Int. J. General Systems, Vol. 17, pp. 211–240
Reprints available directly from the publisher
Photocopying permitted by license only

FUZZINESS VS. PROBABILITY

BART KOSKO

Electrical Engineering Department, Signal and Image Processing Institute, University of Southern California, Los Angeles, California 90089-0272, USA

(*Received 14 November 1989; in final form 2 March 1990*)

So far as the laws of mathematics refer to reality, they are not certain. And so far as they are certain, they do not refer to reality.

Albert Einstein

Fuzziness is explored as an alternative to randomness for describing uncertainty. The new *sets-as-points* geometric view of fuzzy sets is developed. This view identifies a fuzzy set with a point in a unit hypercube and a nonfuzzy set with a vertex of the cube. Paradoxes of two-valued logic and set theory, such as Russell's paradox, correspond to the midpoint of the fuzzy cube. The fundamental questions of fuzzy theory—How fuzzy is a fuzzy set? How much is one fuzzy set a subset of another?—are answered geometrically with the Fuzzy Entropy Theorem, the Fuzzy Subsethood Theorem, and the Entropy-Subsethood Theorem. A new geometric proof of the Subsethood Theorem is given, a corollary of which is that the apparently probabilistic relative frequency n_A/N turns out to be the deterministic subsethood $S(X, A)$, the degree to which the sample space X is contained in its subset A. So the frequency of successful trials is viewed as the degree to which all trials are successful. Recent Bayesian polemics against fuzzy theory are examined in light of the new sets-as-points theorems.

INDEX TERMS: Probability Theory, fuzzy set theory, fuzzy subsethood, geometry of fuzzy sets.

1. FUZZINESS IN A PROBABILISTIC WORLD

Is uncertainty the same as randomness? If we are not sure about something, is it only up to chance? Do the notions of likelihood and probability exhaust our notions of uncertainty?

Many people, trained in probability and statistics, believe so. Some even say so, and say so loudly. These voices are often heard in the Bayesian camp of statistics, where probability is viewed, not as a frequency or other objective testable quantity, but as a subjective *state of knowledge.*

Bayesian physicist E. T. Jaynes says[6] that

> any method of inference in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to Laplace's [probability], or inconsistent.

He claims physicist R. T. Cox[3] has proven this as a theorem, a claim we examine below.

More recently, Bayesian statistician Dennis Lindley[13] issued an explicit challenge:

probability is the only sensible description of uncertainty and is adequate for all problems involving uncertainty. All other methods are inadequate.

Lindley directs his challenge in large part at *fuzzy theory*, the theory that *all things admit degrees*, but admit them deterministically. This article accepts the probabilist's challenge from the fuzzy viewpoint—admitting but ignoring other approaches to uncertainty, such as Dempster–Shafer belief function theory—by defending fuzziness from new geometric first principles and by questioning the reasonableness and the axiomatic status of randomness. The new view is the *sets-as-points view*[11] of fuzzy sets: A fuzzy set is a point in a unit hypercube and a nonfuzzy set is a corner of the hypercube.

There are conceptual and theoretical differences between randomness and fuzziness. Some can be illustrated with examples. Some can be proven with theorems, as we show below.

There are also many similarities. The chief, but superficial, similarity is that both systems describe uncertainty with numbers in the unit interval $[0, 1]$. This ultimately means that both systems describe uncertainty numerically. The structural similarity is that both systems combine sets and propositions associatively, commutatively, and distributively. The key distinction concerns how the systems deal simultaneously with a thing A and its opposite A^c.

Questions raise doubt, and doubt suggests room for change. So to commence the exposition, consider the following two questions, one fuzzy and the other probabilistic:

i) Is it always and everywhere true that $A \cap A^c = \varnothing$?

ii) Who can *derive* the conditional probability operator

$$P(B|A) = \frac{P(A \cap B)}{P(A)}?$$

The second question may appear less fundamental than the first question, which asks whether fuzziness exists. The Entropy-Subsethood Theorem below shows that the first and second questions are connected: How fuzzy a fuzzy set A is can be measured by how much the superset $A \cup A^c$ is a subset of its own subset $A \cap A^c$, a paradoxical relationship unique to fuzzy theory. In contrast, in probability theory this state of affairs is impossible (has zero probability): $P(A \cap A^c | A \cup A^c) = P(\varnothing | X) = 0$, where X is the sample space or "sure event" and the empty set $\varnothing$ is the "impossible event".

The conditioning or subsethood in the second question is at the heart of Bayesian probabilistic systems. The absence of a first-principles derivation of $P(B|A)$ in itself may be acceptable. One simply agrees to take the ratio relationship as an axiom. The problem is that the new sets-as-points view of fuzzy sets *derives* its conditioning operator as a theorem from first principles. The history of science suggests that systems that hold theorems as axioms continue to evolve.

The first question asks whether the law of noncontradiction—one of Aristotle's three "laws of thought" along with the laws of excluded middle, $A \cup A^c = X$, and identity, $A = A$—can be violated. Set fuzziness occurs when, and only when, it is violated. Classical logic and set theory assume that the law of noncontradiction,

and equivalently the law of excluded middle, is never violated. That is what makes the classical theory black or white. Fuzziness begins where Western logic ends.

2. RANDOMNESS VS. AMBIGUITY: WHETHER VS. HOW MUCH

Fuzziness describes *event ambiguity*. It measures the degree to which an event occurs, not whether it occurs. Randomness describes the uncertainty of *event occurrence*. An event occurs or not, and you can bet on it. At issue is the nature of the occurring event: whether it itself is uncertain in any way, in particular whether it can be unambiguously distinguished from its opposite.

Whether an event occurs is "random". To what degree it occurs is fuzzy. Whether an ambiguous event occurs—as when we say there is 20% chance of light rain tomorrow—involves compound uncertainties, the probability of a fuzzy event.

In practice we regularly apply probabilities to fuzzy events: small errors, satisfied customers, A students, safe investments, developing countries, noisy signals, spiking neurons, dying cells, charged particles, nimbus clouds, planetary atmospheres, galactic clusters. We understand that, at least around the edges, some satisfied customers can be somewhat unsatisfied, some A students might equally be B+ students, some stars are as much in a galactic cluster as out of it. Events can more or less smoothly transition to their opposites, making classification hard near the midpoint of the transition. But in theory—in formal descriptions and in textbooks—the events and their opposites are black and white. A hill is a mountain if it is at least x meters tall, not a mountain if it is one micron less than x in height. Every molecule in the universe either is or is not a pencil molecule, even those hovering above the pencil's surface.

Consider some further examples. The probability that this essay gets published is one thing. The degree to which it gets published is another. The essay may be edited in hundreds of ways. Or the essay may be marred with typographical errors, and so on.

Question: Does quantum mechanics deal with the probability that an unambiguous electron occupies spacetime points? Or does it deal with the degree to which an electron, or an electron smear, occurs at spacetime points? Does $|\psi|^2 dV$ measure the probability that a random-point electron occurs in infinitesimal volume dV? Or[12] does it measure the degree to which a deterministic electron cloud occurs in dV? Different interpretation, different universe. Perhaps even existence admits degrees.

Suppose there is 50% chance that there is an apple in the refrigerator (electron in a cell[12]). That is one state of affairs, perhaps arrived at through frequency calculations or a Bayesian state of knowledge. Now suppose there is half an apple in the refrigerator. That is another state of affairs. Both states of affairs are superficially equivalent in terms of their numerical uncertainty. Yet physically, ontologically, they are distinct. One is "random", the other fuzzy.

If events are *assumed* unambiguous, as in balls-in-urns experiments, there is no fuzziness. Only randomness remains. But when discussing the physical universe, every assertion of event ambiguity or nonambiguity is an empirical *hypothesis*. This is habitually overlooked when applying probability theory. Years of such oversight are perhaps responsible for the deeply entrenched sentiment that uncertainty is randomness, and randomness alone. The silent assumption of universal nonambiguity is akin to the pre-relativistic assumption of an uncurved

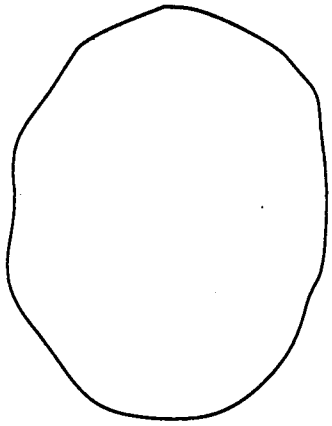

Figure 1 Inexact oval. Which statement better describes the situation: (a) "It is probably an ellipse" or (b) "It is a fuzzy ellipse"?

universe. $A \cap A^c = \varnothing$ is the "parallel postulate" of classical set theory and logic, indeed of Western thought.

If fuzziness is a unique type of uncertainty, if fuzziness exists, the physical consequences are universal, and the sociological consequence is startling: scientists, especially physicists, have overlooked an entire mode of reality.

Fuzziness is a type of deterministic uncertainty. Ambiguity is a property of physical phenomena. Unlike fuzziness, probability dissipates with increasing information. After the fact "randomness" looks like fiction. (This is especially awkward since in general the laws of science are time reversible, invariant if time t is replaced with time $-t$. Where does the randomness go?) Yet there is as much ambiguity after a sample-space experiment as before. Increasing information tends to specify the degrees of occurrence. Even if science had run its course and all the facts were in, a platypus would remain only roughly a mammal, a large hill only roughly a mountain, an oval squiggle only roughly an ellipse. Fuzziness does not require that God play dice.

Consider the inexact oval in Figure 1. Does it make more sense to say that the oval is *probably* a circle (or ellipse), or that it *is* a fuzzy ellipse? There is nothing random about the matter. The situation is deterministic: All the facts are in. Yet uncertainty remains. The uncertainty is due to the simultaneous occurrence of two properties: to some extent the inexact oval is an ellipse and to some extent it is not an ellipse.

More formally, is $m_A(x)$, the degree to which element x belongs to fuzzy set A, simply the probability that x is in A? Is $m_A(x) = \text{Prob}\{x \in A\}$ true? Cardinality-wise, sample spaces cannot be too big. Else a positive measure cannot be both countably additive and finite, and thus a probability measure. The space of all possible oval figures is too big, since there are more of these than real numbers. Almost all sets are too big to define probabilities, yet fuzzy sets can always be defined.

$\text{Prob}\{x \in A\}$ might be interpreted as the probability of a fuzzy event, the probability that element x belongs to fuzzy set A with degree $m_A(x)$. Rarely indeed then should the equality $\text{Prob}\{x \in A\} = m_A(x)$ occur.

But this is not the intended interpretation of the assertion $\text{Prob}\{x \in A\} = m_A(x)$. Instead set A is not fuzzy. The element x either is or is not an element of set A. We do not know which, and we describe this uncertainty with the probability $\text{Prob}\{x \in A\}$. But then surely $\text{Prob}\{x \in A\} \neq m_A(x)$. For example, $\text{Prob}\{x \in A \cap A^c\} = 0$ and $\text{Prob}\{x \in A \cup A^c\} = 1$ for every nonfuzzy set A. Yet $m_{A \cap A^c}(x) > 0$ and $m_{A \cup A^c}(x) < 1$ for every properly fuzzy set A.

Probability theory is a chapter in the book of finite measure theory. Many probabilists do not care for this classification, but they fall back upon it when defining terms.[7] How reasonable is it to believe that finite measure theory—ultimately, the summing of nonnegative numbers to unity—exhaustively describes the (quantum-mechanical) universe? Does it really describe *any* thing?

Surely from time to time every probabilist wonders whether probability describes anything real. From Democritus to Einstein, there has been the suspicion that, as David Hume[5] put it,

> though there be no such thing as *chance* in the world, our ignorance of the real cause of any event has the same influence on the understanding and begets a like species of belief.

When we model noisy processes by extending differential equations to stochastic differential equations, it seems we introduce the formalism only as a working approximation to several underlying unspecified processes, processes that presumably obey deterministic differential equations. In this sense conditional expectations and martingale techniques might seem reasonably applied, for example, to stock options or commodity futures phenomena, where the behavior involved consists of aggregates of aggregates of aggregates. The same techniques seem less reasonably applied to quarks, leptons, and void.

3. THE UNIVERSE AS A FUZZY SET

The world, as Wittgenstein[11] observed, is everything that is the case. In this spirit we can summarize the ontological case for fuzziness: The universe consists of all subsets of the universe. The only subsets of the universe that are not fuzzy are the constructs of classical mathematics. All other sets—sets of particles, cells, tissues, people, ideas, galaxies—in principal contain elements to different degrees. Their membership is partial, graded, inexact, ambiguous, or uncertain.

The same universal circumstance holds at the level of logic and truth. The only logically true or false statements—statements S with truth value $t(S)$ in $\{0, 1\}$—are tautologies, theorems, and contradictions. If S is any statement about the universe, an empirical statement, then $0 < t(S) < 1$ holds by the canons of scientific method and by the lack of a single demonstrated factual statement S with $t(S) = 1$ or $t(S) = 0$. That is the thrust of Einstein's quote above.

Fuzziness arises from the ambiguity between a thing A and its opposite A^c. If we do not know A with certainty, we do not know A^c with certainty either. Else by double negation we would know A with certainty. This produces nondegenerate *overlap*: $A \cap A^c \neq \varnothing$, which breaks the "law of noncontradiction". Equivalently, this also produces nondegenerate *underlap*:[10] $A \cup A^c \neq X$, which breaks the "law of excluded middle". Here X is the ground set or universe of discourse. Recall[4] that these laws are never broken in probabilistic or stochastic logics—$P(A \text{ and not-}A) = 0$ and $P(A \text{ or not-}A) = 1$—even though they are broken with many, perhaps most, human utterances. Nor are probability measures allowed to take such fuzzy

sets as arguments. The sets must first be quantized, rounded off, or defuzzified to the nearest nonfuzzy set. So the question arises: How mathematically natural are fuzzy sets?

4. THE GEOMETRY OF FUZZY SETS: SETS AS POINTS

It helps to see the geometry of fuzzy sets when discussing fuzziness. To date this visual property has been overlooked. The emphasis has instead been on interpreting fuzzy sets as membership functions, mappings m_A from domain X to range $[0, 1]$. But functions are hard to visualize. Membership functions are often pictured as two-dimensional graphs, with the domain X misleadingly represented as one-dimensional. The geometry of fuzzy sets involves both the domain $X = \{x_1, \ldots, x_n\}$ and the range $[0, 1]$ of mappings $m_A: X \to [0, 1]$. The geometry of fuzzy sets is a great aid in understanding fuzziness, defining fuzzy concepts, and proving fuzzy theorems. Visualizing this geometry may by itself be the most powerful argument for fuzziness.

The geometry of fuzzy sets is revealed by asking an odd question: What does the fuzzy power set $F(2^X)$, the set of all fuzzy subsets of X, look like? Answer: A cube. What does a fuzzy set look like? A point in a cube. The set of all fuzzy subsets is the unit hypercube $I^n = [0, 1]^n$. A fuzzy set is any point[11] in the cube I^n. So (X, I^n) is the fundamental measurable space of (finite) fuzzy theory. The theory of fuzzy sets—more accurately, the theory of *continuous* sets—can be taught on a Rubik's cube.

Vertices of the cube I^n are nonfuzzy sets. So the ordinary power set 2^X, the set of all 2^n nonfuzzy subsets of X, is the Boolean n-cube $B^n: 2^X = B^n$. Fuzzy sets fill in the lattice B^n to produce the solid cube $I^n: F(2^X) = I^n$.

Consider the set of two elements $X = \{x_1, x_2\}$. The nonfuzzy power set 2^X contains four sets: $2^X = \{\varnothing, X, \{x_1\}, \{x_2\}\}$. These four sets correspond respectively to the four bit vectors (0 0), (1 1), (1 0), and (0 1). The 1s and 0s indicate the presence or absence of the ith element x_i in the subset. More abstractly, each subset A is uniquely defined by one of the two-valued membership functions $m_A: X \to \{0, 1\}$.

Now consider the fuzzy subsets of X. The fuzzy subset $A = (\frac{1}{3}\ \frac{3}{4})$ can be viewed as one of the continuum-many continuous-valued membership functions $m_A: X \to [0, 1]$. Indeed this is the classical Zadeh[16] *sets-as-functions* definition of fuzzy sets. In this example element x_1 belongs to, or fits in, subset A a little bit—to degree $\frac{1}{3}$. Element x_2 has more membership than not at $\frac{3}{4}$. Analogous to the bit vector representation of finite (countable) sets, we say that A is represented by the *fit vector* $(\frac{1}{3}\ \frac{3}{4})$. The element $m_A(x_i)$ is the ith *fit*[10] or *fuzzy unit* value. The set-as-points view then geometrically represents the fuzzy subset A as a point in I^2, the unit square, as in Figure 2.

The midpoint of the cube I^n is maximally fuzzy. All its membership values are $\frac{1}{2}$. The midpoint is unique in two respects. First, the midpoint is the only set A that not only equals its own opposite A^c but equals its own overlap and underlap as well:

$$A = A \cap A^c = A \cup A^c = A^c.$$

Second, the midpoint is the only point in the cube I^n that is equidistant to each

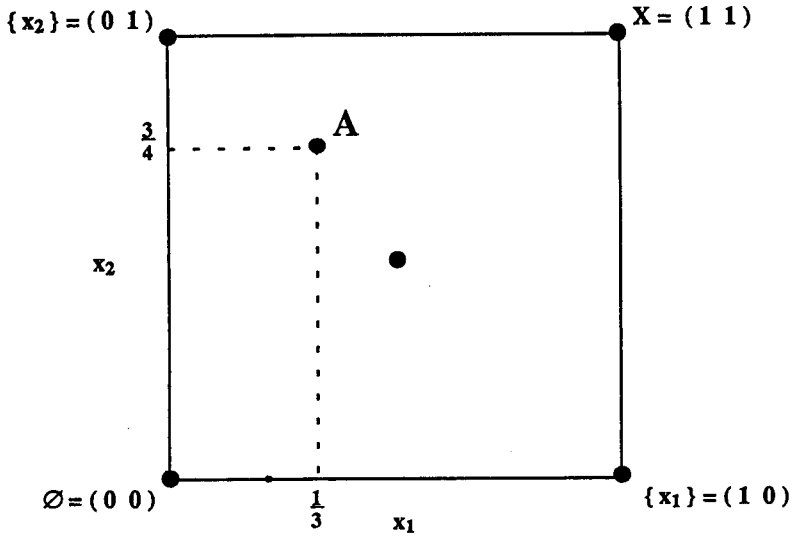

Figure 2 Sets as points. The fuzzy subset A is a point in the unit 2-cube with coordinates or fit values $(\frac{1}{3}\ \frac{3}{4})$. The first element x_1 fits in or belongs to A to degree $\frac{1}{3}$, the element x_2 to degree $\frac{3}{4}$. The cube consists of all possible fuzzy subsets of two elements $\{x_1, x_2\}$. The four corners represent the power set 2^X of $\{x_1, x_2\}$.

of the 2^n vertices of the cube. The nearest corners are also the farthest. This metrical relationship is evident in Figure 2.

Fuzzy sets are combined[16] pairwise with minimum, maximum, and order reversal, as are nonfuzzy sets. Fuzzy set intersection is defined fitwise by pairwise minimum (picking the smaller of the two elements), union by pairwise maximum, and complementation by order reversal. For example:

$$A = (1\ 0.8\ 0.4\ 0.5)$$

$$B = (0.9\ 0.4\ 0\ 0.7)$$

$$A \cap B = (0.9\ 0.4\ 0\ 0.5)$$

$$A \cup B = (1\ 0.8\ 0.4\ 0.7)$$

$$A^c = (0\ 0.2\ 0.6\ 0.5)$$

$$A \cap A^c = (0\ 0.2\ 0.4\ 0.5)$$

$$A \cup A^c = (1\ 0.8\ 0.6\ 0.5).$$

Note that the overlap fit vector $A \cap A^c$ is not the vector of all zeroes and the underlap fit vector $A \cup A^c$ is not the vector of all ones. This is true of all properly fuzzy sets, all points in I^n other than vertex points. Indeed the min-max definitions

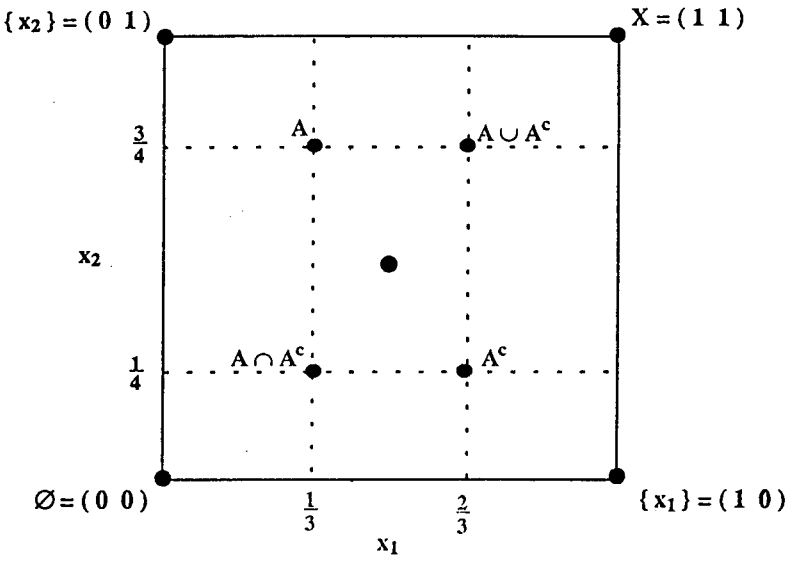

Figure 3 Completing the fuzzy square. The fuzzier A is, the closer A is to the midpoint of the fuzzy cube. As A approaches the midpoint, all four points—A, A^c, $A \cap A^c$, and $A \cup A^c$—contract to the midpoint. The less fuzzy A is, the closer A is to the nearest vertex. As A approaches the vertex, all four points spread out to the four vertices and the bivalent power set 2^X is recovered.

give at once the following fundamental characterization of fuzziness as non-degenerate overlap and nonexhaustive underlap.

PROPOSITION *A is properly fuzzy if and only if $A \cap A^c \neq \emptyset$ and if and only if $A \cup A^c \neq X$.*

An illustration of this fundamental proposition is what we might call *completing the fuzzy square.* Consider again the two-dimensional fuzzy set A defined by the fit vector $(\frac{1}{3}\ \frac{3}{4})$. The corresponding overlap and underlap sets can be found by first finding the complement set A^c and then combining the fit vectors pairwise with minimum and with maximum:

$$A = (\tfrac{1}{3}\ \tfrac{3}{4})$$

$$A^c = (\tfrac{2}{3}\ \tfrac{1}{4})$$

$$A \cap A^c = (\tfrac{1}{3}\ \tfrac{1}{4})$$

$$A \cup A^c = (\tfrac{2}{3}\ \tfrac{3}{4}).$$

The sets-as-points view shows that these four points in the unit square hang together, indeed move together, in a very natural way. Consider the geometry of Figure 3.

In Figure 3 the four fuzzy sets involved in the fuzziness of set A—the sets A, A^c, $A \cap A^c$, and $A \cup A^c$—contract to the midpoint as A becomes maximally fuzzy and

expand out to the Boolean corners of the cube as A becomes minimally fuzzy. The same contraction and expansion occurs in n dimensions for the 2^n fuzzy sets defined by all combinations of $m_A(x_1)$ and $m_{A^c}(x_1), \ldots, m_A(x_n)$ and $m_{A^c}(x_n)$.

At the midpoint nothing is distinguishable. At the vertices everything is distinguishable. These extremes represent the two ends of the spectrum of logic and set theory. In this sense the midpoint is the black hole of set theory.

5. PARADOX AT THE MIDPOINT

The midpoint is full of paradox. It is forbidden to classical logic and set theory. Where midpoint phenomena appear in Western thought they are invariably labeled "paradoxes" or denied altogether. Midpoint phenomena include the half-empty and half-full cup, the Taoist Yin-Yang, the liar from Crete who said that all Cretans are liars, Bertrand Russell's set of all sets that are not members of themselves, and Russell's barber.

Russell's barber is a bewhiskered man who lives in a town and shaves a man if and only if he does not shave himself. So who shaves the barber? If he shaves himself, then by definition he does not. But if he does not shave himself, then by definition he does. So he does and he does not—contradiction ("paradox"). Gaines[4] observed that this paradoxical circumstance can be numerically interpreted as follows.

Let S be the proposition that the barber shaves himself and not-S that he does not. Then since S implies not-S and not-S implies S, the two propositions are logically equivalent: $S = \text{not-}S$. Equivalent propositions have the same truth values:

$$t(S) = t(\text{not-}S)$$

$$= 1 - t(S).$$

Solving for $t(S)$ gives the midpoint point of the truth interval (the one-dimensional cube $[0, 1]$): $t(S) = \frac{1}{2}$. The midpoint is equidistant to the vertices 0 and 1. In the bivalent (two-valued) case, roundoff is impossible and paradox occurs.

In bivalent logic both statements S and not-S must have truth value zero or unity. The fuzzy resolution of the paradox only uses the fact that the truth values are equal. It does not in principle constrain their range. The midpoint value $\frac{1}{2}$ emerges from the structure of the problem and the order-reversing effect of negation.

The paradoxes of classical set theory and logic are part of the price one pays for an arbitrary insistence on bivalence. This insistence is often made in the name of science. In the end, though, if is simply a cultural preference, a reflection of an educational predilection that goes back at least to Aristotle. It takes great faith to insist on bivalence in the face of both bivalent contradictions (paradoxes) and a consistent fuzzy alternative.

Put another way, fuzziness shows that there are limits to logical certainty. We can no longer assert the laws of noncontradiction and excluded middle *for sure*—and *for free*.

The fuzzy theorist must explain why so many people have been wrong for so long. We now have the machinery to offer an explanation. The reason is that *rounding off*, quantizing, simplifies life and often costs little. We agree to call empty

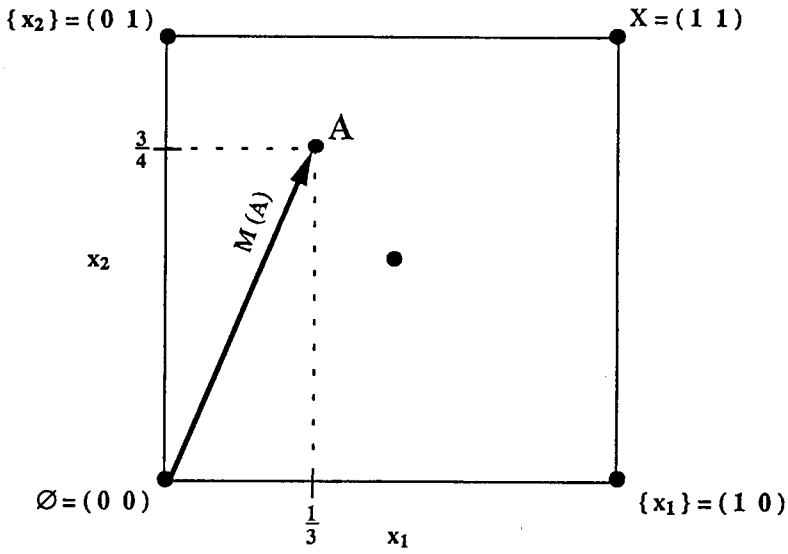

Figure 4 The count $M(A)$ of A is the fuzzy Hamming norm (l^1 norm) of the vector drawn from the origin to A.

the near empty cup, and present the large pulse and absent the small pulse. We round off points inside the fuzzy cube to the nearest vertex. This roundoff heuristic works fine as a first approximation to describing the universe until we get near the midpoint of the cube. These phenomena are harder to roundoff. In the logically extreme case, at the midpoint of the cube, the procedure breaks down completely because every vertex is equally close. Hands are thrown up and paradox declared.

Faced with midpoint phenomena, the fuzzy skeptic is in the same position as the flat-earther, who denies that the earth's surface is curved, when she stands at the north pole, looks at her compass and wants to go south.

6. COUNTING WITH FUZZY SETS

How big is a fuzzy set? The size or cardinality of A, $M(A)$, is the sum of the fit values of A:

$$M(A) = \sum_{i=1}^{n} m_A(x_i).$$

The count of $A = (\frac{1}{3}\ \frac{3}{4})$ is $M(A) = \frac{1}{3} + \frac{3}{4} = \frac{13}{12}$. The cardinality measure M is sometimes called the *sigma-count*.[17] The measure M generalizes[9] the classical counting measure of combinatorics and measure theory. (So (X, I^n, M) is the fundamental measure space of fuzzy theory.) In general the measure M does not give integer values.

The measure M has a natural geometric interpretation in the sets-as-points framework. It is the magnitude of the vector drawn from the origin to the fuzzy set, as illustrated in Figure 4.

Consider the l^p distance between fuzzy sets A and B in I^n:

$$l^p(A, B) = \sqrt[p]{\sum_{i=1}^{n} |m_A(x_i) - m_B(x_i)|^p},$$

where $1 \leqq p \leqq \infty$. The l^2 distance is the physical Euclidean distance actually illustrated in the figures. The simplest distance is the l^1 or *fuzzy Hamming distance*, the sum of the absolute fit differences. We shall use fuzzy Hamming distance throughout, though all results admit a general l^p formulation. Using the fuzzy Hamming distance the count M can be rewritten as the desired l^1 norm:

$$\begin{aligned} M(A) &= \sum_{i=1}^{n} m_A(x_i) \\ &= \sum_i |m_A(x_i) - 0| \\ &= \sum_i |m_A(x_i) - m_{\varnothing}(x_i)| \\ &= l^1(A, \varnothing). \end{aligned}$$

7. THE FUZZY ENTROPY THEOREM

How fuzzy is a fuzzy set? Fuzziness is measured by a *fuzzy entropy* measure. Entropy is a generic notion. It need not be probabilistic. Entropy measures the uncertainty of a system or message. A fuzzy set is a type of system or message. Its uncertainty is its fuzziness.

The fuzzy entropy of A, $E(A)$, varies from 0 to 1 on the unit hypercube I^n. Only the cube vertices have zero entropy, since nonfuzzy sets are unambiguous. The cube midpoint uniquely has maximum entropy one. Fuzzy entropy smoothly increases as one moves from any vertex to the midpoint. The algebraic requirements for fuzzy entropy measures can be found in Klir.[8]

Simple geometric considerations lead[10] to a ratio form for the fuzzy entropy. The closer the fuzzy set A is to the nearest vertex A_{near}, the farther A is from the farthest vertex A_{far}. Opposite the long diagonal from the nearest vertex is the farthest vertex. Let a denote the distance $l^1(A, A_{near})$ to the nearest vertex and let b denote the distance $l^1(A, A_{far})$ to the farthest vertex. Then the fuzzy entropy is simply the ratio of a to b:

$$E(A) = \frac{a}{b} = \frac{l^1(A, A_{near})}{l^1(A, A_{far})}.$$

The sets-as-points interpretation of the fuzzy entropy is shown in Figure 5, where $A = (\frac{1}{3}\ \frac{3}{4})$, $A_{near} = (0\ 1)$, and $A_{far} = (1\ 0)$. So $a = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$ and $b = \frac{2}{3} + \frac{3}{4} = \frac{17}{12}$. So $E(A) = \frac{7}{17}$.

Alternatively, those reading this in a room can imagine that the room is the unit

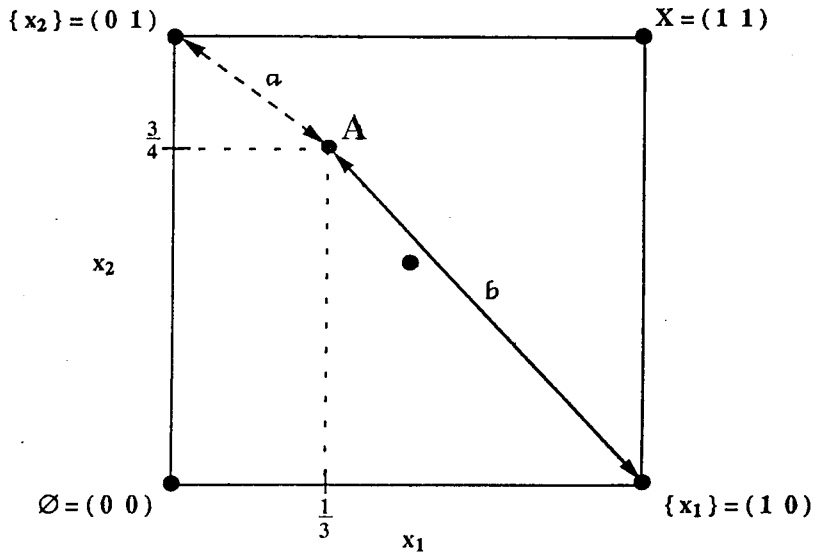

Figure 5 Fuzzy entropy $E(A)=a/b$ as balance between distance to nearest vertex and distance to farthest vertex.

cube I^3 and that their head is a fuzzy set in it. Once the nearest corner of the room is located, the farthest corner is opposite the long diagonal emanating from the nearest corner. If your head is in a corner, $a=0$ and $E(A)=0$. If your head is in the metrical center of the room, every corner is nearest and farthest. So $a=b$ and $E(A)=1$.

Since overlap and underlap characterize fuzziness we can expect them to be involved in the measure of fuzziness. Careful examination of the completed fuzzy square in Figure 3 shows that this is the case. For, by symmetry, each of the four points A, A^c, $A\cap A^c$, and $A\cup A^c$ is equally close to its nearest vertex. The common distance is a. Similarly, each point is equally far from its farthest vertex. The common distance is b. One of the first four distances is the count $M(A\cap A^c)$. One of the second four distances is the count $M(A\cup A^c)$. This gives a geometric proof of the Fuzzy Entropy Theorem,[10,11] which states that fuzziness consists of a balance of counted violations of the law of noncontradiction and counted violations of the law of excluded middle.

FUZZY ENTROPY THEOREM

$$E(A)=\frac{M(A\cap A^c)}{M(A\cup A^c)}.$$

An algebraic proof is straightforward. The geometric proof can be seen by examining the completed fuzzy square in Figure 6.

The Fuzzy Entropy Theorem explains why fuzziness begins where Western logic ends. When the laws of noncontradiction and excluded middle are obeyed, overlap is empty and underlap is exhaustive. So $M(A\cap A^c)=0$ and $M(A\cup A^c)=n$, and thus $E(A)=0$.

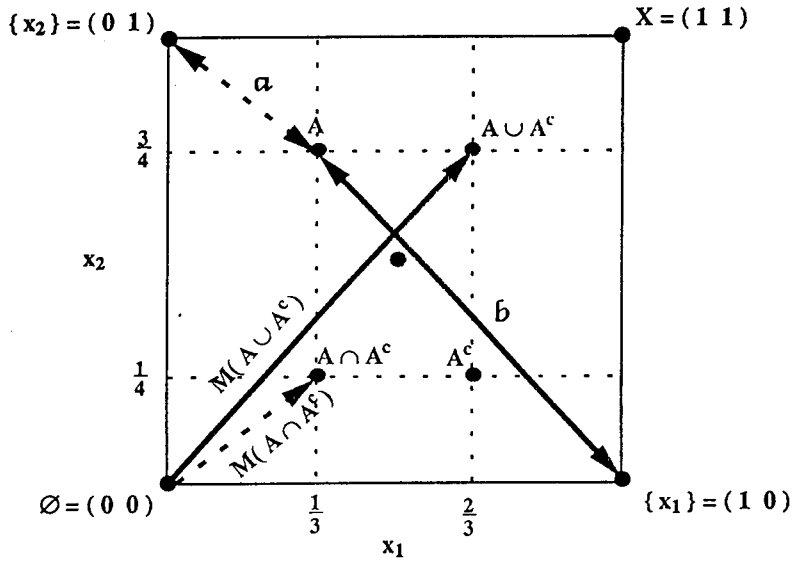

Figure 6 Geometry of the Fuzzy Entropy Theorem. By symmetry each of the four points on the completed fuzzy square is equally close to its nearest vertex and equally far from its farthest vertex.

The Fuzzy Entropy Theorem also provides a first-principles derivation of the basic fuzzy set operations of minimum (intersection), maximum (union), and order reversal (complementation) proposed in 1965 by Zadeh[16] at the inception of fuzzy theory. (Lukasiewicz first proposed these operations for continuous or fuzzy logics in the 1920s.)

For the fuzzy theorist, this result also shows that triangular norms or T-norms,[8] which generalize conjunction or intersection, and the dual triangular co-norms C, which generalize disjunction or union, do not have the first-principles status of min and max. For, the triangular norm inequalities,

$$T(x, y) \leqq \min(x, y) \leqq \max(x, y) \leqq C(x, y),$$

show that replacing min with any T in the numerator term $M(A \cap A^c)$ can only make the numerator smaller. Replacing max with any C in the term $M(A \cup A^c)$ can only make the denominator larger. So any T or C not identically min or max makes the ratio smaller, strictly smaller if A is fuzzy. Then the entropy theorem does not hold and the resulting pseudo-entropy measure does not equal unity at the midpoint, though it continues to be maximized there. This can be easily seen with the product T-norm[14] $T(x, y) = xy$ and its DeMorgan dual co-norm $C(x, y) = 1 - T(1-x, 1-y) = x + y - xy$, or with the bounded sum T-norm $T(x, y) = \max(0, x + y - 1)$ and DeMorgan dual $C(x, y) = \min(1, x + y)$. The Entropy Theorem similarly fails in general if the negation or complementation operator $N(x) = 1 - x$ is replaced by a parameterized operator $N_a(x) = (1-x)/(1+ax)$ for nonzero $a > -1$.

As an aside, note that all probability distributions, all sets A with $M(A) = 1$, in I^n form a $n-1$ dimensional simplex S^n. In the unit square the probability simplex is

the negatively sloped diagonal line. In the unit 3-cube it is a solid triangle. In the unit 4-cube it is a tetrahedron, and so on up.

If no probabilistic fit value p_i is such that $p_i > \frac{1}{2}$, then the Fuzzy Entropy Theorem implies[11] that the distribution P has fuzzy entropy $E(P) = 1/(n-1)$. Else $E(P) < 1/(n-1)$. So the probability simplex S^n is entropically degenerate for large dimensions n. This result also shows that the uniform distribution $(1/n, \ldots, 1/n)$ maximizes fuzzy entropy on S^n but not uniquely. This in turn shows that fuzzy entropy differs from the average-information measure of probabilistic entropy, which is uniquely maximized by the uniform distribution.

The Fuzzy Entropy Theorem also implies that, analogous to $\log 1/p$, a unit of fuzzy information is $f/(1-f)$ or $(1-f)/f$, depending on whether the fit value f obeys $f \leqq \frac{1}{2}$ or $f \geqq \frac{1}{2}$.

The event x can be ambiguous or clear. It is ambiguous if f is approximately $\frac{1}{2}$ and clear if f is approximately 1 or 0. If an ambiguous event occurs, is observed, is disambiguated, etc., then it is maximally informative: $E(f) = E(\frac{1}{2}) = 1$. If a clear event occurs, is observed, etc., it is minimally informative: $E(f) = E(0) = E(1) = 0$. This is in accord with the information interpretation of the probabilistic entropy measure $\log 1/p$, where the occurrence of a sure event ($p = 1$) is minimally informative (zero entropy) and the occurrence of an impossible event ($p = 0$) is maximally informative (infinite entropy).

8. THE SUBSETHOOD THEOREM

Sets contain subsets. A is a *subset* of B, denoted $A \subset B$, if and only if every element of A is an element of B. The power set 2^B contains all of B's subsets. So, alternatively,[1] A is a subset of B just in case A belongs to B's power set:

$$A \subset B \quad \text{if and only if } A \in 2^B.$$

The subset relation corresponds to the implication relation in logic. In classical logic *truth* is a mapping from the set of statements $\{S\}$ to truth values: $t: \{S\} \to \{0, 1\}$. Consider the truth-tabular definition of implication for bivalent propositions P and Q:

P	Q	$P \to Q$
0	0	1
0	1	1
1	0	0
1	1	1

The implication is false if and only if the antecedent P is true and the consequent Q is false—when "truth implies falsehood".

The same holds for subsets. Representing sets as bivalent functions $m_A: X \to \{0, 1\}$, A is a subset of B if there is no element x that belongs to A but not to B, or $m_A(x) = 1$ but $m_B(x) = 0$. This membership-function definition can be rewritten as follows:

$$A \subset B \quad \text{if and only if } m_A(x) \leqq m_B(x) \quad \text{for all } x.$$

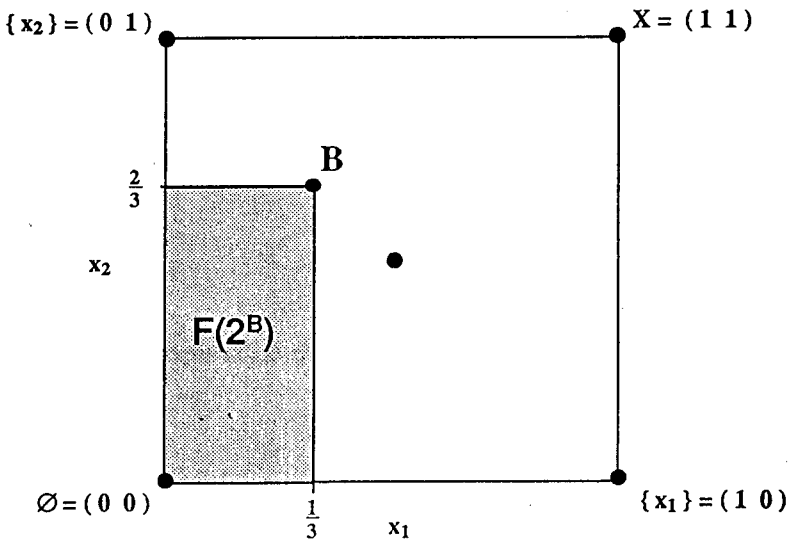

Figure 7 Fuzzy power set $F(2^B)$ as a hyper-rectangle in the fuzzy cube. Side lengths are the fit values $m_B(x_i)$. The size or volume of $F(2^B)$ is the product of the fit values.

Zadeh[16] proposed the same relation for defining when fuzzy set A is a subset of fuzzy set B. We refer to this as the *dominated membership function relationship*. If $A=(0.3\,0\,0.7)$ and $B=(0.4\,0.7\,0.9)$, then A is a fuzzy subset of B but B is not a fuzzy subset of A. A candidate fuzzy set A either is or is not a fuzzy subset of B. This is the problem. The relation of fuzzy subsethood is *not* fuzzy. It is either black or white.

The sets-as-points view asks a geometric question: What do all fuzzy subsets of B look like? What does the *fuzzy power set* of B—$F(2^B)$, the set of all fuzzy subsets of B—look like? The dominated membership function relationship implies that $F(2^B)$ is the hyper-rectangle emanating from the origin with side lengths given by the fit values $m_A(x_i)$. Figure 7 displays the fuzzy power set of the set $B=(\frac{1}{3}\,\frac{2}{3})$. Of course the count of $F(2^B)$ is infinite if B is not empty. For finite-dimensional sets, the size of $F(2^B)$ can be taken[11] as the Lebesgue measure or volume $V(B)$, the product of the fit values:

$$V(B)=\prod_{i=1}^{n} m_B(x_i).$$

Figure 7 illustrates that $F(2^B)$ is not a fuzzy set. A cube point A either is or is not in the hyper-rectangle $F(2^B)$. Some points A outside the hyper-rectangle $F(2^B)$ resemble subsets of B more than other points do. The black–white definition of subsethood ignores this.

The natural generalization is to define fuzzy subsets on $F(2^B)$. Some sets A belong in $F(2^B)$ to different degrees. The abstract membership function $m_{F(2^B)}(A)$ can be any number in $[0, 1]$. Degrees of subsethood are possible.

Let $S(A, B)$ denote the degree to which A is a subset of B:

$$S(A, B) = \text{Degree}(A \subset B)$$

$$= m_{F(2^B)}(A).$$

$S(\cdot, \cdot)$ is the *subsethood measure*. $S(\cdot, \cdot)$ takes values in $[0, 1]$. We will see that it is the fundamental, unifying structure in fuzzy theory.

The current task is to measure $S(A, B)$. We will first present an earlier[10,11] algebraic derivation of the subsethood measure $S(A, B)$. We will then present a new, more fundamental, geometric derivation.

The algebraic derivation is the *fit-violation strategy*.[10] The idea is that you study a law by breaking it. Consider the dominated membership function relationship: $A \subset B$ if and only if $m_A(x) \leqq m_B(x)$ for all x in X.

Suppose element x_v violates the dominated membership function relationship: $m_A(x_v) > m_B(x_v)$. Then A is not a subset of B, at least not totally. Suppose further that the dominated membership inequality holds for all other elements x. Only element x_v violates the relationship. For instance, X may consist of one hundred values: $X = \{x_1, \ldots, x_{100}\}$. The violation might occur, say, with the first element: $x_1 = x_v$. Then intuitively A is largely a subset of B. Suppose now that X contains a thousand elements, or a trillion elements, and only the first element violates the dominated membership function relationship. Then surely A is overwhelmingly a subset of B; perhaps $S(A, B) = 0.999999999999$.

This argument suggests we should count fit violations in magnitude and frequency. The greater the violations in magnitude, $m_A(x_v) - m_B(x_v)$, and the greater the number of violations relative to the size $M(A)$ of A, the less A is a subset of B; equivalently, the more A is a *superset* of B. For, both intuitively and by the dominated-membership definition, supersethood and subsethood are inversely related:

$$SUPERSETHOOD(A, B) = 1 - S(A, B).$$

The simplest way to count violations is to add them. If we sum over all x, the summand should equal $m_A(x_v) - m_B(x_v)$ when this difference is positive, zero when it is nonpositive. So the summand is $\max(0, m_A(x) - m_B(x))$. The unnormalized count is therefore the sum of these maxima:

$$\sum_{x \in X} \max(0, m_A(x) - m_B(x)).$$

The simplest, and most appropriate, normalization factor is the count of A, $M(A)$. We can assume $M(A) > 0$ since $M(A) = 0$ if and only if A is empty. The empty set trivially satisfies the dominated membership function relationship. So it is a subset of every set. Normalization gives the minimal measure of nonsubsethood, of supersethood:

$$SUPERSETHOOD(A, B) = \frac{\sum_x \max(0, m_A(x) - m_B(x))}{M(A)}.$$

Then subsethood is the negation of this ratio. This gives the minimal fit-violation measure of subsethood:

$$S(A,B)=1-\frac{\sum_x \max(0, m_A(x)-m_B(x))}{M(A)}.$$

The subsethood measure may appear ungraceful at first but it behaves as it should. Observe that $S(A,B)=1$ if and only if the dominated membership function relationship holds. For if it holds, zero violations are summed. Then $S(A,B)=1-0=1$. If $S(A,B)=1$, every numerator summand is zero. So no violation occurs. At the other exteme, $S(A,B)=0$ if and only if B is the empty set. So the empty set is the unique set without subsets, fuzzy or nonfuzzy. Degrees of subsethood occur between these extremes.

The subsethood measure also relates to logical implication. Viewed at the 1-dimensional level of fuzzy logic, and so ignoring the normalizing count ($M(A)=1$), the subsethood measure reduces to the Lukasiewicz implication operator:

$$\begin{aligned} S(A,B) &= 1-\max(0, m_A-m_B) \\ &= 1-[1-\min(1-0, 1-(m_A-m_B))] \\ &= \min(1, 1-m_A+m_B) \\ &= t_L(A\rightarrow B), \end{aligned}$$

which clearly generalizes the truth-tabular definition of bivalent implication.

Consider the fit vectors $A=(0.2\ 0\ 0.4\ 0.5)$ and $B=(0.7\ 0.6\ 0.3\ 0.7)$. Neither set is a proper subset of the other. A is almost a subset of B but not quite since $m_A(x_3)-m_B(x_3)=0.4-0.3=0.1>0$. Hence $S(A,B)=1-\frac{0.1}{1.1}=\frac{10}{11}$. Similarly $S(B,A)=1-\frac{1.3}{2.3}=\frac{10}{23}$.

The concept of subsethood applies to nonfuzzy sets. Consider the sets

$$C=\{x_1, x_2, x_3, x_5, x_7, x_9, x_{10}, x_{12}, x_{14}\}$$

and

$$D=\{x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{12}, x_{13}, x_{14}\}$$

with corresponding bit vectors

$$C=(1\ 1\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 1\ 0\ 1\ 0\ 1)$$

$$D=(0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 1\ 1\ 1)$$

C and D are not subsets of each other. But C should very nearly be a subset of D since only x_1 violates the dominated membership function relationship. We find $S(C,D)=1-\frac{1}{9}=\frac{8}{9}$ while $S(D,C)=1-\frac{4}{12}=\frac{2}{3}$. So D is more a subset of C than it is not. This is because the two sets are largely equivalent. They have much overlap: $M(C\cap D)=8$. This observation motivates the Fuzzy Subsethood Theorem presented below. First, though, we present a new geometric derivation of the subsethood measure.

Consider the sets-as-points geometry of subsethood in Figure 7. Set A is either

in the hyper-rectangle $F(2^B)$ or not. Intuitively the subsethood of A in B should be nearly unity when A is arbitrarily close to the fuzzy power set $F(2^B)$. The farther away, the less the subsethood $S(A,B)$ or, equivalently, the greater the supersethood.

So the key idea is metrical: How close is A to $F(2^B)$? Let $d(A,F(2^B))$ denote this l^p distance. There is a distance $d(A,B')$ between A and every point B' in the hyper-rectangle, every subset B' of B. The distance $d(A,F(2^B))$ is the smallest such distance. Since the hyper-rectangle $F(2^B)$ is geometrically well behaved—$F(2^B)$ is closed and bounded (compact) and convex—some subset B^* of B achieves this minimum distance. So the infimum, the greatest lower bound, is the distance $d(A,B^*)$:

$$d(A,F(2^B)) = \inf\{d(A,B'): B' \in F(2^B)\}$$

$$= d(A,B^*).$$

The closest set B^* is easy to locate geometrically. In the Euclidean or ℓ^2 case, this is formally due to the geometric Hahn–Banach Theorem since $F(2^B)$ is convex. If A is a subset of B— if A is in the hyper-rectangle $F(2^B)$—then A itself is the closest subset: $A = B^*$. So suppose A is not a proper subset of B.

The unit cube I^n can be sliced into 2^n hyper-rectangles by extending the sides of $F(2^B)$ to hyperplanes. The hyperplanes intersect perpendicularly (orthogonally), at least in the Euclidean case. $F(2^B)$ is one of the hyper-rectangles. The hyper-rectangle interiors correspond to the 2^n cases whether $m_A(x_i) < m_B(x_i)$ or $m_A(x_i) > m_B(x_i)$ for fixed B and arbitrary A. The edges are the loci of points when some $m_A(x_i) = m_B(x_i)$.

The 2^n hyper-rectangles can be classified as *mixed* or *pure* membership domination. In the pure case either $m_A < m_B$ or $m_A > m_B$ holds in the hyper-rectangle interior for all x and all interior points A. In the mixed case $m_A(x_i) < m_B(x_i)$ holds for some of the coordinates x_i and $m_A(x_j) > m_B(x_j)$ holds for the remaining coordinates x_j in the interior for all interior A. So there are only two pure membership-domination hyper-rectangles, the set of proper subsets $F(2^B)$ and the set of proper supersets.

Figure 8 illustrates how the fuzzy power set $F(2^B)$ of $B = (\frac{1}{3}\ \frac{2}{3})$ can be linearly extended to partition the unit square into $2^2 = 4$ rectangles. The non-subsets A_1, A_2 and A_3 reside in distinct quadrants. The northwest and southeast quadrants are the mixed membership-domination rectangles. The southwest and the northeast quadrants are the pure rectangles.

The nearest set B^* to A in the pure superset hyper-rectangle is B itself. The nearest set B^* in the mixed case is found by drawing a perpendicular (orthogonal) line segment from A to $F(2^B)$. Convexity of $F(2^B)$ is responsible. In Figure 8 the perpendicular lines from A_1 and A_3 intersect line edges (1-dimensional linear subspaces) of the rectangle $F(2^B)$. The line from A_2 to B, the corner of $F(2^B)$, is degenerately perpendicular since B is a zero-dimensional linear subspace.

These "orthogonality" conditions are more pronounced in three dimensions. Let your room again be the unit 3-cube. Consider a large dictionary fit snugly against the floor corner corresponding to the origin. Point B is the dictionary corner farthest from the origin. Extending the three exposed faces of the dictionary partitions the room into 8 octants, one of which is occupied by the dictionary.

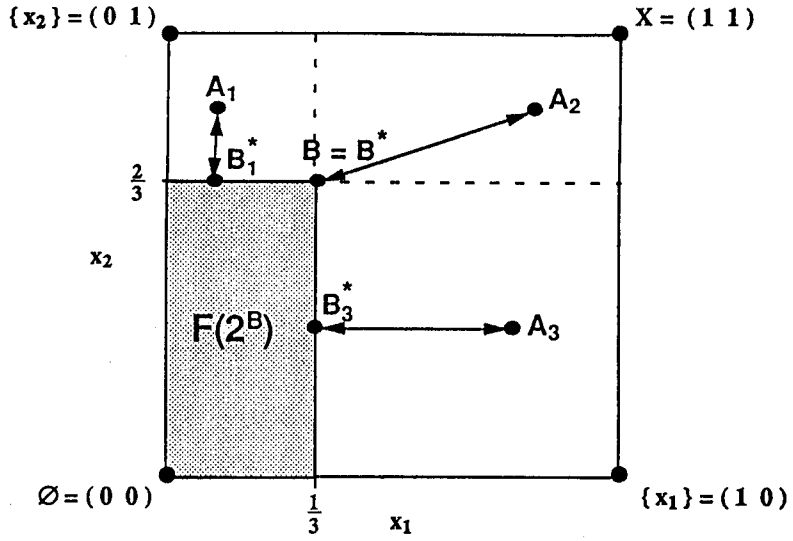

Figure 8 Partition of hypercube I^n into 2^n hyper-rectangles by linearly extending $F(2^B)$. The nearest points B_1^* and B_3^* to points A_1 and A_3 in the northwest and southeast quadrants are found by the normals from $F(2^B)$ to A_1 and A_3. The nearest point B^* to point A_2 in the northeast quadrant is itself. This "orthogonal" optimality condition allows $d(A, B)$ to be given by the general Pythagorean Theorem as the hypotenuse in an ℓ^p "right" triangle.

Points in the other 7 octants are connected to the nearest points on the dictionary by lines that perpendicularly intersect one of the three exposed faces, one of the three exposed edges, or the corner B.

The "orthogonality" condition invokes the ℓ^p-version of the Pythagorean Theorem. For our ℓ^1 purposes:

$$d(A, B) = d(A, B^*) + d(B, B^*).$$

The more familiar l^2-version, actually pictured in Figure 8, requires squaring these distances. For the general ℓ^p case:

$$\|A - B\|^p = \|A - B^*\|^p + \|B^* - B\|^p,$$

or equivalently,

$$\sum_{i=1}^{n} |a_i - b_i|^p = \sum_{i=1}^{n} |a_i - b_i^*|^p + \sum_{i=1}^{n} |b_i^* - b_i|^p.$$

Equality holds for all $p \geqq 1$ since, as is clear from Figure 8 and in general, from the algebraic argument below, either $b_i^* = a_i$ or $b_i^* = b_i$.

This Pythagorean equality is surprising. We have come to think of the Pythagorean Theorem (and orthogonality) as an ℓ^2 or Hilbert space property. Yet here it holds in *every* ℓ^p space—if B^* is the set in $F(2^B)$ closest to A in ℓ^p distance. Of course for other sets strict inequality holds in general if $p \neq 2$. This suggests a

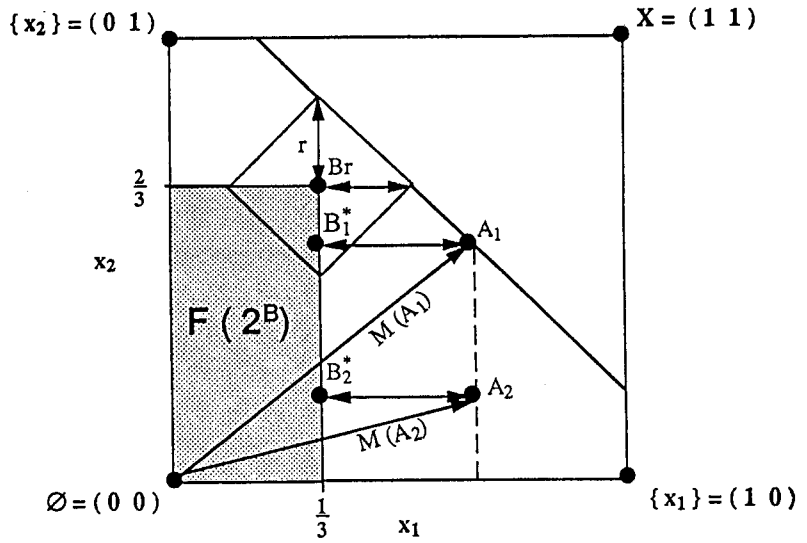

Figure 9 Dependence of subsethood on the count $M(A)$. A_1 and A_2 are equidistant to $F(2^B)$ but A_1 is closer to B than A_2 is; correspondingly, $M(A_1) > M(A_2)$. Loci of points A of constant count $M(A)$ are line segments parallel to the negatively sloping long diagonal. l^1 spheres centered at B are diamond shaped.

special status for the closest set B^*. We shall see below that the Subsethood Theorem confirms this suggestion. We shall use the term "orthogonality" loosely to refer to this ℓ^p Pythagorean relationship, while remembering its customary restriction to ℓ^2 spaces and inner products.

The natural suggestion is to define supersethood as the distance $d(A, F(2^B)) = d(A, B^*)$. Supersethood increases with this distance, subsethood decreases with it. To keep supersethood, and thus subsethood, unit-interval valued, the distance must be suitably normalized.

The simplest way to normalize $d(A, B^*)$ is with a constant: the maximum unit-cube distance, $n^{1/p}$ in the general ℓ^p case and n in our case. This gives the candidate subsethood measure

$$S(A, B) = 1 - \frac{d(A, B^*)}{n}.$$

This candidate subsethood measure fails in the boundary case when B is the empty set. For then $d(A, B^*) = d(A, B) = M(A)$. So the measure gives $S(A, \varnothing) = 1 - (M(A)/n) > 0$. Equality holds exactly when $A = X$. But the empty set has no subsets. The only normalization factor that ensures this is the count $M(A)$. Of course $M(A) = n$ when $A = X$.

Normalizing by n also treats all equidistant points the same. Consider points A_1 and A_2 in Figure 9. Both points are equidistant to their nearest $F(2^B)$ point: $d(A_1, B_1^*) = d(A_2, B_2^*)$. But A_1 is closer to B than A_2. In particular A_1 is closer to the horizontal line defined by the fit value $m_B(x_2) = \frac{2}{3}$. The variable quantity that

reflects this is the count $M(A)$: $M(A_1) > M(A_2)$. The count gap $M(A_1) - M(A_2)$ is due to the fit gap involving x_1, and reflects $d(A_1, B) < d(A_2, B)$. In general the count $M(A)$ relates to this distance, as can be seen by checking extreme cases of closeness of A to B (and drawing some diamond-shaped l^1 spheres centered at B). Indeed if $m_A > m_B$ everywhere, $d(A, B) = M(A) - M(B)$.

Since $F(2^B)$ fits snugly against the origin, the count $M(A)$ in any of the other $2^n - 1$ hyper-rectangles can only be larger than the count $M(B^*)$ of the nearest $F(2^B)$ points. The normalization choice of n leaves the candidate subsethood measure indifferent to which of the $2^n - 1$ hyper-rectangles A is in and to where A is in the hyper-rectangle. Each point in each hyper-rectangle involves a different combination of fit violations and satisfactions. The normalization choice of $M(A)$ reflects this fit violation structure as well as behaves appropriately in boundary cases.

The normalization choice $M(A)$ leads to the subsethood measure

$$S(A, B) = 1 - \frac{d(A, B^*)}{M(A)}.$$

We now show that this measure is equal to the subsethood measure derived algebraically above.

Let B' be any subset of B. Then by definition the nearest subset B^* obeys the inequality:

$$\sqrt[p]{\sum_{i=1}^{n} |a_i - b_i^*|^p} \leqq \sqrt[p]{\sum_{i=1}^{n} |a_i - b_i'|^p},$$

where for convenience $a_i = m_A(x_i)$ and similarly for the b_i fit values. We will assume $p = 1$ but the following characterization of b_i^* is valid for any $p > 1$.

By orthogonality we know that a_i is at least as big as b_i^*. So first suppose $a_i = b_i^*$. This occurs if and only if no violation occurs: $a_i \leqq b_i$. (If this holds for all i, then $A = B^*$.) So $\max(0, a_i - b_i) = 0$. Next suppose $a_i > b_i^*$. This occurs if and only if a violation occurs: $a_i > b_i$. (If this holds for all i, then $B = B^*$.) So $b_i^* = b_i$ since B^* is the subset of B nearest to A. Equivalently, $a_i > b_i$ holds if and only if $\max(0, a_i - b_i) = a_i - b_i$. So the two cases together prove that $\max(0, a_i - b_i) = |a_i - b_i^*|$. Summing over all x_i gives:

$$d(A, B^*) = \sum_{i=1}^{n} \max(0, m_A(x_i) - m_B(x_i)).$$

So the two subsethood measures are equivalent.

This proof also proves a deeper characterization of the optimal subset B^*: $B^* = A \cap B$. For if a violation occurs, $a_i > b_i$ and $b_i = b_i^*$. So $\min(a_i, b_i) = b_i^*$. Otherwise $a_i = b_i^*$, and so $\min(a_i, b_i) = b_i^*$.

This in turn proves that B^* is a point of double optimality. Not only is B^* the subset of B nearest A, B^* is also A^*, the subset of A nearest to B: $d(B, F(2^A)) = d(B, A^*) = d(B, B^*)$. Figure 10 illustrates that $B^* = A \cap B = A^*$ is the set within both

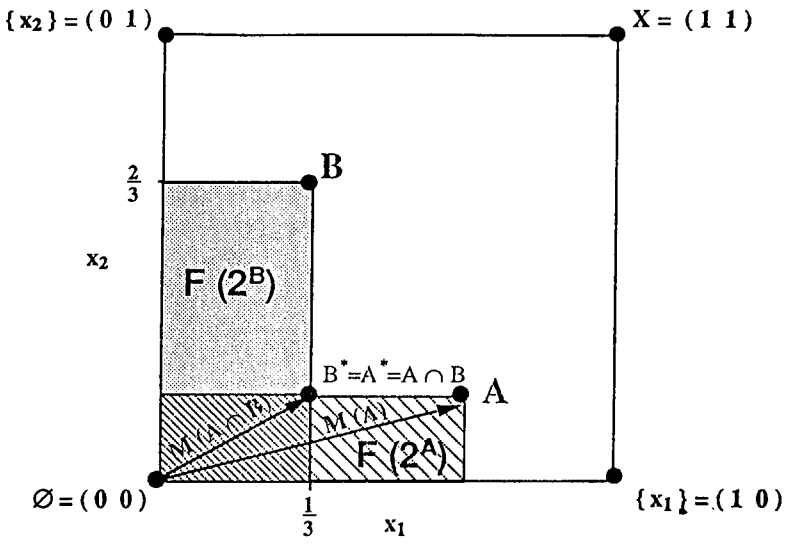

Figure 10 B^* as both the subset of B nearest A and the subset A^* of A nearest B: $B^* = A^* = A \cap B$. The distance $d(A, B^*) = M(A) - M(A \cap B)$ illustrates the Subsethood Theorem.

the hyper-rectangle $F(2^A)$ and the hyper-rectangle $F(2^B)$ that has maximal count $M(A \cap B)$.

Figure 10 also shows that the distance $d(A, B^*)$ is a vector magnitude difference: $d(A, B^*) = M(A) - M(A \cap B)$. Dividing both sides of this equality by $M(A)$ and rearranging proves a surprising and still deeper structural characterization of subsethood, the Subsethood Theorem.

SUBSETHOOD THEOREM

$$S(A, B) = \frac{M(A \cap B)}{M(A)}.$$

The ratio form of the subsethood measure $S(A, B)$ is familiar. It is the same as the ratio form of the conditional probability $P(B|A)$. The fundamental difference is that the ratio form is *derived* for the subsethood measure $S(A, B)$ but *assumed* for the conditional probability $P(B|A)$. This is the difference between showing and telling. The inability to derive conditional probability further suggests that probability is not real. For every probability is a conditional probability, $P(A) = P(A|X)$.

Consider first the physical interpretation of randomness as a relative frequency. The Subsethood Theorem suggests that randomness is a working fiction akin to the luminiferous ether of nineteenth-century physics—the phlogiston of thought. For in one stroke we can now derive the relative frequency definition of probability as $S(X, A)$, the degree to which a bivalent superset X, the sample space, is a subset of its own subset A. The concept of randomness never enters the deterministic framework.

Suppose A and B are nonfuzzy subsets of X. (X, like every observed set, is at

most countably infinite.) Suppose A is a subset of B. In the extreme case $B=X$. Then the degree of subsethood $S(B, A)$ is what has traditionally been called a *relative frequency*:

$$S(B, A)=\frac{M(A)}{M(B)}$$

$$=\frac{n_A}{N},$$

where the N elements of B constitute the *de facto* universe of discourse of the "experiment". (Of course the limit of the ratio $S(B, A)$ can be taken if it mathematically makes sense.[7]) The probability n_A/N has been reduced to degrees of subsethood, a purely fuzzy set-theoretical relationship. An immediate historical speculation is that if set theory had been more carefully worked out first, the notion of "randomness" might never have culturally evolved.

A classical example of relative frequency is the number n_A of successful trials in N trials. A biological example is the number of blue-eyed genes or alleles at all such chromosomal loci in a gene pool. The new way of expressing these relative frequencies $S(B, A)$ is the degree to which all trials are successes or all genes at a specific chromosomal location are for blue-eyedness. If the distinction between successful and unsuccessful trials is not clear cut, the resulting fuzzy relative frequency $S(B, A)$ may be real-valued. The frequency structure remains since A is a subset of B (since $B=X$ invariably in practice).

Where did the "randomness" go? The relative frequency $S(B, A)$ describes a fuzzy state of affairs, the degree to which B belongs to the power set of A: $S(B, A)=m_{2^A}(B)$. (Consider $B=X$ and $A=\{x_2\}$ in the unit square: the frequency $S(X, A)$ corresponds by the Pythagorean Theorem to the ratio of the left cube edge and the long diagonal to X.) Whether $S(B, A)$ is a rational or irrational number seems a technicality, a matter of fineness of quantization, if it is not zero or one. In practice only physical objects like tossed coins and DNA strands are involved. Their individual behavior might be fully determined by a system of differential equations.

The key quantity is the measure of overlap $M(A \cap B)$. This count does not involve "randomness". It counts which elements are identical or similar and to what degree. The phenomena themselves are deterministic. The corresponding frequency number that summarizes the deterministic situation is also deterministic. The same situation always gives the same number. The number may be used also to place bets or to switch a phone line, but it remains part of the description of a specific state of affairs. The deterministic subsethood derivation of relative frequency eliminates the need to invoke an undefined "randomness" to further describe the situation.

The identification of relative frequency with probability is cultural, not logical. This may take getting used to after hundreds of years of casting gambling intuitions as matters of probability and a century of building probability into the description of the universe. It is ironic that to date every assumption of probability—at least in the relative frequency sense of science, engineering, gambling, and daily life—has actually been an invocation of fuzziness.

9. BAYESIAN POLEMICS

Bayesian probabilists interpret probability as a subjective state of knowledge. In practice they use relative frequencies (subsethood degrees) but only to approximate these "states of knowledge".

Bayesianism is a polemical doctrine. Bayesians claim that they, and only they, use all and only the available uncertainty information in the description of uncertain phenomena. This stems from the Bayes Theorem expansion of the "*a posteriori*" conditional probability $P(H_i|E)$, the probability that H_i, the ith of k-many disjoint hypotheses $\{H_j\}$, is true when evidence E is observed:

$$P(H_i|E) = \frac{P(E \cap H_i)}{P(E)}$$

$$= \frac{P(E|H_i)P(H_i)}{P(E)}$$

$$= \frac{P(E|H_i)P(H_i)}{\sum_{j=1}^{k} P(E|H_j)P(H_j)},$$

since the hypotheses partition the sample space X: $H_1 \cup H_2 \cup \cdots \cup H_k = X$ and $H_i \cap H_j = \varnothing$ if $i \neq j$.

Conceptually, Bayesians use all available information in computing this posterior distribution by using the "*a priori*" or prior distribution $P(H_i)$ of the hypotheses. Mathematically, the Bayesian approach clearly stems from the ratio form of the conditional probability.

The Subsethood Theorem trivially implies Bayes Theorem when the hypotheses $\{H_i\}$ and evidence E are nonfuzzy subsets. More important, the Subsethood Theorem implies the Fuzzy Bayes Theorem in the more interesting case when the observed data E is fuzzy:

$$S(E, H_i) = \frac{S(H_i, E)M(H_i)}{\sum_{j=1}^{k} S(H_j, E)M(H_j)}$$

$$= \frac{S(H_i, E)f_i}{\sum_{j=1}^{k} S(H_j, E)f_j},$$

where

$$f_i = \frac{M(H_i)}{M(X)} = \frac{M(H_i)}{n} = S(X, H_i)$$

is the "relative frequency" of H_i, the degree to which all the hypotheses are H_i. So the Subsethood Theorem allows fuzzyists to be "Bayesians" as well.

The Subsethood Theorem implies inequality when the partitioning hypotheses are fuzzy. For instance, if $k = 2$, H^c is the complement of an arbitrary fuzzy set H, and evidence E is fuzzy, then[10] the occurrence of nondegenerate hypothesis overlap and underlap gives a lower bound on the posterior subsethood:

$$S(E,H) \geqq \frac{S(H,E) f_H}{S(H,E) f_H + S(H^c,E) f_{H^c}},$$

where $f_H = S(X,H)$. The lower bound is an increasing function of $M(H)$, a decreasing function of $M(H^c)$. Since a like lower bound holds for $S(E,H^c)$, adding the two posterior subsethoods gives the additive inequality:

$$S(E,H) + S(E,H^c) \geqq 1,$$

an inequality arrived at independently by Zadeh[17] by directly defining a "relative sigma-count" as the subsethood measure given by the Subsethood Theorem. If H is nonfuzzy, equality holds as in the additive law of conditional probability:

$$P(H|E) + P(H^c|E) = 1.$$

The Subsethood Theorem implies a deeper Bayes Theorem for arbitrary fuzzy sets, the Odds-Form Fuzzy Bayes Theorem:

$$\frac{S(A_1 \cap H, A_2)}{S(A_1 \cap H, A_2^c)} = \frac{S(A_2 \cap H, A_1)}{S(A_2^c \cap H, A_1)} \frac{S(H, A_2)}{S(H, A_2^c)}.$$

This theorem is proved directly by replacing the subsethood terms on the righthand side with their equivalent ratios of counts, canceling like terms three times, multiplying by $M(A_1 \cap H)/M(A_1 \cap H)$, rearranging, and applying the Subsethood Theorem a second time.

We have now developed enough fuzzy theory to examine critically the recent anti-fuzzy polemics of Lindley[13] and Jaynes[6] (and thus Cheeseman[2] who uses Jaynes' arguments). To begin we observe four more corollaries of the Subsethood Theorem:

i) $$0 \leqq S(H,A) \leqq 1,$$

ii) $$S(H,A) = 1 \quad \text{if } H \subset A,$$

iii) $$S(H, A_1 \cup A_2) = S(H,A_1) + S(H,A_2) - S(H, A_1 \cap A_2),$$

iv) $$S(H, A_1 \cap A_2) = S(H,A_1) S(A_1 \cap H, A_2).$$

Each relationship follows from the ratio form of $S(A,B)$. The third relationship uses the additivity of the count $M(A)$, which follows from $\min(x,y) + \max(x,y) = x + y$.

Now make the notational identification $S(H,A) = P(A|H)$. We then obtain the defining relationships of conditional probability proposed by Lindley:[13]

Convexity: $$0 \leqq P(A|H) \leqq 1 \quad \text{and } P(A|H) = 1 \quad \text{if } H \text{ implies } A,$$

Addition: $$P(A_1 \cup A_2|H) = P(A_1|H) + P(A_2|H) - P(A_1 \cap A_2|H),$$

Multiplication: $P(A_1 \cap A_2 | H) = P(A_1 | H) P(A_2 | A_1 \cap H)$.

"From these three rules", Lindley[13] tells us,

all of the many, rich and wonderful results of the probability calculus follow. They may be described as the axioms of probability.

Lindley takes these as "unassailable" axioms:

We really have no choice about the rules governing our measurement of uncertainty: they are dictated to us by the inexorable laws of logic.

Lindley proceeds to build a "coherence" argument around the Odds-Form Bayes Theorem, which he correctly deduces from the axioms as the equality:

$$\frac{P(A_2 | A_1 \cap H)}{P(A_2^c | A_1 \cap H)} = \frac{P(A_1 | A_2 \cap H)}{P(A_1 | A_2^c \cap H)} \frac{P(A_2 | H)}{P(A_2^c | H)},$$

where here we interpret A^c as not-A. "Any other procedure", we are told, "is incoherent." This polemic evaporates in the face of the above four subsethood corollaries and the Odds-Form Fuzzy Bayes Theorem. Ironically, rather than establish the primacy of axiomatic probability, Lindley seems to argue that it is fuzziness in disguise.

Another source of Bayesian probability polemic[2] is maximum entropy estimation. Here the axiomatic argument rests on the so-called Cox's Theorem.[3] Cox's Theorem is best presented by its most vocal proponent, physicist E. T. Jaynes.

According to Jaynes:[6]

Cox proved that any method of inference in which we represent degrees of plausibility by real numbers, is necessarily either equivalent to Laplace's, or inconsistent,

where Laplace is cited as an early Bayesian probabilist. In fact Cox used *bivalent* logic (Boolean algebra) and other assumptions to show that, again according to Jaynes, the "conditions of consistency can be stated in the form of functional equations," namely the probabilistic product and sum rules:

$$P(A \cap B | C) = P(A | B \cap C) P(B | C),$$

$$P(B | A) + P(B^c | A) = 1.$$

The Subsethood Theorem implies

$$S(C, A \cap B) = S(B \cap C, A) S(C, B),$$

$$S(A, B) + S(A, B^c) \geqq 1,$$

with, as we have seen, equality holding for the second subsethood relationship when B is nonfuzzy, which is the case in the Cox–Jaynes setting.

In the probabilistic case overlap and underlap are degenerate. So

$$P(A \cap A^c | B) = P(\varnothing | B) = \frac{P(\varnothing)}{P(B)} = 0,$$

and $P(B | A \cap A^c) = P(B | \varnothing)$ is undefined. Yet in general $S(B, A \cap A^c) > 0$ and $S(A \cap A^c, B)$ is defined when A is fuzzy and B is fuzzy or nonfuzzy.

Jaynes' claim is either false or concedes that probability is a special case of fuzziness. For strictly speaking, since the subsethood measure $S(A, B)$ satisfies the multiplicative and additive laws specified by Cox and yet differs from the conditional probability $P(B|A)$, Jaynes' claim is false.

Presumably Jaynes was unaware of fuzzy sets. He seems to suggest that the only alternative uncertainty theory is the frequency theory of probability, a theory we have seen reduced to the subsethood measure $S(X, A)$. So if we restrict consideration to nonfuzzy sets A and B, equality holds in the above subsethood relations and Jaynes is right: probability and fuzziness coincide. But fuzziness exists, indeed abounds, outside this restriction and classical probability theory does not. So fuzzy theory is an extension of probability theory. Equivalently, probability then is a special case of fuzziness.

Incidentally, when one examines Cox's actual arguments,[3] one finds that Cox assumes that the uncertainty combination operators in question are continuously *twice differentiable*! Min and max are not twice differentiable. Technically, Cox's theorem does not apply.

10. THE ENTROPY–SUBSETHOOD THEOREM

The Fuzzy Entropy Theorem and the Subsethood Theorem were independently derived from first principles, from sets-as-points unit-cube geometry. Both theorems involve ratios of cardinalities. A connection is inevitable.

The Entropy–Subsethood Theorem shows that the connection occurs in terms of overlap $A \cap A^c$ and underlap $A \cup A^c$ (what else?). The theorem says fuzzy entropy can be eliminated in favor of subsethood. So subsethood emerges as the fundamental, characterizing quantity of fuzziness—and, arguably, of probability as well.

ENTROPY–SUBSETHOOD THEOREM

$$E(A) = S(A \cup A^c, A \cap A^c).$$

The theorem is proved by replacing B and A in the Subsethood Theorem respectively with overlap $A \cap A^c$ and underlap $A \cup A^c$. Since overlap is a (dominated-membership function) subset of underlap, the intersection of the two sets is just overlap.

The Entropy–Subsethood Theorem is a peculiar relationship. It says that fuzziness is the degree to which the superset $A \cup A^c$ is a subset of its own subset $A \cap A^c$, the extent to which the whole is a part of one of its own parts, a relationship forbidden by Western logic.

This relationship violates our ingrained Venn-diagram intuitions of unambiguous set inclusion. Only the midpoint of I^n yields total containment of underlap in overlap. The cube vertices yield no containment. This parallels in the extreme the relative frequency relationship $S(X, A) = n_A/N$, where a nonfuzzy superset X is to some degree a subset of one of its nonfuzzy subsets A.

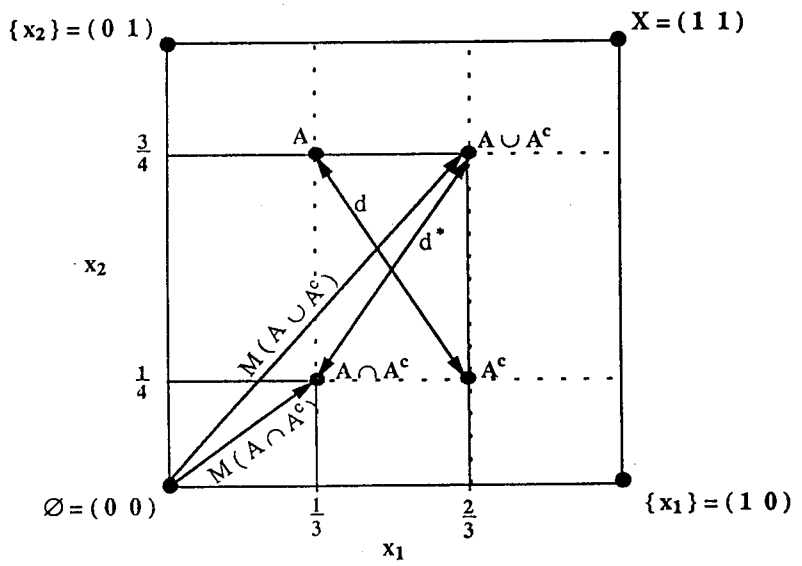

Figure 11 Entropy–Subsethood Theorem in two dimensions. Just as the long diagonals have equal length, $d(A, A^c) = d(A \cup A^c, A \cap A^c) = d^* = M(A \cup A^c) - M(A \cap A^c)$, the shortest distance from $A \cup A^c$ to the fuzzy power set of $A \cap A^c$.

Figure 11 illustrates the Entropy–Subsethood Theorem. It shows that d^*, the shortest distance from underlap $A \cup A^c$ to the hyper-rectangle defining the fuzzy power set of overlap $A \cap A^c$, is equivalent to $d(A \cup A^c, A \cap A^c) = d(A, A^c)$ and to a difference of vector magnitudes: $d^* = M(A \cup A^c) - M(A \cap A^c)$.

The Entropy–Subsethood Theorem implies that no probability measure measures fuzziness. For the moment, suppose not. Suppose fuzzy entropy measures nothing new; fuzziness is simply disguised probability. Suppose, as Lindley[13] claims, that probability theory "is adequate for all problems involving uncertainty." So there exists some probability measure P such that $P = E$. P cannot be identically zero because $P(X) = 1$. Then there is some A such that $P(A) = E(A) > 0$. But in a probability space there is no overlap or underlap: $A \cap A^c = \emptyset$ and $A \cup A^c = X$.

The Entropy–Subsethood Theorem then implies that $0 < P(A) = E(A) = S(A \cup A^c, A \cap A^c) = S(X, \emptyset)$. The only way X can be a subset to any degree of the empty set is if X itself, and hence A, is empty: $X = A = \emptyset$. Then the sure event X is impossible: $P(X) = P(\emptyset) = 0$. Or the impossible event is sure: $P(\emptyset) = 1$. Either outcome is a bivalent contradiction, impervious to normalization. So there exists no probability measure P that measures fuzziness. Fuzziness exists.

This *within-cube* theory can be extended[11] to define a natural fuzzy integral with respect to the fuzzy counting measure M. A more practical extension[11] is to mappings *between* fuzzy cubes, in particular to *fuzzy associative memories.* In short, a fuzzy set is a point in a unit hypercube I^n. A *fuzzy system* $S: I^n \to I^p$ is a mapping between cubes. Fuzzy systems map fuzzy subsets of the input space X to fuzzy subsets of the output space Y. Fuzzy systems are tools of machine intelligence, and can be applied to a wide range of control and decision problems.

11. PRECISE PAST, FUZZY FUTURE[4]

The boat of uncertainty reasoning is being rebuilt at sea. Plank by plank fuzzy theory is beginning to gradually shape its design. Today only a few fuzzy planks have been laid. But a hundred years from now, a thousand years from now, the boat of uncertainty reasoning may little resemble the boat of today. Notions and measures of overlap $A \cap A^c$ and underlap $A \cup A^c$ will have smoothed its rudder. Amassed fuzzy applications, hardware, and products will have broadened its sails. And no one on the boat will believe that there was a time when a concept as simple, as intuitive, as expressive[18] as a *fuzzy set* met with such impassioned denial.

How would the world be different today if fuzziness had been developed, taught, and applied before probability theory? Suppose the fuzzy framework was worked out at the time of Galileo of Laplace. Suppose Isaac Newton included an appendix on the geometry of fuzzy sets in his *Principia.* What would be different today?

Reasoning systems in machine intelligence would surely be different. So would be the range of automatic control devices. There would be many more of them, and they would more accurately reflect our reasoning processes than do our current decision trees and thermostats. Western belief systems might be more Eastern, and vice versa. (How many Westerners can name five Eastern books?) More of social science might be systematized. Historical tendencies would have been easier to articulate and defend. Communication, signal processing, and computational hardware might be built around the *fit.* Our physical explorations of subatomic reality, antimatter, and the spacetime fabric may have led to different times and places. Relative frequencies might be considered the everyday application of fuzzy subsethood. Besides betting on games of chance or frequency, betting on games of degree—perhaps involving simulated chaotic trajectories in unit cubes (or guppies swimming in hand-held cubical aquaria) or real-valued dice—might help support the economy of Las Vegas.

As the total amount of information in society continues to grow exponentially, the velocity of scientific and cultural change increases. Cultural change that once took centuries can now occur in a few years, perhaps soon in a single year. A current engineering example of this velocity of change is Moore's Law, the doubling of silicon-chip transistor density every one to two years.

One tendency of this information acceleration is to leave further behind what has already been explored. The complementary tendency is to soon experiment with systems that may at present seem distant, impractical, even absurd. In this light the recent developments in fuzzy theory and in fuzzy applications and hardware will surely affect the science, engineering, and culture of the future. The question is to what degree.

ACKNOWLEDGMENT

This research was supported by the Air Force Office for Scientific Research (AFOSR-88-0236) and by a grant from the Rockwell Science Center.

REFERENCES

1. W. Bandler and L. Kohout, "Fuzzy power sets and fuzzy implication operators." *Fuzzy Sets and Systems*, **4**, 1980, pp. 13–30.

2. P. Cheeseman, "In defense of probability." *Proc. of the IJCAI-85*, Aug. 1985, pp. 1002–1009.
3. R. T. Cox, "Probability, frequency, and reasonable expectations." *American Journal of Physics*, **14,** No. 1 Jan./Feb. 1946, pp. 1–13.
4. B. R. Gaines, "Precise past, fuzzy future." *International Journal of Man–Machine Studies*, **19,** 1983, pp. 117–134.
5. D. Hume, *An Inquiry Concerning Human Understanding*, 1748.
6. E. T. Jaynes, "Where do we stand on maximum entropy?" In: *The Maximum Entropy Formalism*, edited by Levine and Tribus, MIT Press, Cambridge, Mass., 1979.
7. M. Kac, *Probability and Related Topics in Physical Sciences, Lectures in Applied Mathematics*, Vol. I. Interscience, New York, 1959.
8. G. J. Klir and T. A. Folger, *Fuzzy Sets, Uncertainty, and Information.* Prentice-Hall, Englewood Cliffs, New Jersey, 1988.
9. B. Kosko, "Counting with fuzzy sets." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-8,** July 1986, pp. 556–557.
10. B. Kosko, "Fuzzy entropy and conditioning." *Information Sciences*, **40,** 1986, pp. 165–174.
11. B. Kosko, *Foundations of Fuzzy Estimation Theory.* Ph.D. dissertation, Department of Electrical Engineering, University of California at Irvine, June 1987; Order Number 8801936, University Microfilms International, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.
12. B. Kosko, "Fuzzy quantum states." in preparation, 1990.
13. D. V. Lindley, "The probability approach to the treatment of uncertainty in artificial intelligence and expert systems." *Statistical Science*, **2,** No. 1, Feb. 1987, pp. 17–24.
14. H. Prade, "A computational approach to approximate and plausible reasoning with applications to expert systems." *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **7,** 1985, pp. 260–283.
15. L. Wittgenstein, *Tractacus Logico-Philosophicus.* Routledge & Kegan Paul, London, 1922.
16. L. A. Zadeh, "Fuzzy sets." *Information and Control*, **8,** 1965, 338–353.
17. L. A. Zadeh, "A computational approach to fuzzy quantifiers in natural languages." *Computers and Mathematics*, **9,** No. 1, 1983, pp. 149–184.
18. L. A. Zadeh, "Making computers think like people." *IEEE Spectrum*, Aug. 1984, pp. 26–32.

Bart Kosko is an assistant professor in the electrical engineering department of the University of Southern California. He received bachelor degrees in philosophy and economics from USC, a masters degree in applied mathematics from UC San Diego, and a Ph.D. degree in electrical engineering from UC Irvine. Dr. Kosko is an elected member of the governing board of the International Neural Network Society, managing editor of Springer-Verlag's *Lecture Notes in Neural Computing* monograph series, and associate editor of *IEEE Transactions on Neural Networks*, *Journal of Mathematical Biology*, *Lecture Notes in Biomathematics*, and *Neural Networks*. Dr. Kosko was program chairman of the 1987 and 1988 IEEE International Conferences on Neural Networks and is program co-chairman of the 1990 International Joint Conference on Neural Networks and program chairman of the 1990 International Fuzzy Logic and Neural Network Conference in Iizuka, Japan. Dr. Kosko is a USC Shell Oil Faculty Fellow.