GENERALIZED INFORMATION THEORY (GIT): A Digest

George J. Klir
State University of New York (SUNY)
Binghamton, New York 13902, USA
gklir@binghamton.edu

prepared for International Centre for Information and Uncertainty, Palacky University, Olomouc

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
INVESTMENTS IN EDUCATION DEVELOPMENT

OUTLINE

1. GIT as a Research Program
2. Classical Uncertainty Theories
3. Two-Dimensional Expansion of Classical Theories
4. Levels of Development in Each Uncertainty Theory
5. Uncertainty-Based Information
6. Functional U for Measuring Uncertainty
7. Imprecise Probabilities: Uncertainty Functions u
8. Imprecise Probabilities: Measuring Functionals U
9. Methododological Issues in Imprecise Probabilities
10. Some Open Problems 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.
- GIT was introduced in (Klir,1991).
- In GIT, as in the two classical information theories, uncertainty (predictive, retrodictive, diagnostic, prescriptive, etc.) is the primary concept and information is defined via reduction of uncertainty.
- That is, GIT deals with a special type of information, which is usually referred to as uncertainty-based information.
- Principal reference: (Klir 2006).

UNCERTAINTY THEORIES: General Scenario

- A set of mutually exclusive alternatives is considered (predictions, diagnoses, etc.).
- Only one of the alternatives is true (prediction, diagnosis, etc.), but we are not certain which one it is.
- Uncertainty about the true alternative is expressed differently in each theory.

CLASSICAL UNCERTAINTY THEORIES

- **POSSIBILISTIC:** Uncertainty results from more possible alternatives than one. Information is obtained by any evidence that some of the considered alternatives are not possible.

- **PROBABILISTIC:** Uncertainty results from a distribution of degrees of evidential claims from a fixed value among all considered alternatives. Information is obtained by any evidence that makes the distribution more discriminatory.

CLASSICAL POSSIBILISTIC UNCERTAINTY THEORY

- Given a finite set X of considered alternatives, possibilistic uncertainty is expressed by a possibility function $r: X \rightarrow \{0,1\}$.
- $r(x) = 1$ means that x is possible and $r(x) = 0$ means that, under given evidence, x is not possible, $x \in X$.
- Function r partitions set X into two subsets: X_0 and X_1.
- Information is obtained by any evidence that reduces the subset X_1 of possible alternatives.
- Possibility set function (nonadditive measure):

$$\mathrm{Pos}(A) = \max_{x \in A}\{r(x)\}, \forall A \subseteq X$$

CLASSICAL PROBABILISTIC UNCERTAINTY THEORY

- Given a finite set X of considered alternatives, uncertainty is expressed by a probability function p: $X \rightarrow [0,1]$ such that

$$\sum_{x \in X} p(x) = 1.$$

- Function p distributes the value 1 to alternatives in X according to their relative strength of support by given evidence.
- Information is obtained by any evidence that makes the distribution more discriminatory.
- Probability set function (classical additive measure):

$$\mathrm{Pro}(A) = \sum_{x \in A} p(x), \forall A \subseteq X.$$

TWO-DIMENSIONAL EXPANSION OF CLASSICAL UNCERTAINTY THEORIES

1. By generalizing classical possibility and probability measures via various useful types of monotone measures.

2. By generalizing classical sets via various useful types of fuzzy sets.

DIVERSITY AND COMMONALITY OF UNCERTAINTY THEORIES

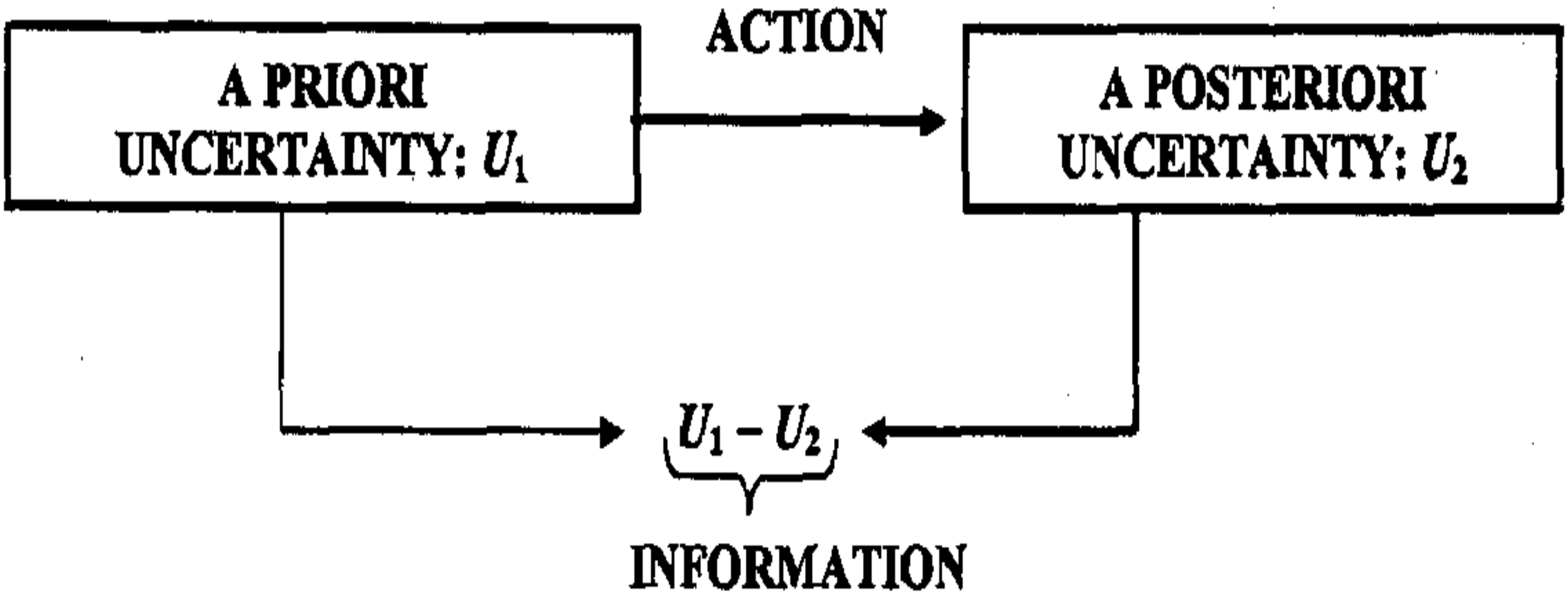
- The large diversity of uncertainty theories in GIT is balanced by some common features they share.
- All recognizable uncertainty theories in GIT can be classified in a useful way so that some properties of the theories in each class are invariant.
- The commonality of uncertainty theories within the individual classes makes it meaningful to work with each class as a whole.
- A significant class of uncertainty theories with some common properties consists of the various theories of imprecise probabilities.

UNCERTAINTY THEORIES:
Levels of Development

1. Formulating axioms of a conceived class of functions u that represent uncertainty in a given theory T (e.g. Kolmogorov's axioms in probability theory).
2. Developing calculus for manipulating functions u (e.g. probability calculus).
3. Establishing justifiable functional U on the class of functions u for measuring the amount of uncertainty captured by each function u (e.g. Shannon entropy).
4. Developing methodological aspects of theory T (e.g. maximum and minimum entropy principles).

UNCERTAINTY-BASED INFORMATION

The amount of information obtained by an action
is equal to
the amount of uncertainty reduced by the action.


ACTION
A PRIORI
UNCERTAINTY: U_1
A POSTERIORI
UNCERTAINTY: U_2
$U_1 - U_2$
INFORMATION

FUNCTIONAL *U* for Measuring Uncertainty : Key Requirements — Part 1

1. Subadditivity: The amount of uncertainty in a joint representation of evidence must be smaller than or equal to the sum of the amounts of uncertainty in the associated marginal representations of uncertainty.
2. Additivity: The equality between the two amounts of uncertainty compared in subadditivity is obtained if and only if the marginal uncertainty functions are independent according to the rules of the uncertainty calculus involved.
3. Range: The range of the amount of uncertainty must be [0, *M*], where 0 is obtained only for the unique uncertainty function in a given uncertainty theory that represent full certainty in the theory and *M* depends on the universal set and on the chosen measurement unit.
4. Continuity: *U* must be a continuous functional.

FUNCTIONAL *U* for Measuring Uncertainty : Key Requirements — Part 2

7. 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 given theory, the amount of uncertainty must be the same (consistent).
7. Monotonicity: When evidence can be ordered in the given theory, the amount of uncertainty must preserve this order.
8. Coordinate Invariance: When evidence is expressed within the Euclidean space, the amount of uncertainty must be invariant under isometric transformation of coordinates.
9. Normalization: A measurement unit for the amount of uncertainty must be chosen by assigning a positive real number to a particular uncertainty function *u* in the given theory.

UNIQUENESS OF FUNCTIONAL U

- For any given uncertainty theory T, functional U is required to be unique under all the requirements formulated in the calculus of theory T.
- The normalization requirement may be formulated in different ways in each theory T. Each possible formulation defines a particular measurement unit of uncertainty when uncertainty is measured by functional U in theory T.
- U is an abstract measuring instrument: a particular function u in theory T is an input, a real number measuring the amount of uncertainty captured by u is the corresponding output.

Basic Equations and Inequalities of Uncertainty Measures U *on* $X \times Y$

- $U(X \mid Y) = U(X \times Y) - U(Y)$
- $U(Y \mid X) = U(X \times Y) - U(X)$
- $T_U(X,Y) = U(X) + U(Y) - U(X \times Y)$
- $U(X \times Y) \leq U(X) + U(Y)$
- $U(X \mid Y) \leq U(X)$ and $U(Y \mid X) \leq U(Y)$
- These equations and inequalities are valid in all theories of uncertainty.

IMPRECISSION IN PROBABILITIES: Why we need it ?

- It allows us to account for the amount of statistical information upon which probabilities are based.
- It allows us to properly represent total ignorance.
- It makes it easier to assess or elicit probabilities.
- It allows us to account for inconsistencies among several sources of information.
- It allows us to determine probabilities approximately under time or computational constraints.
- It allows us to properly represent probabilities outside a given class for which precise probabilities are known.

LIMITATIONS OF PROBABILITY THEORY

I think it wiser to avoid the use of a probability model when we do not have the necessary data than to fill the gaps arbitrarily; arbitrary assumptions yield arbitrary conclusions.

(Terrence L. Fine, *Theories of Probability*, 1973, p. 177)

MONOTONE MEASURES - 1

Given a universal set X and a nonempty class **C** of subsets of X, a monotone measure is a set function

$$\mu\colon \mathbf{C} \to [0, \infty]$$

that satisfies the following requirements 1 & 2 (for finite X) or requirements 1-4 (for infinite X):

1. <u>Boundary requirements</u>: $\mu(\varnothing) = 0$ and, for normalized measures, also $\mu(X) = 1$.
2. <u>Monotonicity</u>: For all $A,B \in \mathbf{C}$, if $A \subseteq B$, then

$$\mu(A) \leq \mu(B).$$

MONOTONE MEASURES - 2

3. Continuity from below: For any increasing sequence $A_1, A_2, \cdots$ of sets in $\mathbf{C}$ if $\bigcup_{i=1}^{n} A_i \in \mathbf{C}$, then $\lim_{i\to\infty} \mu(A_i) = \mu\left(\bigcup_{i=1}^{\infty} A_i\right)$.

4. Continuity from above: For any decreasing sequence $A_1, A_2, \cdots$ of sets in $\mathbf{C}$, if $\bigcap_{i=1}^{\infty} A_i \in \mathbf{C}$, then $\lim_{i\to\infty} \mu(A_i) = \mu\left(\bigcap_{i=1}^{\infty} A_i\right)$.

CHOQUET CAPACITIES OF ORDER k (k-monotone measures)

- 2-monotone measures: for all $A,B \subseteq X$,
  $\mu_*(A \cup B) \geq \mu_*(A) + \mu_*(B) - \mu_*(A \cap B)$.

- 3-monotone measures: for all $A,B,C \subseteq X$,
  $\mu_*(A \cup B \cup C) \geq \mu_*(A) + \mu_*(B) + \mu_*(C)$
  $- \mu_*(A \cap B) - \mu_*(A \cap C) - \mu_*(B \cap C) +$
  $\mu_*(A \cap B \cap C)$.

k-MONOTONE MEASURES

- k-monotone measures ($k \geq 2$) are defined for all families of k subsets of X by the inequality

$$\mu_*\left(\bigcup_{j=1}^{k} A_j\right) \geq \sum_{\varnothing \neq K \subseteq N_k} (-1)^{|K+1|} \mu_*\left(\bigcap_{j \in K} A_j\right),\ N_k = \{1,2,...,k\}.$$

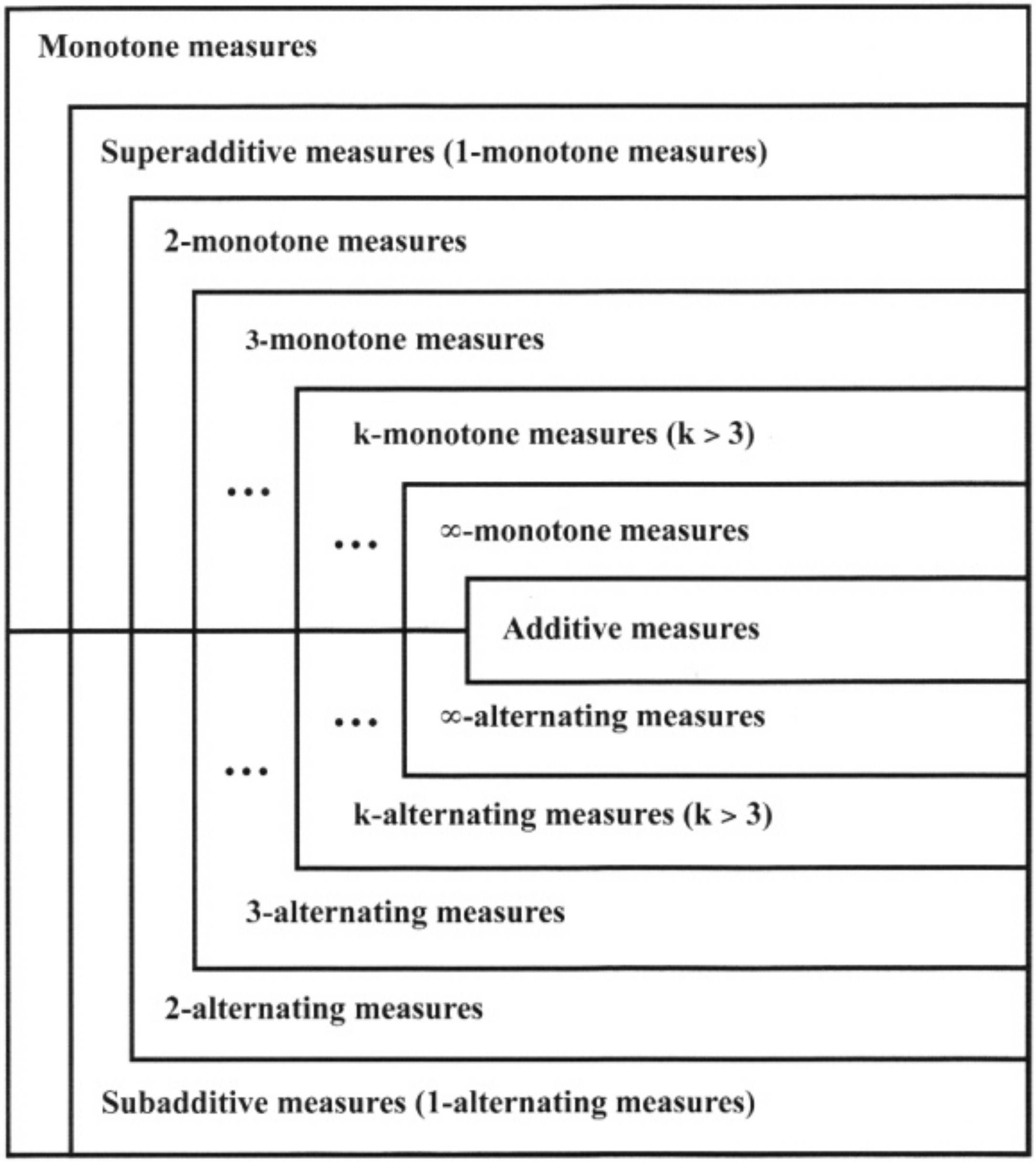
- 1-monotone measures: a convenient name for measures that satisfy for all pairs of sets $A,B \subseteq X$ the inequality

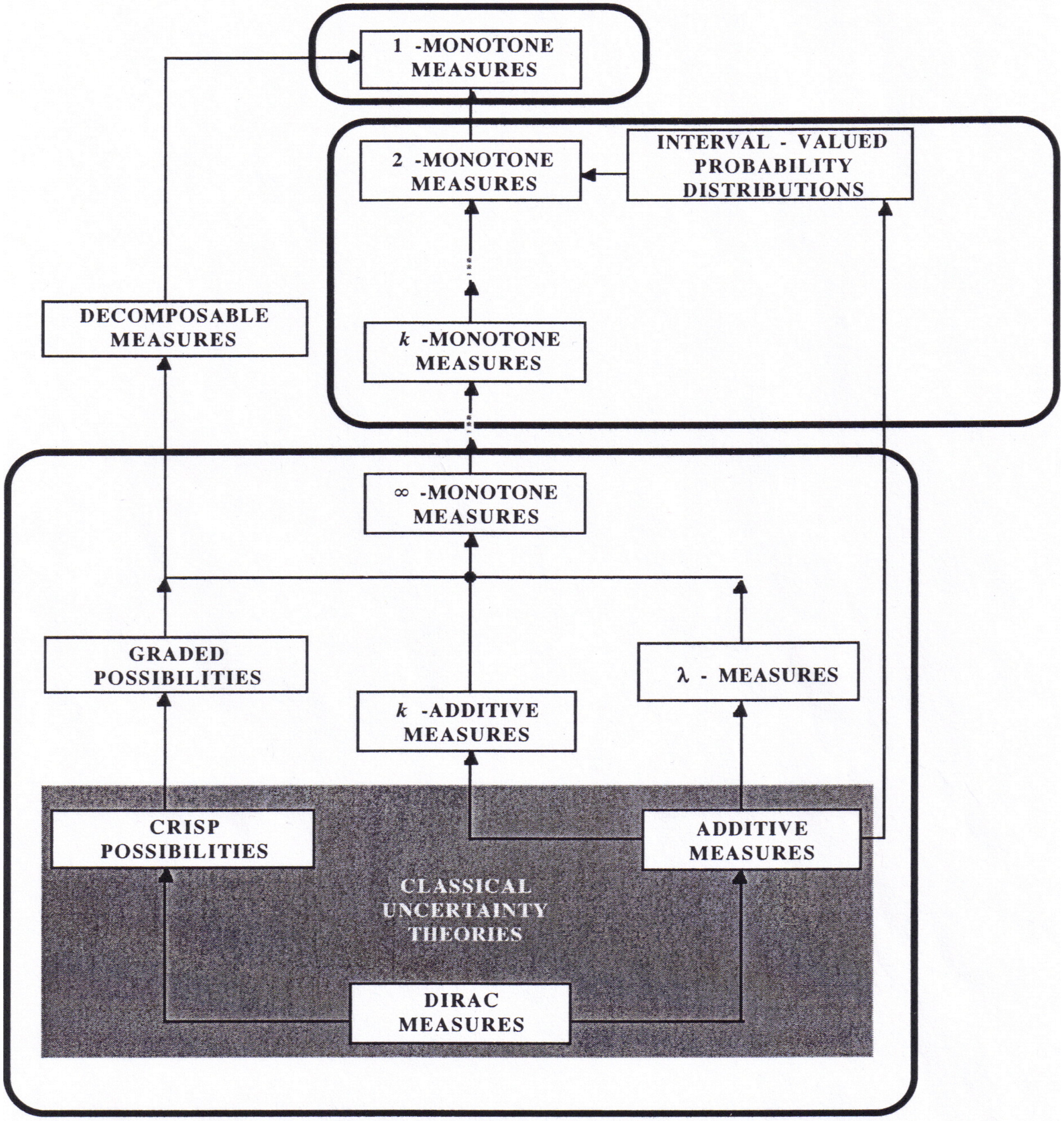
$$\mu_*(A \cup B) \geq \mu_*(A) + \mu_*(B).$$

ALTERNATING CHOQUET CAPACITIES OF ORDER *k*

Defined by the inequality:

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


Monotone measures
Superadditive measures (1-monotone measures)
2-monotone measures
3-monotone measures
k-monotone measures (k > 3)
...
...
∞-monotone measures
Additive measures
...
∞-alternating measures
...
k-alternating measures (k > 3)
3-alternating measures
2-alternating measures
Subadditive measures (1-alternating measures)


1 -MONOTONE MEASURES
2 -MONOTONE MEASURES
INTERVAL - VALUED PROBABILITY DISTRIBUTIONS
DECOMPOSABLE MEASURES
k -MONOTONE MEASURES
∞ -MONOTONE MEASURES
GRADED POSSIBILITIES
k -ADDITIVE MEASURES
λ - MEASURES
CRISP POSSIBILITIES
ADDITIVE MEASURES
CLASSICAL UNCERTAINTY THEORIES
DIRAC MEASURES

IMPRECISE PROBABILITIES: Canonical Representations

1. Lower probability functions: μ_*
2. Upper probability functions: μ^*
3. Möbius representations: m
4. Convex sets of probability distributions (also called credal sets): $\mathbf{P}$

LOWER PROBABILITY FUNCTIONS

- These functions are always monotone measures that are also superadditive: $\mu_*(A \cup B) \geq \mu_*(A) + \mu_*(B)$ whenever A, B, $A \cup B \in \mathbf{C}$ and $A \cap B = \varnothing$.

- For singletons of X, these functions always satisfy the inequality

$$\sum_{x \in X} \mu_*(\{x\}) \leq 1.$$

UPPER PROBABILITY FUNCTIONS

- These functions are always monotone measures that are subadditive: $\mu^*(A \cup B) \leq \mu^*(A) + \mu^*(B)$ whenever A, B, $A \cup B \in \mathbf{C}$.

- For singletons of X, these functions also satisfy the inequality

$$\sum_{x \in X} \mu^*(\{x\}) \geq 1.$$

MÖBIUS REPRESENTATIONS

- Möbius representations are set functions, m, that satisfy the following two requirements:

$$M(\varnothing) = 0; \text{ and}$$

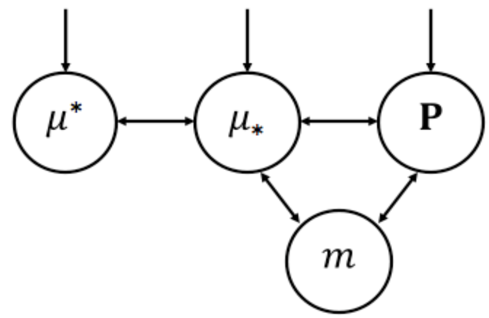
$$\sum_{A \subseteq X} m(A) = 1.$$

CONVEX SETS OF PROBABILITY DISTRIBUTION (CREDAL SETS)

- Each credal set **P** is expressed in terms of a convex polytope with a finite number, *r*, of extreme points: $\{p_i = (p_{i1}, p_{i2}, \ldots, p_{in}): i \in N_r\}$.

- **P** is the convex hull of the extreme points:

$$\mathbf{P} = \left\{ \left(\sum_{i=1}^{r} p_{ik}\lambda_i : k \in N_n \right) : \lambda_i \in [0,1], i \in N_r, \sum_{i=1}^{r} \lambda_i = 1 \right\}.$$



μ_*: lower probability

μ^*: upper probability (dual of μ_*)

P : closed convex set of probability distributions

m : Möbius representation

CONVERSIONS - 1

- From **P** to μ_* (many-to-one functions):

$$\mu_*(A) = \inf_{p \in \mathbf{P}} \left\{ \sum_{x \in A} p(x) \right\}, \forall A$$

- From **P** to μ^* (many-to-one functions):

$$\mu^*(A) = \sup_{p \in \mathbf{P}} \left\{ \sum_{x \in A} p(x) \right\}, \forall A$$

- From μ_* to μ^* and vice versa by their duality:

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

CONVERSIONS - 2

- From μ_* to m via Möbius transform:

$$m(A) = \sum_{B \subseteq A} (-1)^{|A-B|} \mu_*(B), \forall A$$

- From m to μ_* via inverse Möbius transform:

$$\mu_*(A) = \sum_{B \subseteq A} m(B), \forall A$$

CONVERSIONS - 3

- From μ_* to **P**:

$$\mathbf{P}(\mu_*) = \left\{ p : \mu_*\left(A\right) \leq \sum\nolimits_{x \in A} p(x), \forall A \right\}.$$

- From μ^* to **P**:

$$\mathbf{P}(\mu^*) = \left\{ p : \mu^*\left(A\right) \geq \sum\nolimits_{x \in A} p(x), \forall A \right\}.$$

CONVERSIONS - 4

- Let $X = \{x_1, x_2, \ldots, x_n\}$ and let $(\pi(x_1), \pi(x_2), \ldots, \pi(x_n))$ denote a permutation by which elements of X are reordered.
- Given any lower probability measure μ_* on the power set of X that is at least 2-monotone, the convex polytope of all probability distributions that dominate μ_*, $\mathbf{P}(\mu_*)$, is determined by its vertices (extreme points), which are probability distributions p_π computed as follows:

CONVERSIONS - 5

- Probability distribution for permutation π:

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

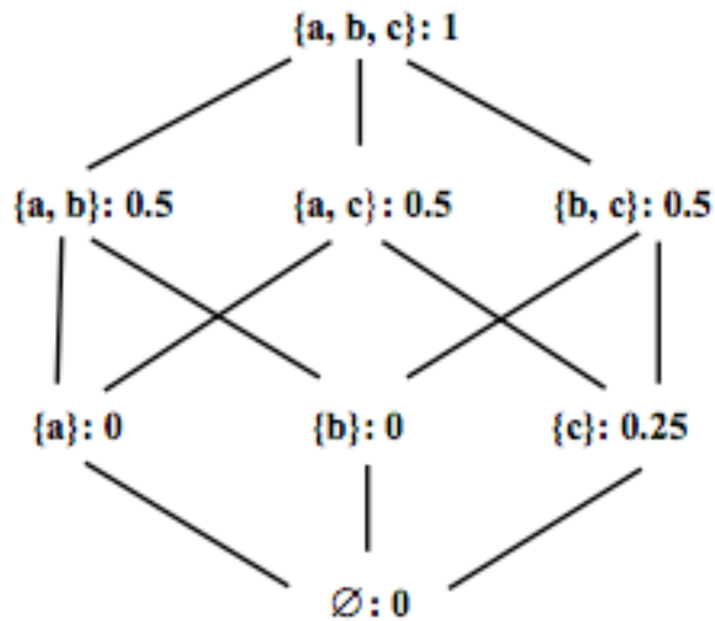
$$p_\pi(\pi(x_2)) = \mu_*(\{\pi(x_1), \pi(x_2)\}) - \mu_*(\{\pi(x_1)\}),$$

$$p_\pi(\pi(x_3)) = \mu_*(\{\pi(x_1), \pi(x_2), \pi(x_3)\}) - \mu_*(\{\pi(x_1), \pi(x_2)\}),$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

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

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


{a, b, c}: 1
{a, b}: 0.5
{a, c}: 0.5
{b, c}: 0.5
{a}: 0
{b}: 0
{c}: 0.25
∅: 0

THEORY OF REACHABLE PROBABILITY DISTRIBUTION INTERVALS

- Lower and upper probability functions are determined by intervals $[\underline{p}(x), \overline{p}(x)]$ of probabilities on singletons.
- Given a set of reachable (nonredundant) probability intervals for all $x \in X$, I, the associated convex set **P** of probability distribution functions p on X is given by:

$$\mathbf{P}(I) = \left\{ p(x) : x \in X, p(x) \in \left[\underline{p}(x), \overline{p}(x) \right], \sum_{x \in X} p(x) = 1 \right\}.$$

- **P**(I) are ($n-1$)-dimensional polytopes, where $n = |X|$, whose number c of extreme points is bounded:

$$n \leq c \leq n(n-1)$$

Hartley functional for measuring classical possibilistic uncertainty, which is usually referred to as nonspecificity, and the associated uncertainty-based information:

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

$$I_H(A) = \log_2 |X| - \log_2 |A|$$

HARTLEY-LIKE MEASURE IN n-DIMENSIONAL EUCLIDEAN SPACE

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

- A is a convex subset of $\mathbf{R}^n$.
- μ denotes the Lebesgue measure.
- T denotes the set of all isometric transformations from one orhogonal coordinate system to another.
- The i-th projection of A within the coordinate system t is denoted by A_{i_t}

Shannon functional for measuring classical probabilistic uncertainty (conflict) and the associated uncertainty-based information:

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

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

$$I_S(p(X) : x \in X) = \log_2 |X| - S(p(x) : x \in X)$$

GENERALIZED UNCERTAINTY MEASURES: A Historical Overview

- Generalization of the Hartley measure to graded possibilities (Higashi and Klir, 1983):

$$GH(r(x) : x \in X) = \int_0^1 \log_2 |A_\alpha| \, d\alpha, \text{ where}$$

$$A_\alpha = \{x \in X : r(x) \geq \alpha\}, \; r(x) \in [0,1], \; \max_{x \in X} r(x) = 1.$$

- Uniqueness proof in (Klir and Mariano, 1987).

From *H* to *GH* in Dempster-Shafer Theory

- Further generalization to the Dempster-Shafer theory on finite sets (Dubois and Prade, 1985):

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

$$m(A) \in [0,1] \text{ for all } A \subseteq X,$$

$$m(\varnothing) = 0 \text{ and } \sum_{A \subseteq X} m(A) = 1.$$

- Uniqueness proof in (Ramer, 1987).

The issue of generalizing the Shannon entropy

- Recognizing that the Shannon entropy measures the conflict among evidential claims in a probability distribution, at least six functionals were proposed in 1982-1990 as generalized Shannon entropies in the Dempster-Shafer theory.

- Each of the proposed functionals was eventually found to violate the essential property of subadditivity.

- The unsuccessful attempts to generalize the Shannon entropy are outlined in (Klir, 2006).

Total Aggregated Uncertainty Measure in DST

- A total measure of uncertainty in DST that aggregates nonspecificity and conflict and satisfies all essential requirements was found in the early 1990s by several groups (Harmanec and Klir, 1994):

$$AU(Bel) = \max_{p(x) \in \mathbf{P}_{Bel}} \left\{ -\sum_{x \in X} p(x) \log_2 p(x) \right\},$$

where $\mathbf{P}_{Bel}$ is the set of distributions p that dominate Bel:

$$Bel(A) \leq \sum_{x \in A} p(x) \text{ for all } A \subseteq X.$$

Aggregate Measure of Total Uncertainty for all Types of Imprecise Probabilities

$$\overline{S}(\mathbf{P}) = \max_{p \in \mathbf{P}} \{S(p)\}$$

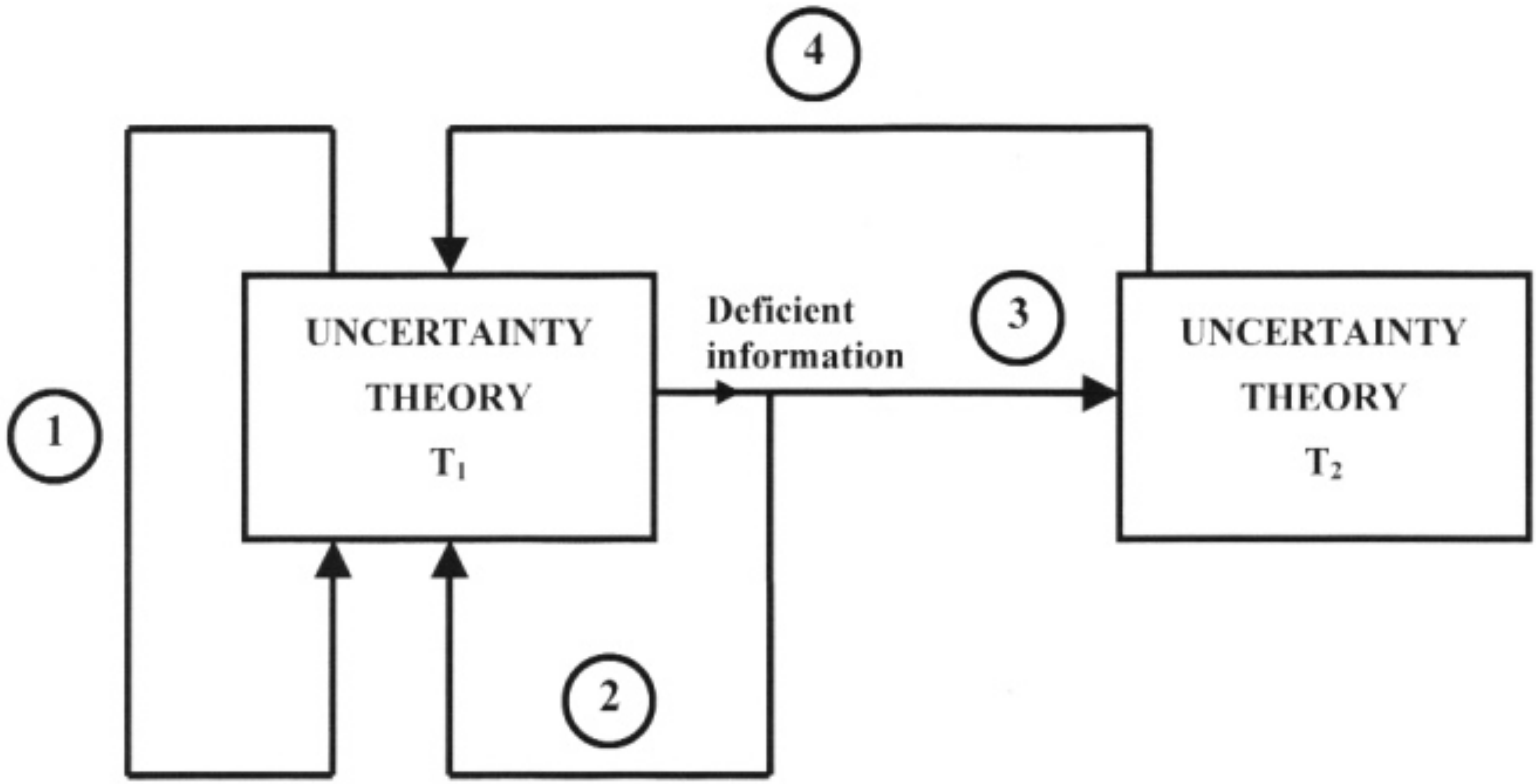
Disaggregated Measure of Total Uncertainty for All Types of Imprecise Probabilities

$$TU(\mathbf{P}) = \left\langle \underline{S}(\mathbf{P}), \overline{S}(\mathbf{P}) - \underline{S}(\mathbf{P}) \right\rangle$$

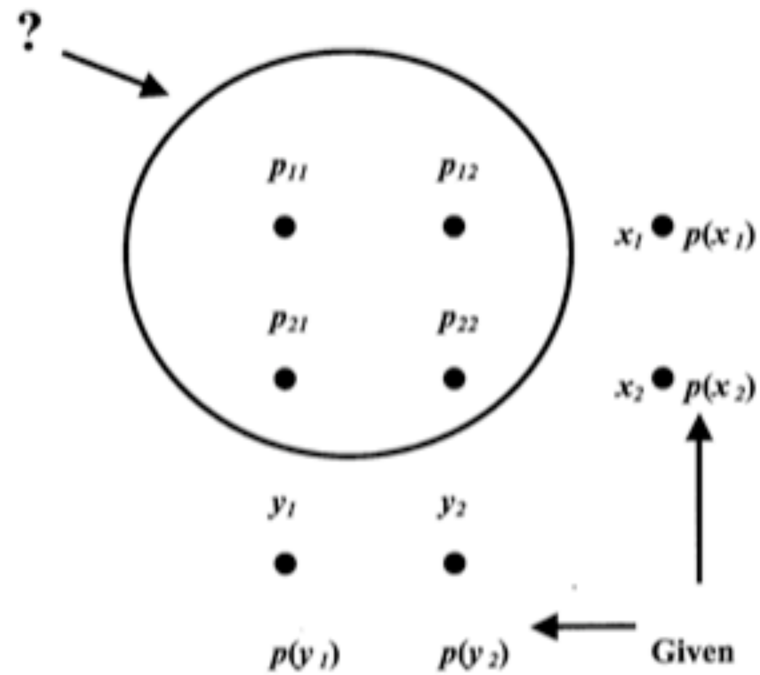
$$\underline{S}(\mathbf{P}) = \min_{p \in \mathbf{P}} \{S(p)\}$$

PRINCIPLES OF UNCERTAINTY BASED ON RELEVANT UNCERTAINTY MEASURE *U*

- Principle of minimum uncertainty
- Principle of maximum uncertainty
- Principle of uncertainty invariance
- Principle of requisite generalization



4
UNCERTAINTY
THEORY
T_1
Deficient
information
3
UNCERTAINTY
THEORY
T_2
1
2



Example: $p(x_1) = 0.8$, $p(x_2) = 0.2$
$p(y_1) = 0.6$, $p(y_2) = 0.4$

PRINCIPLE OF REQUISITE GENERALIZATION

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

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

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

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

VIEWING UNCERTAINY PRINCIPLES AS OPERATIONALIZED WISDOM PRINCIPLES:
Some Relevant Quotations

Knowing ignorance is strength.

Ignoring knowledge is sickness.

(Lao Tsu, *Tao Te Ching*, 6th century B.C.

There is nothing better to know that you don' t know.

Not knowing yet thinking you know -- this is sickness.

(An alternative translation)

Ignorance is preferable to the illusion of knowledge.

(Thomas Sowell)

UNCERTAINTY IN DAILY LIFE

Whenever you find yourself getting angry about a difference in opinion, be on your guard; you will probably find, on examination, that your belief is getting beyond what the evidence warrants.

(Bertrand Russell
Unpopular Essays)

UNCERTAINTY IN SCIENCE

At a fundamental level, scientific uncertainty begins when we make measurements. … The very term 'experiment' implies uncertainty.

(Henry N. Pollack, *Uncertain Science … Uncertain World*, 2003, p. 63)

SOME OPEN PROBLEMS

- To develop uncertainty theories that are based on decomposable measures.
- To fuzzify those uncertainty theories that have not been fuzzified as yet.
- Is the functional *GH* subadditive for imprecise probabilities based on two-monotone measures?
- How can negative values of *m* be interpreted?
- What are meaningful and useful disaggregations of the total aggregated uncertainty S^*?
- Unresolved problem in (Bronevich and Klir, 2010).