original discoverer."[34, p. 277]

[17] There was no need for Shannon to present his entropy concept axiomatically since it was based on a standard

special case for a specific parameter value in several cases. During the 1960's, Aczél and Daróczy [1] developed the *generalized entropies of degree* α:

$$H_n^\alpha(p_1,...,p_n) = \frac{\sum_i p_i^\alpha - 1}{(2^{1-\alpha}-1)}$$

and the logical entropy occurred as half the value for $\alpha = 2$. That formula also appeared as Havrda-Charvat's *structural* α*-entropy* [16]:

$$S(p_1,...,p_n,;\alpha) = \frac{2^{\alpha-1}}{2^{\alpha-1}-1}\left(1-\sum_i p_i^\alpha\right)$$

and the special case of $\alpha = 2$ was considered by Vajda [36].

Patil and Taillie [25] defined the *diversity index of degree* β in 1982:

$$\Delta_\beta = \frac{1-\sum_i p_i^{\beta+1}}{\beta}$$

and Tsallis [35] independently gave the same formula as an entropy formula in 1988:

$$S_q(p_1,...,p_n) = \frac{1-\sum_i p_i^q}{q-1}$$

where the logical entropy formula occurs as a special case ($\beta = 1$ or $q = 2$). While the generalized parametric entropies may be interesting as axiomatic exercises, our purpose is to emphasize the specific logical interpretation of the logical entropy formula (or its complement).

From the logical viewpoint, two elements from $U = \{u_1,...,u_n\}$ are either identical or distinct. Gini [10] introduced d_{ij} as the "distance" between the i^{th} and j^{th} elements where $d_{ij} = 1$ for $i \neq j$ and $d_{ii} = 0$. Since $1 = (p_1 + ... + p_n)(p_1 + ... + p_n) = \sum_i p_i^2 + \sum_{i\neq j} p_i p_j$, the logical entropy, i.e., Gini's index of mutability, $h(p) = 1 - \sum_i p_i^2 = \sum_{i\neq j} p_i p_j$, is the average logical distance between a pair of independently drawn elements. But one might generalize by allowing other distances $d_{ij} = d_{ji}$ for $i \neq j$ (but always $d_{ii} = 0$) so that $Q = \sum_{i\neq j} d_{ij} p_i p_j$ 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] (which was later rediscovered in the biodiversity literature as the "Avalanche Index" by Ganeshaish et al. [9]). In many domains, it is quite reasonable to move beyond the bare-bones logical distance of $d_{ij} = 1$ for $i \neq j$ so that Rao's quadratic entropy is a useful and easily interpreted generalization of logical entropy.

3 Relationship between the Logical and Shannon Entropies

3.1 The Search Approach to Find the "Sent Message"

The logical entropy $h(\pi) = \sum_{B\in\pi} p_B(1-p_B)$ in this form as an average over blocks allows a direct comparison with Shannon's entropy $H(\pi) = \sum_{B\in\pi} p_B \log_2(\frac{1}{p_B})$ of the partition which is also an average over the blocks. What is the connection between the *block entropies* $h(B) = 1 - p_B$ and $H(B) = \log_2\left(\frac{1}{p_B}\right)$? Shannon uses reasoning (shared with Hartley) to arrive at a notion of entropy or information content for an element out of a subset (e.g., a block in a partition as a set of blocks). Then for a partition π, Shannon averages the block values to get the partition value $H(\pi)$. Hartley and Shannon start with the question of the information required to single an element u out of a set U, e.g., to single out the sent message from the set of possible messages. Alfred Renyi has

concrete interpretation (expected number of binary partitions needed to distinguish a designated element) which could then be generalized. The axiomatic development encouraged the presentation of other "entropies" as if the axioms eliminated or, at least, relaxed any need for an interpretation of the "entropy" concept.

also emphasized this "search-theoretic" approach to information theory (see [28], [29], or numerous papers in [30]).[18]

One intuitive measure of the information obtained by determining the designated element in a set U of equiprobable elements would just be the cardinality $|U|$ of the set, and, as we will see, that leads to a multiplicative "block-count" version of Shannon's entropy. But Hartley and Shannon wanted the additivity that comes from taking the logarithm of the set size $|U|$. If $|U| = 2^n$ then this allows the crucial Shannon interpretation of $\log_2(|U|) = n$ as the minimum number of yes-or-no questions (binary partitions) it takes to single out any designated element (the "sent message") of the set. In a mathematical version of the game of twenty questions (like Rényi's Hungarian game of "Bar-Kochba"), think of each element of U as being assigned a unique binary number with n digits. Then the minimum n questions can just be the questions asking for the i^{th} binary digit of the hidden designated element. Each answer gives one *bit* (short for "binary digit") of information. With this motivation for the case of $|U| = 2^n$, Shannon and Hartley take $\log(|U|)$ as the measure of the information required to single out a hidden element in a set with $|U|$ equiprobable elements.[19] That extends the "minimum number of yes-or-no questions" motivation from $|U| = 2^n$ to any finite set U with $|U|$ equiprobable elements. If a partition π had equiprobable blocks, then the Shannon entropy would be $H(B) = \log(|\pi|)$ where $|\pi|$ is the number of blocks.

To extend this basic idea to sets of elements which are not equiprobable (e.g., partitions with unequal blocks), it is useful to use an old device to restate any positive probability as a chance among equiprobable elements. If $p_i = 0.02$, then there is a 1 in $50 = \frac{1}{p_i}$ chance of the i^{th} outcome occurring in any trial. It is "as if" the outcome was one among $1/p_i$ equiprobable outcomes.[20] Thus each positive probability p_i has an associated *equivalent number* $1/p_i$ which is the size of the hypothetical set of equiprobable elements so that the probability of drawing any given element is p_i.[21]

Given a partition $\{B\}_{B\in\pi}$ with unequal blocks, we motivate the block entropy $H(B)$ for a block with probability p_B by taking it as the entropy for a hypothetical *numbers-equivalent partition* π_B with $\frac{1}{p_B}$ equiprobable blocks, i.e.,

$$H(B) = \log(|\pi_B|) = \log\left(\tfrac{1}{p_B}\right).$$

With this motivation, the Shannon entropy of the partition is then defined as the arithmetical average of the block entropies:

$$\text{Shannon's entropy: } H(\pi) = \sum_{B\in\pi} p_B H(B) = \sum_{B\in\pi} p_B \log\left(\tfrac{1}{p_B}\right).$$

This can be directly compared to the logical entropy $h(\pi) = \sum_{B\in\pi} p_B(1-p_B)$ which arose from quite different distinction-based reasoning (e.g., where the search of a single designated element played no role). Nevertheless, the formula $\sum_{B\in\pi} p_B(1-p_B)$ can be viewed as an average over the quantities which play the role of "block entropies" $h(B) = (1-p_B)$. But this "block entropy" cannot be directly interpreted as a (normalized) dit count since there is no such thing as the dit count for a single block. The dits are the pairs of elements in *distinct* blocks.

[18] In Gian-Carlo Rota's teaching, he supposed that the Devil had picked an element out of U and would not reveal its identity. But when given a binary partition (i.e., a yes-or-no question), the Devil had to truthfully tell which block contained the hidden element. Hence the problem was to find the minimum number of binary partitions needed to force the Devil to reveal the hidden element.

[19] Hartley used logs to the base 10 but here all logs are to base 2 unless otherwise indicated. Instead of considering whether the base should be 2, 10, or e, it is perhaps more important to see that there is a natural base-free variation $H_m(\pi)$ on Shannon's entropy (see "block-count entropy" defined below).

[20] Since $1/p_i$ need not be an integer (or even rational), one could interpret the equiprobable "number of elements" as being heuristic or one could restate it in continuous terms. The continuous version is the uniform distribution on the real interval $[0, 1/p_i]$ where the probability of an outcome in the unit interval $[0,1]$ is $1/(1/p_i) = p_i$.

[21] In continuous terms, the *numbers-equivalent* is the length of the interval $[0, 1/p_i]$ with the uniform distribution on it.

For comparison purposes, we may nevertheless carry over the heuristic reasoning to the case of logical entropy. For each block B, we take the same hypothetical numbers-equivalent partition π_B with $\frac{|U|}{|B|} = \frac{1}{p_B}$ equal blocks of size $|B|$ and then take the desired block entropy $h(B)$ as the normalized dit count $h(\pi_B)$ for that partition. Each block contributes $p_B(1-p_B)$ to the normalized dit count and there are $|U|/|B| = 1/p_B$ blocks in π_B so the total normalized dit count simplifies to: $h(\pi_B) = \frac{1}{p_B} p_B(1-p_B) = 1 - p_B = h(B)$, which we could take as the *logical block entropy*. Then the average of these logical block entropies gives the logical entropy $h(\pi) = \sum_{B\in\pi} p_B h(B) = \sum_{B\in\pi} p_B(1-p_B)$ of the partition π, all in the manner of the heuristic development of Shannon's $H(\pi) = \sum_{B\in\pi} p_B \log\left(\frac{1}{p_B}\right)$.

There is, however, no need to go through this reasoning to arrive at the logical entropy of a partition as the average of block entropies. The interpretation of the logical entropy as the normalized dit count survives the averaging even though all the blocks of π might have different sizes, i.e., the interpretation "commutes" with the averaging of block entropies. Thus $h(\pi)$ is the *actual* dit count (normalized) for a partition π, not just the average of block entropies $h(B)$ that could be interpreted as the normalized dit counts for hypothetical partitions π_B.

The interpretation of the Shannon measure of information as the minimum number of binary questions it takes to single out a designated block does not commute with the averaging over the set of different-sized blocks in a partition. Hence the Shannon entropy of a partition is the *expected* number of bits it takes to single out the designated block while the logical entropy of a partition on a set is the *actual* number of dits (normalized) distinguished by the partition.

The last step in connecting Shannon entropy and logical entropy is to rephrase the heuristics behind Shannon entropy in terms of "making all the distinctions" rather than "singling out the designated element."

3.2 Distinction-based Treatment of Shannon's Entropy

The search-theoretic approach was the heritage of the original application of information theory to communications where the focus was on singling out a designated element, the sent message. In the "twenty questions" version, one person picks a hidden element and the other person seeks the minimum number of binary partitions on the set of possible answers to single out the answer. But it is simple to see that the focus on the single designated element was unnecessary. The essential point was *to make all the distinctions to separate the elements*–since any element could have been the designated one. If the join of the minimum number of binary partitions did not distinguish all the elements into singleton blocks, then one could not have picked out the hidden element if it was in a non-singleton block. Hence the distinction-based treatment of Shannon's entropy amounts to rephrasing the above heuristic argument in terms of "making all the distinctions" rather than "making the distinctions necessary to single out any designated element."

In the basic example of $|U| = 2^n$ where we may think of the 2^n like or equiprobable elements as being encoded with n binary digit numbers, then $n = \log\left(\frac{1}{1/2^n}\right)$ is the minimum number of binary partitions (each partitioning according to one of the n digits) necessary *to make all the distinctions* between the elements, i.e., the minimum number of binary partitions whose join is the discrete partition with singleton blocks (each block probability being $p_B = 1/2^n$). Generalizing to any set U of equiprobable elements, the minimum number of bits necessary to distinguish all the elements from each other is $\log\left(\frac{1}{1/|U|}\right) = \log(|U|)$. Given a partition $\pi = \{B\}_{B\in\pi}$ on U, the block entropy $H(B) = \log\left(\frac{1}{p_B}\right)$ is the minimum number of bits necessary to distinguish all the blocks in the numbers-equivalent partition π_B, and the average of those block entropies gives the Shannon entropy: $H(\pi) = \sum_{B\in\pi} p_B \log\left(\frac{1}{p_B}\right)$.

The point of rephrasing the heuristics behind Shannon's definition of entropy in terms of the

average bits needed to "make all the distinctions" is that it can then be directly compared with the logical definition of entropy which is simply the total number of distinctions normalized by $|U|^2$. Thus the two definitions of entropy boil down to two different ways of measuring the totality of distinctions. A third way to measure the totality of distinctions, called the "block-count entropy," is defined below. Hence we have our overall theme that these three notions of entropy boil down to three ways of "counting distinctions."

3.3 Relationships Between the Block Entropies

Since the logical and Shannon entropies have formulas presenting them as averages of block-entropies, $h(\pi) = \sum_{B\in\pi} p_B (1 - p_B)$ and $H(\pi) = \sum_{B\in\pi} p_B \log\left(\frac{1}{p_B}\right)$, the two notions are precisely related by their respective block entropies, $h(B) = 1 - p_B$ and $H(B) = \log\left(\frac{1}{p_B}\right)$. Solving each for p_B and then eliminating it yields the:

$$\boxed{\text{Block entropy relationship: } h(B) = 1 - \tfrac{1}{2^{H(B)}} \text{ and } H(B) = \log\left(\tfrac{1}{1-h(B)}\right).}$$

The block entropy relation, $h(B) = 1 - \frac{1}{2^{H(B)}}$, has a simple probabilistic interpretation. Thinking of $H(B)$ as an integer, $H(B)$ is the Shannon entropy of the discrete partition on U with $|U| = 2^{H(B)}$ elements while $h(B) = 1 - \frac{1}{2^{H(B)}} = 1 - p_B$ is the logical entropy of that partition since $1/2^{H(B)}$ is the probability of each block in that discrete partition. The probability that a random pair is distinguished by a discrete partition is just the probability that the second draw is distinct from the first draw. Given the first draw from a set of $2^{H(B)}$ individuals, the probability that the second draw (with replacement) is different is $1 - \frac{1}{2^{H(B)}} = h(B)$.

To summarize the comparison up to this point, the logical theory and Shannon's theory start by posing different questions which then turn out to be precisely related. Shannon's statistical theory of communications is concerned with determining the sent message out of a set of possible messages. In the basic case, the messages are equiprobable so it is abstractly the problem of determining the hidden designated element out of a set of equiprobable elements which, for simplicity, we can assume has 2^n elements. The process of determining the hidden element can be conceptualized as the process of asking binary questions which split the set of possibilities into equiprobable parts. The answer to the first question determines which subset of 2^{n-1} elements contains the hidden element and that provides 1 bit of information. An independent equal-blocked binary partition would split each of the 2^{n-1} element blocks into equal blocks with 2^{n-2} elements each. Thus 2 bits of information would determine which of those 2^2 blocks contained the hidden element, and so forth. Thus n independent equal-blocked binary partitions would determine which of the resulting 2^n blocks contains the hidden element. Since there are 2^n elements, each of those blocks is a singleton so the hidden element has been determined. Hence the problem of finding a designated element among 2^n equiprobable elements requires $\log(2^n) = n$ bits of information.

The logical theory starts with the basic notion of a distinction between elements and defines the logical information in a set of distinct 2^n elements as the (normalized) number of distinctions that need to be made to distinguish the 2^n elements. The distinctions are counted as ordered rather than unordered pairs (in order to better apply the machinery of probability theory) and the number of distinctions or dits is normalized by the number of all ordered pairs. Hence a set of 2^n distinct elements would involve $|U \times U - \Delta_U| = 2^n \times 2^n - 2^n = 2^{2n} - 2^n = 2^n(2^n - 1)$ distinctions which normalizes to $\frac{2^{2n}-2^n}{2^{2n}} = 1 - \frac{1}{2^n}$.

There is, however, no need to motivate Shannon's entropy by focusing on the search for a designated element. The task can equivalently be taken as distinguishing all elements from each other rather than distinguishing a designated element from all the other elements. The connection between the two approaches can be seen by computing the total number of distinctions made by intersecting the n independent equal-blocked binary partitions in Shannon's approach.

Example of counting distinctions: Doing the computation, the first partition which creates two sets of 2^{n-1} elements each thereby creates $2^{n-1} \times 2^{n-1} = 2^{2n-2}$ distinctions as unordered pairs and $2 \times 2^{2n-2} = 2^{2n-1}$ distinctions as ordered pairs. The next binary partition splits each of those blocks into equal blocks of 2^{n-2} elements. Each split block creates $2^{n-2} \times 2^{n-2} = 2^{2n-4}$ new distinctions as unordered pairs and there were two such splits so there are $2 \times 2^{2n-4} = 2^{2n-3}$ additional unordered pairs of distinct elements created or 2^{2n-2} new ordered pair distinctions. In a similar manner, the third partition creates 2^{2n-3} new dits and so forth down to the n^{th} partition which adds 2^{2n-n} new dits. Thus in total, the intersection of the n independent equal-blocked binary partitions has created $2^{2n-1}+2^{2n-2}+...+2^{2n-n} = 2^n\left(2^{n-1}+2^{n-2}+...+2^0\right) = 2^n\left(\frac{2^n-1}{2-1}\right) = 2^n\left(2^n-1\right)$ (ordered pair) distinctions which are all the dits on a set with 2^n elements. This is the instance of the block entropy relationship $h(B) = 1 - \frac{1}{2^{H(B)}}$ when the block B is a singleton in a 2^n element set so that $H(B) = \log\left(\frac{1}{1/2^n}\right) = \log\left(2^n\right) = n$ and $h(B) = 1 - \frac{1}{2^{H(B)}} = 1 - \frac{1}{2^n}$.

Thus the Shannon entropy as the number of independent equal-blocked binary partitions it takes to single out a hidden designated element in a 2^n element set is *also* the number of independent equal-blocked binary partitions it takes to distinguish all the elements of a 2^n element set from each other.

The connection between Shannon entropy and logical entropy boils down to two points.

1. The first point is the basic fact that for binary partitions to single out a hidden element ("sent message") in a set is the same as the partitions distinguishing any pair of distinct elements (since if a pair was left undistinguished, the hidden element could not be singled out if it were one of the elements in that undistinguished pair). This gives what might be called the *distinction interpretation* of Shannon entropy as a count of the binary partitions necessary to distinguish between all the distinct messages in the set of possible messages in contrast to the usual *search interpretation* as the binary partition count necessary to find the hidden designated element such as the sent message.

2. The second point is that in addition to the Shannon count of the binary partitions necessary to make all the distinctions, we may use the logical measure that is simply the (normalized) count of the distinctions themselves.

3.4 A Coin-Weighing Example

The logic of the connection between joining independent equal-blocked partitions and efficiently creating dits is not dependent on the choice of base 2. Consider the coin-weighing problem where one has a balance scale and a set of 3^n coins all of which look alike but one is counterfeit (the hidden designated element) and is lighter than the others. The coins might be numbered using the n-digit numbers in mod 3 arithmetic where the three digits are 0, 1, and 2. The n independent ternary partitions are arrived at by dividing the coins into three piles according to the i^{th} digit as $i = 1, ..., n$. To use the n partitions to find the false coin, two of the piles are put on the balance scale. If one side is lighter, then the counterfeit coin is in that block. If the two sides balance, then the light coin is in the third block of coins not on the scale. Thus n weighings (i.e., the join of n independent equal-blocked ternary partitions) will determine the n ternary digits of the false coin, and thus the ternary Shannon entropy is $\log_3\left(\frac{1}{1/3^n}\right) = \log_3\left(3^n\right) = n$ trits. As before we can interpret the joining of independent partitions not only as the most efficient way to find the hidden element (e.g., the false coin or the sent message) but as the most efficient way to make all the distinctions between the elements of the set.

The first partition (separating by the first ternary digit) creates 3 equal blocks of 3^{n-1} elements each so that creates $3 \times 3^{n-1} \times 3^{n-1} = 3^{2n-1}$ unordered pairs of distinct elements or $2 \times 3^{2n-1}$

ordered pair distinctions. The partition according to the second ternary digit divides each of these three blocks into three equal blocks of 3^{n-2} elements each so the additional unordered pairs created are $3 \times 3 \times 3^{n-2} \times 3^{n-2} = 3^{2n-2}$ or $2 \times 3^{2n-2}$ ordered pair distinctions. Continuing in this fashion, the n^{th} ternary partition adds $2 \times 3^{2n-n}$ dits. Hence the total number of dits created by joining the n independent partitions is:

$$2 \times \left[3^{2n-1} + 3^{2n-2} ... + 3^n\right] = 2 \times \left[3^n \left(3^{n-1} + 3^{n-2} ... + 1\right)\right] = 2 \times \left[3^n \frac{(3^n-1)}{3-1}\right] = 3^n \left(3^n - 1\right)$$

which is the total number of ordered pair distinctions between the elements of the 3^n element set. Thus the Shannon measure in trits is the minimum number of ternary partitions needed to create all the distinctions between the elements of a set. The base-3 Shannon entropy is $H_3(\pi) = \sum_{B \in \pi} p_B \log_3 \left(\frac{1}{p_B}\right)$ which for this example of the discrete partition on a 3^n element set U is $H_3\left(\widehat{1}\right) = \sum_{u \in U} \frac{1}{3^n} \log_3 \left(\frac{1}{1/3^n}\right) = \log_3 (3^n) = n$ which can also be thought of as the block value entropy for a singleton block so that we may apply the block value relationship. The logical entropy of the discrete partition on this set is: $h\left(\widehat{1}\right) = \frac{3^n(3^n-1)}{3^{2n}} = 1 - \frac{1}{3^n}$ which could also be thought of as the block value of the logical entropy for a singleton block. Thus the entropies for the discrete partition stand in the block value relationship which for base 3 is:

$$h(B) = 1 - \frac{1}{3^{H_3(B)}}.$$

The example helps to show how the logical notion of a distinction underlies the Shannon measure of information, and how a complete procedure for finding the hidden element (e.g., the sent message) is equivalent to being able to make all the distinctions in a set of elements. But this should not be interpreted as showing that the Shannon's information theory "reduces" to the logical theory. The Shannon theory is addressing an additional question of finding the unknown element. One can have all the distinctions between elements, e.g., the assignment of distinct base-3 numbers to the 3^n coins, without knowing which element is the designated one. Information theory becomes a theory of the *transmission* of information, i.e., a theory of communication, when that second question of "receiving the message" as to which element is the designated one is the focus of analysis. In the coin example, we might say that the information about the light coin was always there in the nature of the situation (i.e., taking "nature" as the sender) but was unknown to an observer (i.e., on the receiver side). The coin weighing scheme was a way for the observer to elicit the information out of the situation. Similarly, the game of twenty questions is about finding a way to uncover the hidden answer—which was all along distinct from the other possible answers (on the sender side). It is this question of the transmission of information (and the noise that might interfere with the process) that carries Shannon's statistical theory of communications well beyond the bare-bones logical analysis of information in terms of distinctions.

3.5 Block-count Entropy

The fact that the Shannon motivation works for other bases than 2 suggests that there might be a base-free version of the Shannon measure (the logical measure is already base-free). Sometimes the reciprocal $\frac{1}{p_B}$ of the probability of an event B is interpreted as the "surprise-value information" conveyed by the occurrence of B. But there is a better concept to use than the vague notion of "surprise-value information." For any positive probability p_0, we defined the reciprocal $\frac{1}{p_0}$ as the *equivalent number* of (equiprobable) elements (always "as it were" since it need not be an integer) since that is the number of equiprobable elements in a set so that the probability of choosing any particular element is p_0. The "big surprise" as a small probability event occurs means it is "as if" a particular element was picked from a big set of elements. For instance, for a block probability $p_B = \frac{|B|}{|U|}$, its numbers-equivalent is the number of blocks $\frac{|U|}{|B|} = \frac{1}{p_B}$ in the hypothetical equal-blocked

partition π_B with each block equiprobable with B. Our task is to develop this number-of-blocks or block-count measure of information for partitions.

The *block-count block entropy* $H_m(B)$ is just the number of blocks in the hypothetical number-of-equivalent-blocks partition π_B where B is one of $\frac{|U|}{|B|} = \frac{1}{p_B}$ associated similar blocks so that $H_m(B) = \frac{1}{p_B}$.

If events B and C were independent, then $p_{B\cap C} = p_B p_C$ so the equivalent number of elements associated with the occurrence of both events is the product $\frac{1}{p_{B\cap C}} = \frac{1}{p_B}\frac{1}{p_C}$ of the number of elements associated with the separate events. This suggests that the average of the block entropies $H_m(B) = \frac{1}{p_B}$ should be the multiplicative average (or geometric mean) rather than the arithmetical average.

Hence we define the *number-of-equivalent blocks entropy* or, in short, *block-count entropy* of a partition π (which does not involve any choice of a base for logs) as the geometric mean of block entropies:

$$\boxed{\text{Block-count entropy: } H_m(\pi) = \prod_{B\in\pi} H_m(B)^{p_B} = \prod_{B\in\pi} \left(\frac{1}{p_B}\right)^{p_B} \text{ blocks}}.$$

Finding the designated block in π is the same on average as finding the designated block in a partition with $H_m(\pi)$ equal blocks. But since $H_m(\pi)$ need not be an integer, one might take the reciprocal to obtain the probability interpretation: finding the designated block in π is the same on average as the occurrence of an event with probability $1/H_m(\pi)$.

Given a finite-valued random variable X with the values $\{x_1, ..., x_n\}$ with the probabilities $\{p_1, ..., p_n\}$, the *additive expectation* is: $E[X] = \sum_{i=1}^{n} p_i x_i$ and the *multiplicative expectation* is: $E_m[X] = \prod_{i=1}^{n} x_i^{p_i}$. Treating the block probability as a random variable defined on the blocks of a partition, all three entropies can be expressed as expectations:

$$\begin{aligned} H_m(\pi) &= E_m\left[\frac{1}{p_B}\right] \\ H(\pi) &= E\left[\log\left(\frac{1}{p_B}\right)\right] \\ h(\pi) &= E[1 - p_B] = 1 - E[p_B]. \end{aligned}$$

The usual (additive) Shannon entropy is then obtained as the $\log_2$ version of this "log-free" block-count entropy:

$$\log_2(H_m(\pi)) = \log\left(\prod_{B\in\pi}\left(\frac{1}{p_B}\right)^{p_B}\right) = \sum_{B\in\pi}\log\left(\left(\frac{1}{p_B}\right)^{p_B}\right) = \sum_{B\in\pi} p_B \log\left(\frac{1}{p_B}\right) = H(\pi).$$

Or viewed the other way around, $H_m(\pi) = 2^{H(\pi)}$.[22] The base 3 entropy encountered in the coin-weighing example is obtained by taking logs to that base: $H_3(\pi) = \log_3(H_m(\pi))$, and similarly for the Shannon entropy with natural logs: $H_e(\pi) = \log_e(H_m(\pi))$, or with common logs: $H_{10}(\pi) = \log_{10}(H_m(\pi))$.

Note that this relation $H_m(\pi) = 2^{H(\pi)}$ is a result, not a definition. The block-count entropy was defined from "scratch" in a manner similar to the usual Shannon entropy (which thus might be called the "$\log_2$-of-block-count entropy" or "binary-partition-count entropy"). In a partition of individual organisms by species, the interpretation of $2^{H(\pi)}$ (or $e^{H_e(\pi)}$ when natural logs are used) is the "number of equally common species" [24, p. 514]. MacArthur argued that this block-count

[22] Thus we expect the number-of-blocks entropy to be multiplicative where the usual Shannon entropy is additive (e.g., for stochastically independent partitions) and hence the subscript on $H_m(\pi)$.

entropy (where a block is a species) will "accord much more closely with our intuition..." (than the usual Shannon entropy).

The block-count entropy is the information measure that takes the count of a set (of like elements) as the measure of the information in the set. That is, for the discrete partition on U, each p_B is $\frac{1}{|U|}$ so the block-count entropy of the discrete partition is $H_m\left(\widehat{1}\right) = \prod_{u \in U} |U|^{1/|U|} = |U|$ which could also be obtained as $2^{H(\widehat{1})}$ since $H\left(\widehat{1}\right) = \log(|U|)$ is the $\log_2$-of-block-count Shannon entropy of $\widehat{1}$. Hence, the natural choice of unit for the block-count entropy is "blocks" (as in $H_m\left(\widehat{1}\right) = |U|$ blocks in the discrete partition on U). The block-count entropy of the discrete partition on an equiprobable 3^n element set is 3^n blocks. Hence the Shannon entropy with base 3 would be the $\log_3$-of-block-count entropy: $\log_3\left(H_m\left(\widehat{1}\right)\right) = \log_3(3^n) = n$ trits as in the coin-weighing example above. The block value relationship between the block-count entropy and the logical entropy in general is:

$$h(B) = 1 - p_B = 1 - \frac{1}{1/p_B} = 1 - \frac{1}{H_m(B)} = 1 - \frac{1}{2^{H(B)}} = 1 - \frac{1}{3^{H_3(B)}} = 1 - \frac{1}{e^{H_e(B)}} = 1 - \frac{1}{10^{H_{10}(B)}}$$

where $H_m(B) = 1/p_B = 2^{H(B)} = 3^{H_3(B)} = e^{H_e(B)} = 10^{H_{10}(B)}$.

4 Analogous Concepts for Shannon and Logical Entropies

4.1 Independent Partitions

It is sometimes asserted that "information" should be additive for independent[23] partitions but the underlying mathematical fact is that the block-count is multiplicative for independent partitions and Shannon chose to use the logarithm of the block-count as his measure of information.

If two partitions $\pi = \{B\}_{B \in \pi}$ and $\sigma = \{C\}_{C \in \sigma}$ are independent, then the block counts (i.e., the block entropies for the block-count entropy) multiply, i.e., $H_m(B \cap C) = \frac{1}{p_{B \cap C}} = \frac{1}{p_B}\frac{1}{p_C} = H_m(B) H_m(C)$. Hence for the multiplicative expectations we have:

$$H_m(\pi \vee \sigma) = \prod_{B,C} H_m(B \cap C)^{p_{B \cap C}} = \prod_{B,C} [H_m(B) H_m(C)]^{p_B p_C} = \left(\prod_{B \in \pi} H_m(B)^{p_B}\right)\left(\prod_{C \in \sigma} H_m(C)^{p_C}\right) = H_m(\pi) H_m(\sigma),$$

or taking logs to any desired base such as 2:

$$H(\pi \vee \sigma) = \log_2(H_m(\pi \vee \sigma)) = \log_2(H_m(\pi) H_m(\sigma)) = \log_2(H_m(\pi)) + \log_2(H_m(\sigma)) = H(\pi) + H(\sigma).$$

Thus for independent partitions, the block-count entropies multiply and the log-of-block-count entropies add. What happens to the logical entropies? We have seen that when the information in a partition is represented by its dit set $\operatorname{dit}(\pi)$, then the overlap in the dit sets of any two non-blob partitions is always non-empty. The dit set of the join of two partitions is just the union, $\operatorname{dit}(\pi \vee \sigma) = \operatorname{dit}(\pi) \cup \operatorname{dit}(\sigma)$, so that union is never a disjoint union (when the dit sets are non-empty). We have used the motivation of thinking of a partition-as-dit-set $\operatorname{dit}(\pi)$ as an "event" in a sample space $U \times U$ with the probability of that event being the logical entropy of the partition. The following proposition shows that this motivation extends to the notion of independence.

Proposition 3 *If π and σ are (stochastically) independent partitions, then their dit sets* $\operatorname{dit}(\pi)$ *and* $\operatorname{dit}(\sigma)$ *are independent as events in the sample space $U \times U$ (with equiprobable points).*

[23] Recall the "independent" means stochastic independence so that partitions π and σ are *independent* if for all $B \in \pi$ and $C \in \sigma$, $p_{B \cap C} = p_B p_C$.

For independent partitions π and σ, we need to show that the probability $m(\pi,\sigma)$ of the event $\mathrm{Mut}(\pi,\sigma) = \mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma)$ is equal to the product of the probabilities $h(\pi)$ and $h(\sigma)$ of the events $\mathrm{dit}(\pi)$ and $\mathrm{dit}(\sigma)$ in the sample space $U \times U$. By the assumption of independence, we have $\frac{|B\cap C|}{|U|} = p_{B\cap C} = p_B p_C = \frac{|B||C|}{|U|^2}$ so that $|B \cap C| = |B|\,|C| \,/\, |U|$. By the previous structure theorem for the mutual information set: $\mathrm{Mut}(\pi,\sigma) = \bigcup_{B\in\pi, C\in\sigma} (B - (B\cap C)) \times (C - (B \cap C))$, where the union is disjoint so that:

$$\begin{aligned}
|\mathrm{Mut}(\pi,\sigma)| &= \sum_{B\in\pi, C\in\sigma} (|B| - |B\cap C|)\,(|C| - |B\cap C|) \\
&= \sum_{B\in\pi, C\in\sigma} \left(|B| - \frac{|B|\,|C|}{|U|}\right)\left(|C| - \frac{|B|\,|C|}{|U|}\right) \\
&= \frac{1}{|U|^2} \sum_{B\in\pi, C\in\sigma} |B|\,(|U| - |C|)\,|C|\,(|U| - |B|) \\
&= \frac{1}{|U|^2} \sum_{B\in\pi} |B|\,|U - B| \sum_{C\in\sigma} |C|\,|U - C| \\
&= \frac{1}{|U|^2}\,|\mathrm{dit}(\pi)|\,|\mathrm{dit}(\sigma)|\,.
\end{aligned}$$

Hence under independence, the normalized dit count $m(\pi,\sigma) = \frac{|\mathrm{Mut}(\pi,\sigma)|}{|U|^2} = \frac{\mathrm{dit}(\pi)}{|U|^2}\frac{\mathrm{dit}(\sigma)}{|U|^2} = h(\pi)\,h(\sigma)$ of the mutual information set $\mathrm{Mut}(\pi,\sigma) = \mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma)$ is equal to product of the normalized dit counts of the partitions:

$$m(\pi,\sigma) = h(\pi)\,h(\sigma) \text{ if } \pi \text{ and } \sigma \text{ are independent.}\blacksquare$$

4.2 Mutual Information

For each of the major concepts in the information theory based on the usual Shannon measure, there should be a corresponding concept based on the normalized dit counts of logical entropy.[24] In the following sections, we give some of these corresponding concepts and results.

The logical mutual information of two partitions $m(\pi,\sigma)$ is the normalized dit count of the intersection of their dit-sets:

$$m(\pi,\sigma) = \tfrac{|\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)|}{|U\times U|}.$$

For Shannon's notion of mutual information, we might apply the Venn diagram heuristics using a block $B \in \pi$ and a block $C \in \sigma$. We saw before that the information contained in a block B was $H(B) = \log\left(\frac{1}{p_B}\right)$ and similarly for C while $H(B\cap C) = \log\left(\frac{1}{p_{B\cap C}}\right)$ would correspond to the union of the information in B and in C. Hence the overlap or "mutual information" in B and C could be motivated as the sum of the two informations minus the union:

$$I(B;C) = \log\left(\tfrac{1}{p_B}\right) + \log\left(\tfrac{1}{p_C}\right) - \log\left(\tfrac{1}{p_{B\cap C}}\right) = \log\left(\tfrac{1}{p_B p_C}\right) + \log(p_{B\cap C}) = \log\left(\tfrac{p_{B\cap C}}{p_B p_C}\right).$$

Then the (Shannon) *mutual information* in the two partitions is obtained by averaging over the mutual information for each pair of blocks from the two partitions:

[24] See Cover and Thomas' book [6] for more background on the standard concepts. The corresponding notions for the block-count entropy are obtained from the usual Shannon entropy notions by taking antilogs.

$$I(\pi;\sigma) = \sum_{B,C} p_{B\cap C} \log\left(\frac{p_{B\cap C}}{p_B p_C}\right).$$

The mutual information can be expanded to verify the Venn diagram heuristics:

$$I(\pi;\sigma) = \sum_{B\in\pi, C\in\sigma} p_{B\cap C} \log\left(\frac{p_{B\cap C}}{p_B p_C}\right) =$$
$$\sum_{B,C} p_{B\cap C} \log\left(p_{B\cap C}\right) + \sum_{B,C} p_{B\cap C} \log\left(\frac{1}{p_B}\right) + \sum_{B,C} p_{B\cap C} \log\left(\frac{1}{p_C}\right)$$
$$= -H(\pi\vee\sigma) + \sum_{B\in\pi} p_B \log\left(\frac{1}{p_B}\right) + \sum_{C\in\sigma} p_C \log\left(\frac{1}{p_C}\right) = H(\pi) + H(\sigma) - H(\pi\vee\sigma).$$

We will later see an important inequality, $I(\pi;\sigma) \geq 0$ (with equality under independence), and its logical version.

In the logical theory, the corresponding "modular law" follows from the inclusion-exclusion principle applied to dit-sets: $|\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)| = |\mathrm{dit}(\pi)| + |\mathrm{dit}(\sigma)| - |\mathrm{dit}(\pi)\cup\mathrm{dit}(\sigma)|$. Normalizing yields:

$$m(\pi,\sigma) = \frac{|\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)|}{|U|^2} = \frac{|\mathrm{dit}(\pi)|}{|U|^2} + \frac{|\mathrm{dit}(\sigma)|}{|U|^2} - \frac{|\mathrm{dit}(\pi)\cup\mathrm{dit}(\sigma)|}{|U|^2} = h(\pi) + h(\sigma) - h(\pi\vee\sigma).$$

Since the formulas concerning the logical and Shannon entropies often have similar relationships, e.g., $I(\pi;\sigma) = H(\pi) + H(\sigma) - H(\pi\vee\sigma)$ and $m(\pi,\sigma) = h(\pi) + h(\sigma) - h(\pi\vee\sigma)$, it is useful to also emphasize some crucial differences. One of the most important special cases is for two partitions that are (stochastically) independent. For independent partitions, it is immediate that $I(\pi;\sigma) = \sum_{B,C} p_{B\cap C} \log\left(\frac{p_{B\cap C}}{p_B p_C}\right) = 0$ but we have already seen that for the logical mutual information, $m(\pi,\sigma) > 0$ so long as neither partition is the blob $\widehat{0}$. However for independent partitions we have;

$$m(\pi,\sigma) = h(\pi)\, h(\sigma)$$

so the logical mutual information behaves like the probability of both events occurring in the case of independence (as it must since logical entropy concepts have direct probabilistic interpretations). For independent partitions, the relation $m(\pi,\sigma) = h(\pi)\, h(\sigma)$ means that the probability that a random pair is distinguished by both partitions is the same as the probability that it is distinguished by one partition *times* the probability that it is distinguished by the other partition. In simpler terms, for independent π and σ, the probability that π and σ distinguishes is the probability that π distinguishes times the probability that σ distinguishes.

It is sometimes convenient to think in the complementary terms of an equivalence relation "identifying." rather than a partition distinguishing. Since $h(\pi)$ can be interpreted as the probability that a random pair of elements from U are distinguished by π, i.e., as a distinction probability, its complement $1 - h(\pi)$ can be interpreted as an *identification probability*, i.e., the probability that a random pair is identified by π (thinking of π as an equivalence relation on U). In general,

$$[1-h(\pi)]\,[1-h(\sigma)] = 1 - h(\pi) - h(\sigma) + h(\pi)\, h(\sigma) = [1-h(\pi\vee\sigma)] + [h(\pi)\, h(\sigma) - m(\pi,\sigma]$$

which could also be rewritten as:

$$[1-h(\pi\vee\sigma)] - [1-h(\pi)]\,[1-h(\sigma)] = m(\pi,\sigma) - h(\pi)\, h(\sigma).$$

Hence:

$$\text{if } \pi \text{ and } \sigma \text{ are independent: } [1-h(\pi)]\,[1-h(\sigma)] = [1-h(\pi\vee\sigma)].$$

Thus if π and σ are independent, then the probability that the join partition $\pi \vee \sigma$ identifies is the probability that π identifies times the probability that σ identifies. In summary, if π and σ are independent, then:

$$\begin{aligned}\text{Binary-partition-count (Shannon) entropy: } & H(\pi \vee \sigma) = H(\pi) + H(\sigma) \\ \text{Block-count entropy: } & H_m(\pi \vee \sigma) = H_m(\pi) H_m(\sigma) \\ \text{Normalized-dit-count (logical) entropy: } & h(\pi \vee \sigma) = 1 - [1 - h(\pi)][1 - h(\sigma)].\end{aligned}$$

4.3 Cross Entropy and Divergence

Given a set partition $\pi = \{B\}_{B \in \pi}$ on a set U, the "natural" or Laplacian probability distribution on the blocks of the partition was $p_B = \frac{|B|}{|U|}$. The set partition π also determines the set of distinctions $\text{dit}(\pi) \subseteq U \times U$ and the logical entropy of the partition was the Laplacian probability of the dit-set as an event, i.e., $h(\pi) = \frac{|\text{dit}(\pi|}{|U \times U|} = \sum_B p_B(1 - p_B)$. But we may also "kick away the ladder" and generalize all the definitions to any finite probability distributions $p = \{p_1, ..., p_n\}$. A probability distribution p might be given by finite-valued random variables X on a sample space U where $p_i = \text{Prob}(X = x_i)$ for the finite set of distinct values x_i for $i = 1, ..., n$. Thus the logical entropy of the random variable X is: $h(X) = \sum_{i=1}^n p_i(1 - p_i) = 1 - \sum_i p_i^2$. The entropy is only a function of the probability distribution of the random variable, not its values, so we could also take it simply as a function of the probability distribution p, $h(p) = 1 - \sum_i p_i^2$. Taking the sample space as $\{1, ..., n\}$, the logical entropy is still interpreted as the probability that two independent draws will draw distinct points from $\{1, ..., n\}$. The further generalizations replacing probabilities by probability density functions and sums by integrals are straightforward but beyond the scope of this paper (which is focused on conceptual foundations rather than mathematical developments).

Given two probability distributions $p = \{p_1, ..., p_n\}$ and $q = \{q_1, ..., q_n\}$ on the same sample space $\{1, ..., n\}$, we can again consider the drawing of a pair of points but where the first drawing is according to p and the second drawing according to q. The probability that the pair of points is distinct would be a natural and more general notion of logical entropy which we will call the:

$$\textit{logical cross entropy: } h(p\|q) = \sum_i p_i(1 - q_i) = 1 - \sum_i p_i q_i = \sum_i q_i(1 - p_i) = h(q\|p)$$

which is symmetric. The logical cross entropy is the same as the logical entropy when the distributions are the same, i.e.,

$$\text{if } p = q, \text{ then } h(p\|q) = h(p).$$

The notion of *cross entropy* in conventional information theory is: $H(p\|q) = \sum_i p_i \log\left(\frac{1}{q_i}\right)$ which is not symmetrical due to the asymmetric role of the logarithm, although if $p = q$, then $H(p\|q) = H(p)$. Then the *Kullback-Leibler divergence* $D(p\|q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$ is defined as a measure of the distance or divergence between the two distributions where $D(p\|q) = H(p\|q) - H(p)$. The *information inequality* is: $D(p\|q) \geq 0$ with equality if and only if $p_i = q_i$ for $i = 1, ..., n$ [6, p. 26]. Given two partitions π and σ, the inequality $I(\pi; \sigma) \geq 0$ is obtained by applying the information inequality to the two distributions $\{p_{B \cap C}\}$ and $\{p_B p_C\}$ on the sample space $\{(B, C) : B \in \pi, C \in \sigma\} = \pi \times \sigma$:

$$I(\pi; \sigma) = \sum_{B,C} p_{B \cap C} \log\left(\frac{p_{B \cap C}}{p_B p_C}\right) = D(\{p_{B \cap C}\} \| \{p_B p_C\}) \geq 0 \text{ with equality under independence.}$$

But starting afresh, one might ask: "What is the natural measure of the difference or distance between two probability distributions $p = \{p_1, ..., p_n\}$ and $q = \{q_1, ..., q_n\}$ that would always be

non-negative, and would be zero if and only they are equal?" The (Euclidean) distance between the two points in $\mathbb{R}^n$ would seem to be the "logical" answer—so we take that distance (squared) as the definition of the:

$$\textit{logical divergence} \text{ (or } \textit{logical relative entropy}\text{): } d(p\|q) = \sum_i (p_i - q_i)^2,$$

which is symmetric and non-negative. We have component-wise:

$$0 \leq (p_i - q_i)^2 = p_i^2 - 2p_i q_i + q_i^2 = 2\left[\tfrac{1}{n} - p_i q_i\right] - \left[\tfrac{1}{n} - p_i^2\right] - \left[\tfrac{1}{n} - q_i^2\right]$$

so that taking the sum for $i = 1, ..., n$ gives:

$$0 \leq d(p\|q) = \sum_i (p_i - q_i)^2 = 2\left[1 - \sum_i p_i q_i\right] - \left[1 - \sum_i p_i^2\right] - \left[1 - \sum_i q_i^2\right] = 2h(p\|q) - h(p) - h(q).$$

Thus we have the:

$$\boxed{0 \leq d(p\|q) = 2h(p\|q) - h(p) - h(q) \text{ with equality if and only if } p_i = q_i \text{ for } i = 1, ..., n}$$

Logical information inequality.

If we take $h(p\|q) - \frac{1}{2}[h(p) + h(q)]$ as the *Jensen difference* [26, p. 25] between the two distributions, then the logical divergence is twice the Jensen difference. The half-and-half probability distribution $\frac{p+q}{2}$ that mixes p and q has the logical entropy of $h\left(\frac{p+q}{2}\right) = \frac{h(p\|q)}{2} + \frac{h(p)+h(q)}{4}$ so that:

$$d(p\|q) = 4\left[h\left(\tfrac{p+q}{2}\right) - \tfrac{1}{2}\{h(p) + h(q)\}\right] \geq 0.$$

The logical information inequality tells us that "mixing increases logical entropy" (or, to be precise, mixing does not decrease logical entropy) which also follows from the fact that logical entropy $h(p) = 1 - \sum_i p_i^2$ is a concave function.

An important special case of the logical information inequality is when $p = \{p_1, ..., p_n\}$ is the uniform distribution with all $p_i = \frac{1}{n}$. Then $h(p) = 1 - \frac{1}{n}$ where the probability that a random pair is distinguished (i.e., the random variable X with $\text{Prob}(X = x_i) = p_i$ has different values in two independent samples) takes the specific form of the probability $1 - \frac{1}{n}$ that the second draw gets a different value than the first. It may at first seem counterintuitive that in this case the cross entropy is $h(p\|q) = h(p) + \sum_i p_i (p_i - q_i) = h(p) + \sum_i \frac{1}{n}\left(\frac{1}{n} - q_i\right) = h(p) = 1 - \frac{1}{n}$ for *any* $q = \{q_1, ..., q_n\}$. But $h(p\|q)$ is the probability that the two points, say i and i', in the sample space $\{1, ..., n\}$ are distinct when one draw was according to p and the other according to q. Taking the first draw according to q, the probability that the second draw is distinct from whatever point was determined in the first draw is indeed $1 - \frac{1}{n}$ (regardless of probability q_i of the point drawn on the first draw). Then the divergence $d(p\|q) = 2h(p\|q) - h(p) - h(q) = \left(1 - \frac{1}{n}\right) - h(q)$ is a non-negative measure of how much the probability distribution q diverges from the uniform distribution. It is simply the difference in the probability that a random pair will be distinguished by the uniform distribution and by q. Also since $0 \leq d(p\|q)$, this shows that among all probability distributions on $\{1, ..., n\}$, the uniform distribution has the maximum logical entropy. In terms of partitions, the n-block partition with $p_B = \frac{1}{n}$ has maximum logical entropy among all n-block partitions. In the case of $|U|$ divisible by n, the equal n-block partitions make more distinctions than any of the unequal n-block partitions on U.

For any partition π with the n block probabilities $\{p_B\}_{B \in \pi} = \{p_1, ..., p_n\}$:

$$h(\pi) \leq 1 - \tfrac{1}{n} \text{ with equality if and only if } p_1 = ... = p_n = \tfrac{1}{n}.$$

For the corresponding results in the Shannon's information theory, we can apply the information inequality $D(p\|q) = H(p\|q) - H(p) \geq 0$ with q as the uniform distribution $q_1 = ... = q_n = \frac{1}{n}$. Then $H(p\|q) = \sum_i p_i \log\left(\frac{1}{1/n}\right) = \log(n)$ so that: $H(p) \leq \log(n)$ or in terms of partitions:

$$H(\pi) \leq \log_2(|\pi|) \text{ with equality if and only if the probabilities are equal}$$

or, in base-free terms,

$$H_m(\pi) \leq |\pi| \text{ with equality if and only if the probabilities are equal.}$$

The three entropies take their maximum values (for fixed number of blocks $|\pi|$) at the partitions with equiprobable blocks.

In information theory texts, it is customary to graph the case of $n = 2$ where the entropy is graphed as a function of $p_1 = p$ with $p_2 = 1 - p$. The Shannon entropy function $H(p) = -p\log(p) - (1-p)\log(1-p)$ looks somewhat like an inverted parabola with its maximum value of $\log(n) = \log(2) = 1$ at $p = .5$. The logical entropy function $h(p) = 1 - p^2 - (1-p)^2 = 2p - 2p^2 = 2p(1-p)$ is an inverted parabola with its maximum value of $1 - \frac{1}{n} = 1 - \frac{1}{2} = .5$ at $p = .5$. The block-count entropy $H_m(p) = \left(\frac{1}{p}\right)^p \left(\frac{1}{1-p}\right)^{1-p} = 2^{H(p)}$ is an inverted U-shaped curve that starts and ends at $1 = 2^{H(0)} = 2^{H(1)}$ and has its maximum at $2 = 2^{H(.5)}$.

4.4 Summary of Analogous Concepts and Results

	Shannon Entropy	Logical Entropy
Block Entropy	$H(B) = \log(1/p_B)$	$h(B) = 1 - p_B$
Relationship	$H(B) = \log\left(\frac{1}{1-h(B)}\right)$	$h(B) = 1 - \frac{1}{2^{H(B)}}$
Entropy	$H(\pi) = \sum p_B \log(1/p_B)$	$h(\pi) = \sum p_B(1-p_B)$
Mutual Information	$I(\pi;\sigma) = H(\pi) + H(\sigma) - H(\pi \vee \sigma)$	$m(\pi,\sigma) = h(\pi) + h(\sigma) - h(\pi \vee \sigma)$
Independence	$I(\pi;\sigma) = 0$	$m(\pi,\sigma) = h(\pi)h(\sigma)$
Independence & Joins	$H(\pi \vee \sigma) = H(\pi) + H(\sigma)$	$h(\pi \vee \sigma) = 1 - [1-h(\pi)][1-h(\sigma)]$
Cross Entropy	$H(p\|\|q) = \sum p_i \log(1/q_i)$	$h(p\|\|q) = \sum p_i(1-q_i)$
Divergence	$D(p\|\|q) = H(p\|\|q) - H(p)$	$d(p\|\|q) = 2h(p\|\|q) - h(p) - h(q)$
Information Inequality	$D(p\|\|q) \geq 0$ with $=$ iff $p_i = q_i \forall i$	$d(p\|\|q) \geq 0$ with $=$ iff $p_i = q_i \forall i$
Info. Ineq. Sp. Case	$I(\pi;\sigma) = D(\{p_{B\cap C}\} \parallel \{p_B p_C\}) \geq 0$ with equality under independence	$d(\{p_{B\cap C}\} \parallel \{p_B p_C\}) \geq 0$ with equality under independence.

5 Concluding Remarks

In the duality of subsets of a set with partitions on a set, we found that the elements of a subset were dual to the distinctions (dits) of a partition. Just as the finite probability theory for events started by taking the size of a subset ("event") S normalized to the size of the finite universe U as the probability $\text{Prob}(S) = \frac{|S|}{|U|}$, so it would be natural to consider the corresponding theory that would associate with a partition π on a finite U, the size $|\text{dit}(\pi)|$ of the set of distinctions of the partition normalized by the total number of ordered pairs $|U \times U|$. This number $h(\pi) = \frac{|\text{dit}(\pi)|}{|U \times U|}$ was called the logical entropy of π and could be interpreted as the probability that a randomly picked (with replacement) pair of elements from U is distinguished by the partition π, just as $\text{Prob}(s) = \frac{|S|}{|U|}$ is the probability that a randomly picked element from U is an element of the subset S. Hence this notion of logical entropy arises naturally out of the logic of partitions that is dual to the usual logic of subsets.

The question immediately arises of the relationship with Shannon's concept of entropy. Following Shannon's definition of entropy, there has been a veritable plethora of suggested alternative entropy concepts [20]. Logical entropy is *not* an alternative entropy concept intended to displace Shannon's concept any more than is the block-count entropy concept. Instead, I have argued that the dit-count, block-count, and binary-partition-count concepts of entropy should be seen as three ways to

measure that same "information" expressed in its most atomic terms as distinctions. The block-count entropy, although it can be independently defined, is trivially related to Shannon's binary-partition-count concept—just take antilogs. The relationship of the logical concept of entropy to the Shannon concept is a little more subtle but is quite simple at the level of blocks $B \in \pi$: $h(B) = 1 - p_B$, $H_m(B) = \frac{1}{p_B}$, and $H(B) = \log\left(\frac{1}{p_B}\right)$ so that eliminating the probability, we have:

$$\begin{aligned} h(B) &= 1 - \frac{1}{H_m(B)} \\ &= 1 - \frac{1}{2^{H(B)}}. \end{aligned}$$

Then the logical and additive entropies for the whole partition are obtained by taking the (additive) expectation of the block entropies while the block-count entropy is the multiplicative expectation of the block entropies:

$$\begin{aligned} H_m(\pi) &= \prod_{B\in\pi} \left(\frac{1}{p_B}\right)^{p_B} \\ H(\pi) &= \sum_{B\in\pi} p_B \log\left(\frac{1}{p_B}\right) \\ h(\pi) &= \sum_{B\in\pi} p_B (1 - p_B). \end{aligned}$$

In conclusion, the simple root of the matter is three different ways to "measure" the distinctions that generate an n-element set. Consider a 4 element set. One measure of the distinctions that distinguish a set of 4 elements is its cardinality 4, and that measure leads to the block-count entropy. Another measure of that set is $\log_2(4) = 2$ which can be interpreted as the minimal number of binary partitions necessary: (a) to single out any designated element as a singleton (search interpretation) or, equivalently, (b) to distinguish all the elements from each other (distinction interpretation). That measure leads to Shannon's entropy formula. And the third measure is the (normalized) count of distinctions (counted as ordered pairs) necessary to distinguish all the elements from each other, i.e., $\frac{4\times4-4}{4\times4} = \frac{12}{16} = \frac{3}{4}$, which yields the logical entropy formula. These measures stand in the block value relationship: $\frac{3}{4} = 1 - \frac{1}{4} = 1 - \frac{1}{2^2}$. It is just a matter of:

1. counting the elements distinguished (block-count entropy),
2. counting the binary partitions needed to distinguish the elements (Shannon entropy), or
3. counting the (normalized) distinctions themselves (logical entropy).

References

[1] Aczél, J. and Z. Daróczy 1975. *On Measures of Information and Their Characterization*. New York: Academic Press.

[2] Adelman, M. A. 1969. Comment on the H Concentration Measure as a Numbers-Equivalent. *Review of Economics and Statistics*. 51: 99-101.

[3] Baclawski, Kenneth and Gian-Carlo Rota 1979. *An Introduction to Probability and Random Processes*. Unpublished typescript. 467 pages. Download available at: http://www.ellerman.org.

[4] Bhargava, T. N. and V. R. R. Uppuluri 1975. On an Axiomatic Derivation of Gini Diversity, With Applications. *Metron*. 33: 41-53.

[5] Birkhoff, Garrett 1948. *Lattice Theory*. New York: American Mathematical Society.

[6] Cover, Thomas and Joy Thomas 1991. *Elements of Information Theory*. New York: John Wiley.

[7] Finberg, David, Matteo Mainetti and Gian-Carlo Rota 1996. The Logic of Commuting Equivalence Relations. In *Logic and Algebra*. Aldo Ursini and Paolo Agliano eds., New York: Marcel Dekker: 69-96.

[8] Friedman, William F. 1922. *The Index of Coincidence and Its Applications in Cryptography*. Geneva IL: Riverbank Laboratories.

[9] Ganeshaiah, K. N., K. Chandrashekara and A. R. V. Kumar 1997. Avalanche Index: A new measure of biodiversity based on biological heterogeneity of communities. *Current Science*. 73: 128-33.

[10] Gini, Corrado 1912. *Variabilità e mutabilità*. Bologna: Tipografia di Paolo Cuppini.

[11] Gini, Corrado 1955. Variabilità e mutabilità. In *Memorie di metodologica statistica*. E. Pizetti and T. Salvemini eds., Rome: Libreria Eredi Virgilio Veschi.

[12] Good, I. J. 1979. A.M. Turing's statistical work in World War II. *Biometrika*. 66 (2): 393-6.

[13] Good, I. J. 1982. Comment (on Patil and Taillie: Diversity as a Concept and its Measurement). *Journal of the American Statistical Association*. 77 (379): 561-3.

[14] Gray, Robert M. 1990. *Entropy and Information Theory*. New York: Springer-Verlag.

[15] Hartley, Ralph V. L. 1928. Transmission of information. *Bell System Technical Journal*. 7 (3, July): 535-63.

[16] Havrda, J. H. and F. Charvat 1967. Quantification Methods of Classification Processes: Concept of Structural α-Entropy. *Kybernetika (Prague)*. 3: 30-35.

[17] Herfindahl, Orris C. 1950. *Concentration in the U.S. Steel Industry*. Unpublished doctoral dissertation, Columbia University.

[18] Hirschman, Albert O. 1945. *National power and the structure of foreign trade*. Berkeley: University of California Press.

[19] Hirschman, Albert O. 1964. The Paternity of an Index. *American Economic Review*. 54 (5): 761-2.

[20] Kapur, J.N. 1994. *Measures of Information and Their Applications*. New Delhi: Wiley Eastern.

[21] Kolmogorov, A.N. 1956. *Foundations of the Theory of Probability*. New York: Chelsea.

[22] Kullback, Solomon 1976. *Statistical Methods in Cryptanalysis*. Walnut Creek CA: Aegean Park Press.

[23] Lawvere, F. William and Robert Rosebrugh 2003. *Sets for Mathematics*. Cambridge: Cambridge University Press.

[24] MacArthur, Robert H. 1965. Patterns of Species Diversity. *Biol. Rev.* 40: 510-33.

[25] Patil, G. P. and C. Taillie 1982. Diversity as a Concept and its Measurement. *Journal of the American Statistical Association*. 77 (379): 548-61.

[26] Rao, C. Radhakrishna 1982. Diversity and Dissimilarity Coefficients: A Unified Approach. *Theoretical Population Biology*. 21: 24-43.

[27] Rejewski, M. 1981. How Polish Mathematicians Deciphered the Enigma. *Annals of the History of Computing*. 3: 213-34.

[28] Rényi, Alfréd 1965. On the Theory of Random Search. *Bull. Am. Math. Soc.* 71: 809-28.

[29] Rényi, Alfréd 1970. *Probability Theory*. Laszlo Vekerdi (trans.), Amsterdam: North-Holland.

[30] Rényi, Alfréd 1976. *Selected Papers of Alfréd Rényi: Volumes 1,2, and 3*. Pal Turan (editor), Budapest: Akademiai Kiado.

[31] Ricotta, Carlo and Laszlo Szeidl 2006. Towards a unifying approach to diversity measures: Bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology*. 70: 237-43.

[32] Shannon, Claude E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*. 27: 379-423; 623-56.

[33] Simpson, Edward Hugh 1949. Measurement of Diversity. *Nature*. 163: 688.

[34] Stigler, Stephen M. 1999. *Statistics on the Table*. Cambridge: Harvard University Press.

[35] Tsallis, C. 1988. Possible Generalization for Boltzmann-Gibbs statistics. *J. Stat. Physics*. 52: 479-87.

[36] Vajda, I. 1969. A Contribution to Informational Analysis of Patterns. In *Methodologies of Pattern Recognition*. Satosi Watanabe ed., New York: Academic Press: 509-519.