GENERALIZED INFORMATION THEORY

Tutorial at ISIPTA 2007 in Prague

George J. Klir
Department of Systems Science & Industrial Engineering
Binghamton University (State University of New York)
Binghamton, New York 13902, USA
gklir@binghamton.edu

OUTLINE

Part 1: Overall Characterization of Generalized Information Theory

1. Generalized Information Theory (GIT): An Overview
2. Two-Dimensional Expansion of Classical Information Theory
3. Uncertainty Theories Viewed as Theories of Imprecise Probabilities
4. Canonical Representation of Imprecise Probabilities

Part 2: Measuring Uncertainty and Information

5. Required Properties of Measures of Uncertainty and Information
6. An Overview of Measures of Uncertainty and Information Developed Within GIT
7. Some Methodological Issues Within GIT
8. Current and Future Research in GIT

GENERALIZED INFORMATION THEORY (GIT)

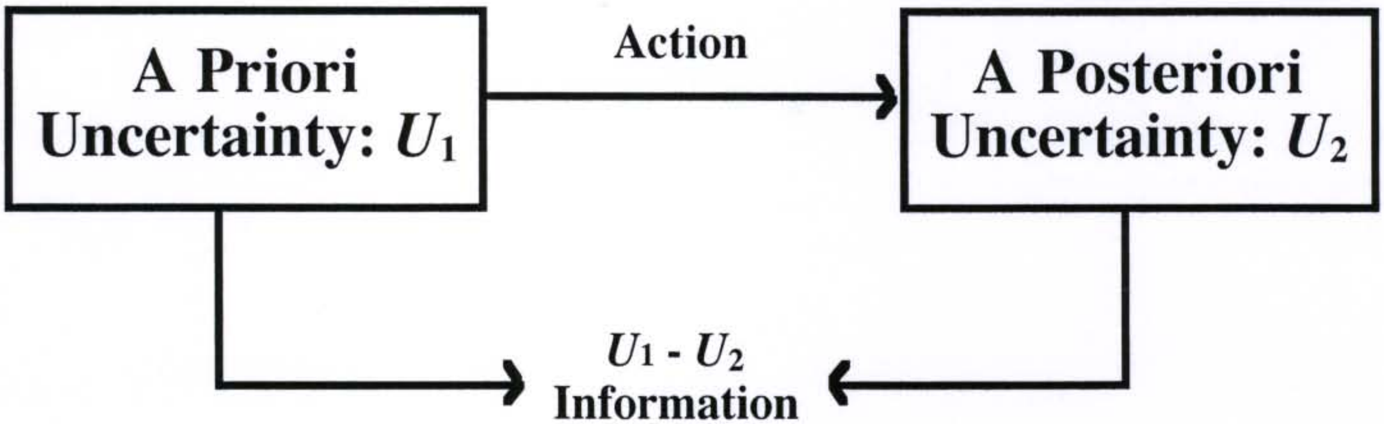
- **GIT is a research program whose objective is to develop a formal treatment of the interrelated concepts of uncertainty and information in all their varieties; it is a generalization of two distinct branches of classical information theory, which are based, respectively, on the notions of possibility (crisp) and probability.**

- **In GIT, as in classical information theory, uncertainty (predictive, retrodictive, diagnostic, prescriptive, etc.) is viewed as a manifestation of some information deficiency, while information is viewed as the capacity to reduce uncertainty. That is, GIT deals with information-based uncertainty and uncertainty-based information.**

- **The aims of GIT were introduced in 1991 in my paper "Generalized Information Theory" [*Fuzzy Sets and Systems,* 40(1), pp. 127-142].**

- **Comprehensive and up-to date coverage of results obtained by research within GIT is contained in the text Uncertainty and Information [John Wiley, Hoboken, NJ, 2006].**

Uncertainty-based Information



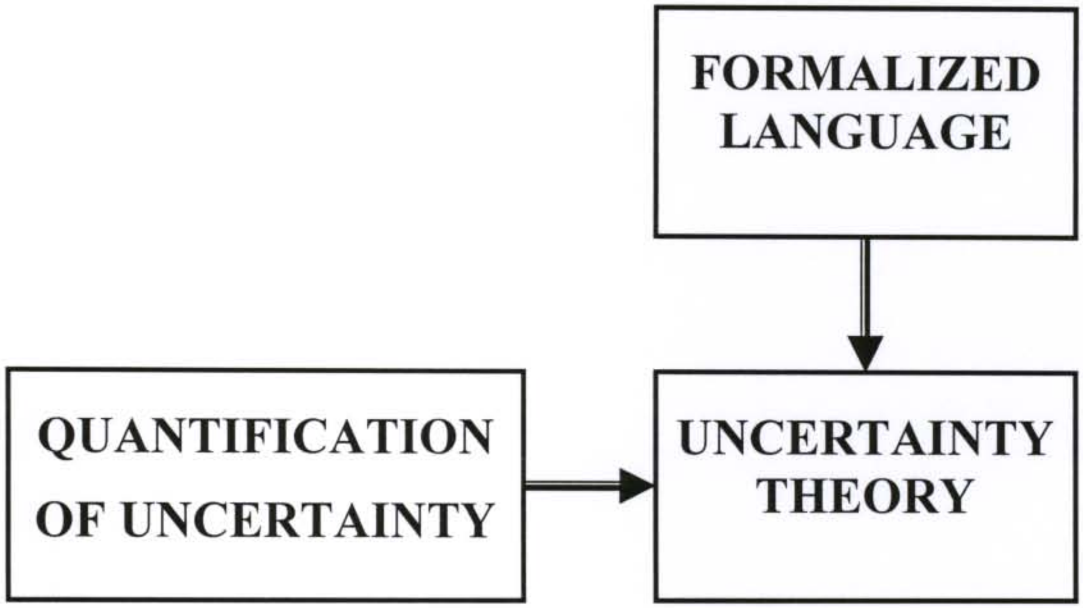
The amount of information obtained by an action

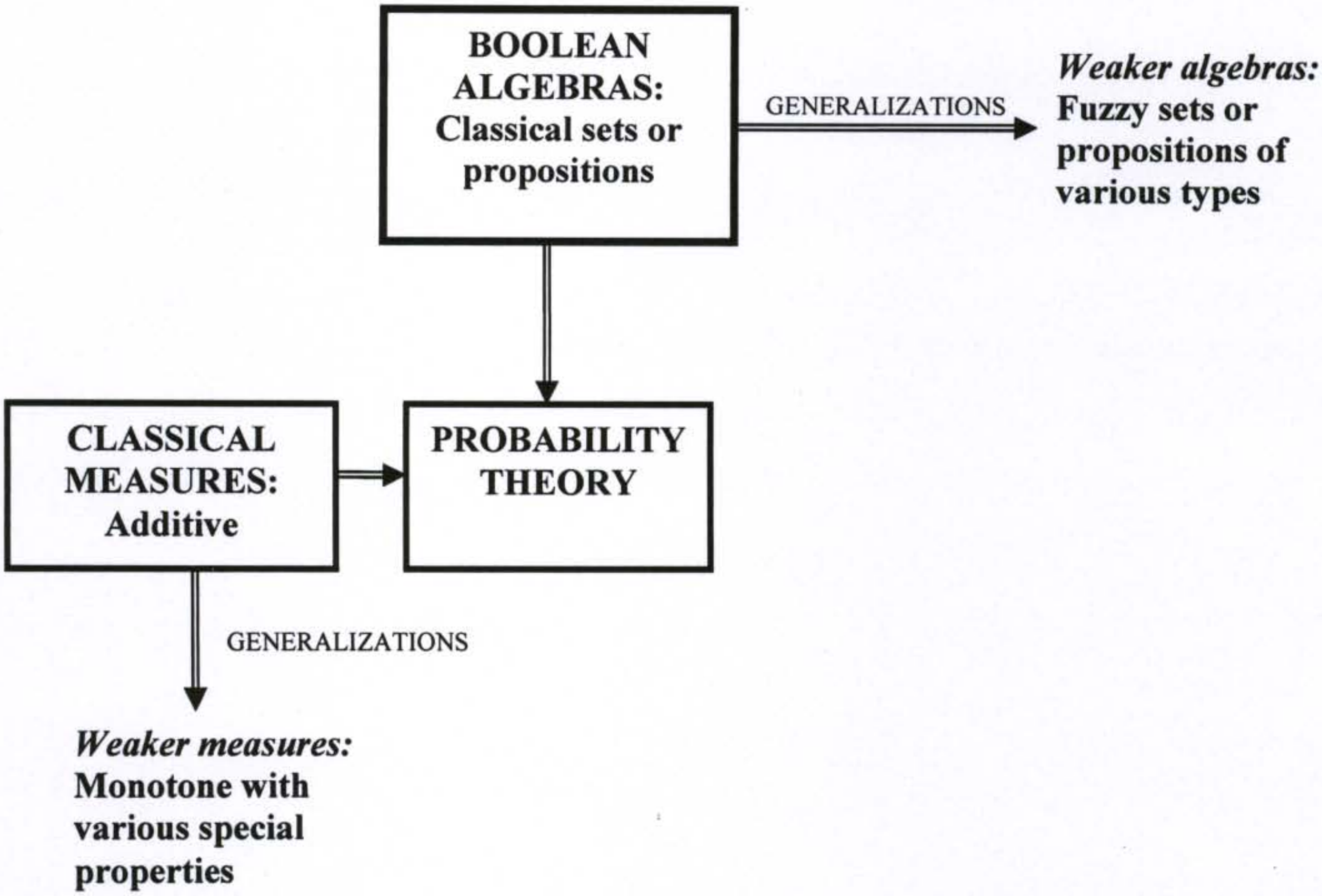
=

The amount of uncertainty reduced by the action

The mathematical theory of information had come into being when it was realized that the flow of information can be represented numerically in the same way as distance, mass, temperature, etc.

(Alfréd Rényi)


FORMALIZED LANGUAGE
QUANTIFICATION OF UNCERTAINTY
UNCERTAINTY THEORY


BOOLEAN ALGEBRAS: Classical sets or propositions
GENERALIZATIONS
Weaker algebras: Fuzzy sets or propositions of various types
CLASSICAL MEASURES: Additive
PROBABILITY THEORY
GENERALIZATIONS
Weaker measures: Monotone with various special properties

MONOTONE MEASURES

Given a universal set X and a non-empty family $\mathcal{C}$ of subsets of X (usually with an appropriate algebraic structure), a *monotone measure* (also called a fuzzy measure), μ, on ⟨X, $\mathcal{C}$⟩ is a function

$$\mu: \mathcal{C} \rightarrow [0, \infty]$$

that satisfies the following requirements:

(1) $\mu(\varnothing) = 0$ (*vanishing at the empty set*);

(2) for all A, B $\in \mathcal{C}$, if $A \subseteq B$, then $\mu(A) \leq \mu(B)$

(*monotonicity*);

(3) for any increasing sequence $A_1 \subseteq A_2 \subseteq \ldots$ of sets in $\mathcal{C}$,

$$\text{if } \bigcup_{i=1}^{\infty} A_i \in \mathcal{C}, \text{ then } \lim_{i\to\infty} \mu(A_i) = \mu\left(\bigcup_{i=1}^{\infty} A_i\right)$$

(*continuity from below*);

(4) for any decreasing sequence $A_1 \supseteq A_2 \supseteq \ldots$ of sets in $\mathcal{C}$,

$$\text{if } \bigcap_{i=1}^{\infty} A_i \in \mathcal{C}, \text{ then } \lim_{i\to\infty} \mu(A_i) = \mu\left(\bigcap_{i=1}^{\infty} A_i\right)$$

(*continuity from above*).

FUZZY SETS: Basic Characteristics

- **Sets whose boundaries are not sharp.**
- **Sets that allow to distinguish degrees (or grades) of membership.**
- **Sets that are fully characterized by membership functions of the form $F: D \rightarrow R$.**
- **Distinct categories of fuzzy sets are distinguished by distinct types of sets D (domains) and R (ranges) that are employed in defining their membership functions.**
- **For each $x \in D$, $F(x)$ is viewed as the degree of membership of object x in fuzzy set F.**
- **$F(x)$ may also be interpreted as the degree of compatibility of object x with a given concept represented by fuzzy set F.**
- **Membership functions of standard fuzzy sets have the form $F: X \rightarrow [0, 1]$, where X is a classical (crisp) set (universal set) whose elements are not fuzzy sets.**

α-Cuts of Standard Fuzzy Sets

- **For each $\alpha \in [0,1]$, the set ${}^{\alpha}A = \{x \in X \mid A(x) \geq \alpha\}$ is called an α-cut of standard fuzzy set A whose membership function has the form $A: X \rightarrow [0,1]$.**

- **Any standard fuzzy set is uniquely represented by its α-cuts for all $\alpha \in [0,1]$.**

- **Properties of classical sets can be extended to fuzzy sets (fuzzified) by requiring that they be preserved in all α-cuts. This kind of fuzzification is called a cutworthy fuzzification.**

UNCERTAINTY THEORIES			FORMALIZED LANGUAGES						
			CLASSICAL SETS	NONCLASSICAL SETS					
				STANDARD FUZZY SETS	NONSTANDARD FUZZY SETS				
					INTERVAL VALUED	TYPE 2	LEVEL 2	LATTICE BASED	•••
MONOTONE MEASURES	ADDITIVE	CLASSICAL NUMERICAL PROBABILITY							
	NONADDITIVE	POSSIBILITY/ NECESSITY							
		SUGENO λ-MEASURES							
		BELIEF/ PLAUSIBILITY (CAPACITIES OF ORDER ∞)							
		CAPACITIES OF VARIOUS FINITE ORDERS							
		INTERVAL-VALUED PROBABILITY DISTRIBUTIONS							
		• • •							
		GENERAL LOWER AND UPPER PROBABILITIES							

DIVERSITY AND UNITY OF UNCERTAINTY THEORIES

- It is significant that the enormous and ever increasing diversity of uncertainty theories subsumed under GIT is balanced by some common features they share.

- The diversity of uncertainty theories is in some sense desirable. It enables us to focus on the development of those theories that are promising from the standpoint of various applications. However, it makes it increasingly more difficult to get oriented within the many existing or emerging theories and the many prospective theories.

- It turns out that all recognizable uncertainty theories in GIT can be classified in a useful way so that some properties of the theories are invariant within each class. This unity of uncertainty within each class allows us to work within the class as a whole.

- A significant class of uncertainty theory with common properties consists of the various theories of imprecise probabilities defined on finite classical sets.

THEORIES OF UNCERTAINTY

In order to develop a <u>fully operational theory</u>, T, for dealing with uncertainty of some conceived type requires that a host of issues be addressed at <u>the following four levels</u>:

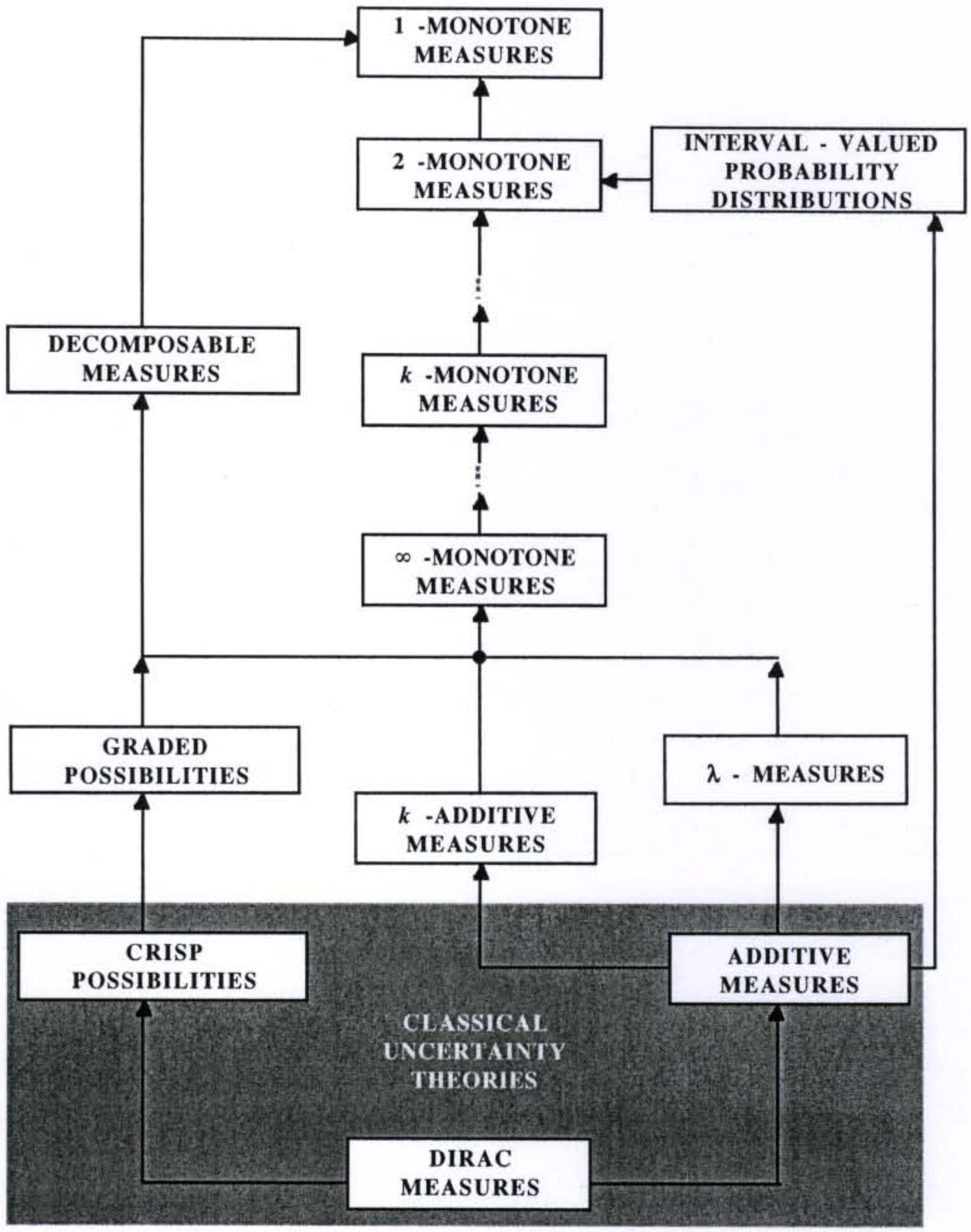
- **LEVEL 1 --- we need to find an appropriate <u>mathematical representation</u> of the conceived type of uncertainty, which is achieved by characterizing, via appropriate <u>axioms</u>, a class of <u>uncertainty functions</u>, say functions u, that represent uncertainty in theory T.**

- **LEVEL 2 --- we need to develop <u>operating rules</u> <u>(calculus)</u> for manipulating the uncertainty functions u in theory T.**

- **LEVEL 3 --- we need to find a meaningful way of <u>measuring</u> the amount of relevant uncertainty in any situation formalizable in theory T, which is achieved by finding a justifiable functional, U, which for each uncertainty function u in theory T measures the amount of uncertainty associated with it.**

- **LEVEL 4 --- we need to develop <u>methodological aspects</u> of theory T by utilizing functional U as an abstract <u>measuring instrument</u>.**

Uncertainty Theories



CHOQUET CAPACITIES OF ORDER k

($k = 2, 3, \ldots, \infty$)

- **Alternative name: k-monotone measures ($k \geq 2$).**
- **2-monotone measures are defined for all pairs A, B of subsets of X by the inequality:**

$$\mu(A \cup B) \geq \mu(A) + \mu(B) - \mu(A \cap B).$$

- **3-monotone measures are defined for all triples A, B, C of subsets of X by the inequality**

$$\mu(A \cup B \cup C) \geq \mu(A) + \mu(B) + \mu(C) - \mu(A \cap B) - \mu(A \cap C) - \mu(B \cap C) + \mu(A \cap B \cap C)$$

- **k-monotone measures are defined for all families of k subsets of X by the inequality**

$$\mu\left(\bigcup_{j=1}^{k} A_j\right) \geq \sum_{\substack{K \subseteq N_k \\ K \neq \varnothing}} (-1)^{|K|+1} \mu\left(\bigcap_{j \in K} A_j\right)$$

- **1-monotone measures is a convenient name for superadditive measures that satisfy for all disjoint pairs of subsets A and B of X the inequality**

$$\mu(A \cup B) \geq \mu(A) + \mu(B).$$

IMPRECISE PROBABILITIES: Canonical Representations

1. **Lower probability function: μ_***
2. **Upper probability function: μ^***
3. **Möbius functions: m**
4. **Convex set of probability distributions: D**

CONVERSIONS

- **1 $\Leftrightarrow$ 2: duality equation**
- **1 $\Leftrightarrow$ 3: Möbius transform**
- **1 $\Rightarrow$ 4: constructing extreme points of D**
- **4 $\Rightarrow$ 1: $\mu_*(A) = \inf_{p \in D}\{\sum_{x \in A} p(x)\}, \forall A$**

LOWER AND UPPER PROBABILITY MEASURES ASSOCIATED WITH A CONVEX SET OF PROBABILITY DISTRIBUTIONS D on X

Notation: $^{D}\mu_{*}$ denotes the lower probability measure associated with D

$^{D}\mu^{*}$ denotes the upper probability measure associated with D

Basic formulas:

$$^{D}\mu_{*}(A) = \inf_{p\in D}\{\textstyle\sum_{x\in A} p(x)\}, \ \forall A$$

$$^{D}\mu^{*}(A) = \sup_{p\in D}\{\textstyle\sum_{x\in A} p(x)\}, \ \forall A$$

Duality of $^{D}\mu_{*}$ and $^{D}\mu^{*}$:

$$^{D}\mu^{*}(A) = 1 - {}^{D}\mu_{*}(\overline{A}), \ \forall A$$

CONVEX SET OF PROBABILITY DISTRIBUTIONS ASSOCIATED WITH A GIVEN LOWER PROBABILITY MEASURE

- **Let $X = \{x_1, x_2, ..., x_n\}$ and let $\sigma = (\sigma(x_1), \sigma(x_2), ..., \sigma(x_n))$ denote a permutation by which elements of X are reordered.**

- **Given any lower probability measure μ_* on the power set of X that is 2-monotone, the convex set of all probability distributions that dominate this measure, $D(\mu_*)$, is determined by its extreme points, which are probability distributions p_σ computed as follows:**

$$p_\sigma(\sigma(x_1)) = \mu_*(\{\sigma(x_1)\}),$$

$$p_\sigma(\sigma(x_2)) = \mu_*(\{\sigma(x_1), \sigma(x_2)\}) - \mu_*(\{\sigma(x_1)\}),$$

....

$$p_\sigma(\sigma(x_{n-1})) = \mu_*(\{\sigma(x_1), ... ,\sigma(x_{n-1})\}) - \mu_*(\{\sigma(x_1), ... ,\sigma(x_{n-2})\})$$

$$p_\sigma(\sigma(x_n)) = \mu_*(\{\sigma(x_1), ... ,\sigma(x_n)\}) - \mu_*(\{\sigma(x_1), ... ,\sigma(x_{n-1})\})$$

- **Each permutation defines an extreme point of $D(\mu_*)$, but different permutations can give rise to the same point.**

- **$D(\mu_*)$ is the convex hull of the extreme points.**

MÖBIUS REPRESENTATION OF LOWER PROBABILITIES

- **Given any lower probability $\underline{\mu}$ on P(X) its Möbius representation, m, is obtained for all A $\in$ P(X) via the formula (Möbius transform)**

$$m(A) = \sum_{B \mid B \subseteq A} (-1)^{|A-B|} \underline{\mu}(B).$$

- **It is guaranteed that m(∅) = 0 and**

$$\sum_{A \in P(X)} m(A) = 1.$$

- **The inverse transform is given for all A $\in$ P(X) by the formula**

$$\underline{\mu}(A) = \sum_{B \mid B \subseteq A} m(B).$$

UNCERTAINTY MEASURES: Key Requirements

1. **Subadditivity: The amount of uncertainty in a joint representation of evidence cannot be greater than the sum of the amounts of uncertainty in the associated marginal representations of uncertainty.**

2. **Additivity: The two amounts of uncertainty considered under subadditivity become equal when the marginal representations of evidence are noninteractive according to the rules of the uncertainty calculus involved.**

3. **Range: The range of uncertainty is [0, M], where 0 must be assigned to the unique uncertainty function that describes full certainty and M depends on the cardinality of the universal set involved and on the chosen unit of measurement.**

4. **Continuity: Any measure of uncertainty must be continuous.**

5. **Expansibility: Expanding the universal set by alternatives that are not supported by evidence must not affect the amount of uncertainty.**

6. **Branching/Consistency: When uncertainty can be computed in several distinct ways, each conforming to the calculus of the theory, the results must be the same (consistent).**

7. **Monotonicity: When evidence can be ordered in the theory, the measure of uncertainty must preserve this ordering.**

8. **Coordinate invariance: When evidence is expressed within some Euclidean space, uncertainty must not change under isometric transformation of coordinates.**

HARTLEY MEASURE OF UNCERTAINTY

$$H(A) = \log_2 |A|$$

- **Among a given universal set X of all considered alternatives (predictions, retrodictions, diagnoses, etc.), only alternatives in set $A \subseteq X$ are <u>possible</u> according to given <u>evidence</u>.**

- **That is, alternatives in the complement of A are <u>not possible</u> according to given evidence.**

- $0 \leq H(A) \leq \log_2 |X|$

- **H(A) measures the degree of <u>nonspecificity</u> (or <u>imprecision</u>).**

- **H(A) has been extended to fuzzy sets A via their α-cut representations by the formula**

$$H(A) = \int_{\alpha \in [0,1]} \log_2 |{}^{\alpha}A| \, d\alpha.$$

- **<u>α-cut</u> of fuzzy set A:** ${}^{\alpha}A = \{x \in X \mid A(x) \geq \alpha\}$.

HARTLEY MEASURES ON $X \times Y$

Basic Types

- $H(X \times Y)$ — joint
- $H(X)$, $H(Y)$ — marginal (or simple)
- $H(X|Y)$, $H(Y|X)$ — conditional
- $T_H(X,Y)$ — information transmission

Basic Equations and Inequalities

- $H(X|Y) = H(X \times Y) - H(Y)$
- $H(Y|X) = H(X \times Y) - H(X)$
- $T_H(X,Y) = H(X) + H(Y) - H(X \times Y)$
- $H(X \times Y) \leq H(X) + H(Y)$
- $H(X|Y) \leq H(X)$ and $H(Y|X) \leq H(Y)$

Additivity under Independence

- $H(X \times Y) = H(X) + H(Y)$
- $H(X|Y) = H(X)$ and $H(Y|X) = H(Y)$
- $T_H(X,Y) = 0$

SHANNON ENTROPY

$$S(p(x)|x\in X) = -\sum_{x\in X} p(x)\log_2 p(x)$$

$$= -\sum_{x\in X} p(x)\log_2[1 - \sum_{y\neq x} p(y)]$$

- **$\text{Con}(x) = \sum_{y\neq x} p(y) \in [0, 1]$ for each $x \in X$ expresses the total conflict (aggregated) between the evidential claim focusing on x and all the other evidential claims expressed by the probability distribution $\langle p(x)|x\in X\rangle$.**

- **Function $-\log_2[1 - \text{Con}(x)]$ is monotone increasing with Con(x) and extend the range of Con(x) from [0, 1] to [0, ∞). Hence, it also expresses the total conflict within any given probability distribution $\langle p(x)|x\in X\rangle$, but in a different scale from Con(x).**

- **This alternative representation of the total conflict is needed to satisfy the additivity requirement of uncertainty measures.**

- **Shannon entropy can thus be viewed as a measure of the total conflict among evidential claims associated with a probability distribution.**

JOINT, MARGINAL, AND CONDITIONAL UNCERTAINTIES: Basic Formulas

$$U(X|Y) = U(X,Y) - U(Y)$$
$$U(Y|X) = U(X,Y) - U(X)$$
$$U(X_1, X_2,\cdots, X_n) = U(X_1) + U(X_2|X_1) + U(X_3|X_1,X_2) + \cdots + U(X_n|X_1, X_2,\cdots, X_{n-1})$$

$$\left.\begin{array}{l} U(X|Y) \leq U(X) \\ U(Y|X) \leq U(Y) \\ U(X,Y) \leq U(X) + U(Y) \end{array}\right\}$$

The equalities are obtained only in the case of noninteraction.

Information transmission:

$$T(X, Y) = U(X) + U(Y) - U(X,Y)$$
$$T(X, Y) = U(X) - U(X|Y)$$
$$T(X, Y) = U(Y) - U(Y|X)$$

HARTLEY-LIKE MEASURE IN n-DIMENSIONAL EUCLIDEAN SPACE

$$HL(A) = \min_{t \in T} \left\{ \log_2 \left[\prod_{i=1}^{n} [1 + \mu(A_{i_t})] + \mu(A) - \prod_{i=1}^{n} \mu(A_{i_t}) \right] \right\}.$$

HARTLEY-LIKE MEASURE

<u>n = 1</u>

$$HL(A) = \log_2[1 + \mu(A)]$$

<u>n = 2</u>

$$HL(A) = \min_{t \in T} \log_2[1 + \mu(A_1) + \mu(A_2) + \mu(A)]$$

GENERALIZED HARTLEY MEASURE IN DST

$$GH(m) = \sum_{A \in F} m(A) \log_2 |A|$$

Conditional forms:

$$GH(X|Y) = GH(X,Y) - GH(Y)$$
$$GH(Y|X) = GH(X,Y) - GH(X)$$

Measures of Uncertainty on Finite Sets X

- **Hartley measure of nonspecificity in classical possibility theory:**

$$H(A) = \log_2|A|\ ,\ A \subseteq X.$$

- **Shannon measure (entropy) of conflict in classical probability theory:**

$$S(p(x) \mid x \in X) = -\sum_{x\in X} p(x) \log_2 p(x).$$

- **Generalized Hartley measure in Dempster-Shafer Theory (DST):**

$$GH(m) = \sum_{A\subseteq X} m(A) \log_2|A|$$

- **All intuitively reasonable candidates for a generalized Shannon measure failed the subadditivity requirement.**

- **Aggregated total uncertainty in any theory of imprecise probabilities:**

$$S^*(D) = \max_{p\in D}\{-\sum_{x\in X} p(x) \log_2 p(x)\}$$

This measure satisfies all the essential requirements for measures of uncertainty, but it is insensitive to changes in evidence.

ENTROPY-LIKE MEASURE IN DST: Attempts

- $$C(m) = -\sum_{A \in \mathcal{F}} m(A) \log_2 Bel(A) = -\sum_{A \in \mathcal{F}} m(A) \log_2 \left[1 - \sum_{B \not\subset A} m(B)\right]$$

  Confusion: Höhle (1982)

- $$E(m) = -\sum_{A \in \mathcal{F}} m(A) \log_2 Pl(A) = -\sum_{A \in \mathcal{F}} m(A) \log_2 \left[1 - \sum_{A \cap B = \emptyset} m(B)\right]$$

  Dissonance: Yager (1983)

- $$D(m) = -\sum_{A \in \mathcal{F}} m(A) \log_2 \left[1 - \sum_{B \in \mathcal{F}} m(B) \frac{|B - A|}{|B|}\right]$$

  Discord : Klir & Ramer (1990); conjunctive set-valued statements

- $$ST(m) = -\sum_{A \in \mathcal{F}} m(A) \log_2 \left[1 - \sum_{B \in \mathcal{F}} m(B) \frac{|A - B|}{|A|}\right]$$

  Strive: Klir & Parviz (1992); disjunctive set-valued statements

DISAGGREGATED UNCERTAINTIES

- **Disaggregated total uncertainty with two components:**

$$TU = (GH, S^* - GH)$$

- **An alternative disaggregated total uncertainty with two components:**

$$^{a}TU = (S^* - S_*, S_*)$$

- **Global disaggregated uncertainty that consists of three components:**

$$GU = (GH, S^* - S_*, S_*)$$

➢ $S_*(D) = \min_{p \in D}\{-\sum_{x \in X} p(x) \log_2 p(x)\}$

$S^*(D) = \max_{p \in D}\{-\sum_{x \in X} p(x) \log_2 p(x)\}$

$GH(m) = \sum_{A \in F} m(A) \log_2 |A|$

INDEPENDENCE OF MARGINAL UNCERTAINTIES VIA THE CLASSICAL PROBABILISTIC INDEPENDENCE APPLIED TO CONVEX SETS OF PROBABILITY DISTRIBUTIONS (STRONG OR CLASSICAL INDEPENDENCE)

- **Marginal sets: $X = \{x_i \mid i \in N_n\}$ and $Y = \{y_j \mid j \in N_m\}$**

- **Joint set: $Z = X \times Y$**

- **Given: convex sets of marginal probability distributions D_X and D_Y**

- **Under the assumption of strong independence, the joint set of probability distributions on Z, denoted by D, is defined by applying classical probabilistic independence to D_X and D_Y:**

  $$D = \{p \mid p(z_{ij}) = p_X(x_i).p_Y(y_j),\ p_X \in D_X,\ p_Y \in D_Y,\ i \in N_n,\ j \in N_m\}$$

- $\underline{\mu}(A) = \inf_{p \in D}\{\sum_{i,j} p(z_{ij}) \mid z_{ij} \in A\}, \forall A$

Möbius independence (also called mass independence):

$$m(C) = \begin{cases} m_X(A) \cdot m_Y(B) & \text{when } C = A \times B \\ 0 & \text{otherwise.} \end{cases}$$

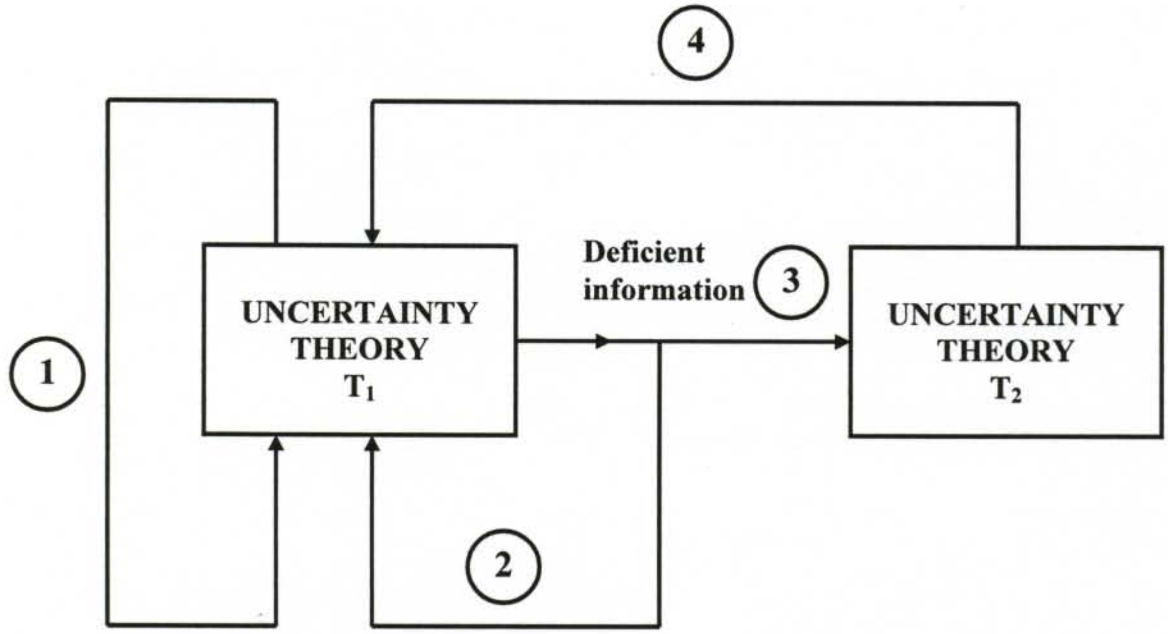
Possibilistic independence:

$$r(x,y) = \min\{r_X(x), r(y)\}, x \in X, y \in Y$$

	GH	S_*	S^*	$S^* - GH$	$S^* - S_*$
Strong independence	**No**	**Yes**	**Yes**	**No**	**Yes**
Mass independence	**Yes**	**No**	**Yes**	**Yes**	**No**

Principles of Uncertainty

- **Principle of Minimum Uncertainty**
- **Principle of Maximum Uncertainty**
- **Principle of Uncertainty Invariance**
- **Principle of Requisite Generalization**

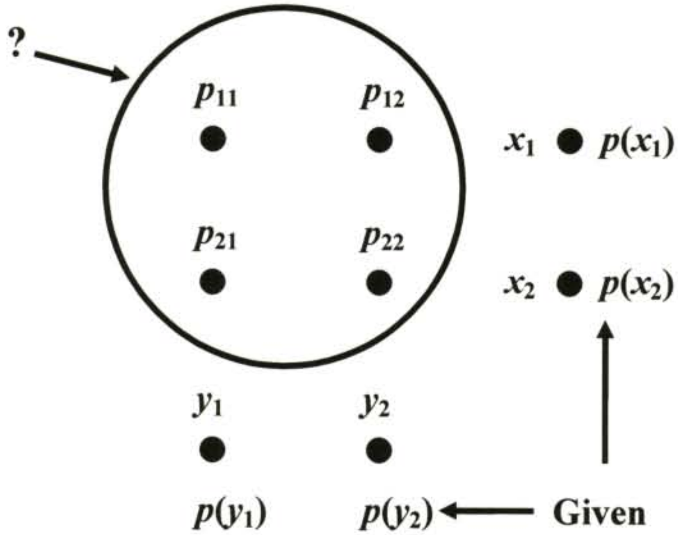


<u>Assumption</u>: **Theory T_2 is more general than theory T**

1. **Minimum – uncertainty principle**
2. **Maximum – uncertainty principle**
3. **Requisite generalization**
4. **Uncertainty – invariance principle**

THE MARGINAL PROBLEM

Given marginal probabilities (or marginal lower probabilities) what are the joint probabilities (or the joint lower probabilities)?



Example: $p(x_1) = 0.8$, $p(x_2) = 0.2$

$p(y_1) = 0.6$, $p(y_2) = 0.4$

- **Remaining in probability theory: maximum entropy principle**

$$p_{11} = p(x_1) \cdot p(y_1) = 0.48,\ \ p_{12} = p(x_1) \cdot p(y_2) = 0.32$$

$$p_{21} = p(x_2) \cdot p(y_1) = 0.12,\ \ p_{22} = p(x_2) \cdot p(y_2) = 0.08.$$

$$TU = \langle 0, 1.693 \rangle$$

- **Requisite generalization: the following convex set of joint probability distributions (epistemologically honest representation of evidence – the marginal probability distributions):**

$$p_{11} \in [\max\{0, p(x_1) + p(y_1) - 1\}, \min\{p(x_1), p(y_1)\}] = [0.4, 0.6]$$

$$p_{12} = p(x_1) - p_{11} = 0.8 - p_{11}$$

$$p_{21} = p(y_1) - p_{11} = 0.6 - p_{11}$$

$$p_{22} = 1 - p(x_1) - p(y_1) + p_{11} = p_{11} - 0.4$$

$$TU = \langle 0.332, 1.683 \rangle$$

Uncertainty Theories