Conditional probability and fuzzy information

Giulianella Coletti\textsuperscript{a}, Romano Scozzafava\textsuperscript{b,*}

\textsuperscript{a}Dipartimento Matematica e Informatica, Università di Perugia, Via Vanvitelli 1, 06123 Perugia, Italy
\textsuperscript{b}Dipartimento Metodi e Modelli Matematici, Università di Roma “La Sapienza”, via Scarpa 16, 00161 Roma, Italy

Available online 15 May 2006

Abstract

The main subject of this paper is the embedding of fuzzy set theory—and related concepts—in a coherent conditional probability scenario. This allows to deal with perception-based information—in the sense of Zadeh—and with a rigorous treatment of the concept of likelihood, dealing also with its role in statistical inference. A coherent conditional probability is looked on as a general non-additive “uncertainty” measure $m(\cdot) = P(E|\cdot)$ of the conditioning events. This gives rise to a clear, precise and rigorous mathematical frame, which allows to define fuzzy subsets and to introduce in a very natural way the counterparts of the basic continuous $T$-norms and the corresponding dual $T$-conorms, bound to the former by coherence. Also the ensuing connections of this approach to possibility theory and to information measures are recalled.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Coherent conditional probability; Likelihood function; Fuzzy theory; Perception-based information

1. Introduction

Many papers, starting with the pioneering one by Zadeh (1965), have been devoted—during past years—to support the negative view maintaining that probability is inadequate to capture what is usually treated by fuzzy theory. This view is still supported by Zadeh in recent years: in particular we mention the paper Zadeh (2002), which contains also a number of relevant references. His thesis is that PT—standard probability theory—is not fit to offer solutions for many simple problems in which a key role is played by a “perception-based information”.

We agree with Zadeh’s position, inasmuch he specifies that by PT he means “standard probability theory of the kind found in textbooks and taught in courses”. Actually, many traditional aspects of probability theory are not so essential as they are usually considered; for example, the requirement that the set of all possible “outcomes” should be endowed with a beforehand given algebraic structure—such as a Boolean algebra or $\sigma$-algebra—or the aim at getting, for these outcomes, uniqueness of their probability values, with the ensuing introduction of suitable relevant assumptions—such as $\sigma$-additivity, maximum entropy, conditional independence, . . .—or interpretations, such as a strict frequentist one, which unnecessarily restricts the domain of applicability.

This paper is based on the approach to probability expounded, e.g., in Coletti and Scozzafava (1996, 1999b): see also the book Coletti and Scozzafava (2002a). The leading tool is that of coherence, a concept that goes back to de Finetti (1970). Our starting point is a synthesis of the available information, expressed by one or more events: to this purpose,
the concept of event must be given its more general meaning, i.e. it must not be looked on just as a possible outcome—a subset of the so-called “sample space”, as it is usually done in PT—but expressed by a proposition. Moreover, events play a twofold role, since we must consider not only those events which are the direct object of study, but also those which represent the relevant “state of information”: in fact a bunch of conditional events, together with a relevant “partial” assessment of conditional probability, are the tools that allow to manage specific—conditional—situations and to update degrees of belief on the basis of the evidence. The role of coherence is in fact that of ruling this extension process; a similar theory—but only for unconditional events—is the probabilistic logic by Nilsson (1986), which is just a re-phrasing—with different terminology—of de Finetti’s theory, as Nilsson (1993) himself acknowledges.

The main points of our approach are recalled and summarized in the Appendix.

Notice that what is usually emphasized in the relevant literature—when a conditional probability \( P(E|H) \) is taken into account—is only the fact that \( P(E|H) \) is a probability for any given \( H \): this is a very restrictive—and misleading—view of conditional probability, corresponding trivially to just a modification of the “sample space” \( \Omega \).

It is instead essential to regard also the conditioning event \( H \) as a “variable”, i.e. the “status” of \( H \) in \( E|H \) is not just that of something representing a given fact, but that of an uncertain event—like \( E \)—for which the knowledge of its truth value is not required. In other words, even if beliefs may come from various sources, they can be treated in the same way, since the relevant conditioning events—including both statistical data and any perception-based information—can always be considered as being assumed propositions: this means, using a terminology due to Koopman (1940), that \( H \) must be looked on—even if asserted—as being contemplated.

The concept of conditional event plays a central role for this approach to probabilistic reasoning. A conditional event \( E|H \), with \( H \neq \emptyset \) (the impossible event), is a logical entity which is true when both \( E \) and \( H \) are true, false when \( H \) is true and \( E \) is false, while, when \( H \) is false, it is a function of the ordered pair \((E, H)\) which measures the degree of belief in the conditional event \( E|H \). Under natural conditions it turns out, as explained in detail in Coletti and Scozzafava (1999b, 2001b), that this function can be the conditional probability \( P(E|H) \) in its most general sense, i.e. satisfying the classic axioms as given by de Finetti (1949) and Popper (1959). Due to its direct assignment as a whole, the knowledge—or the assessment—of the “joint” and “marginal” unconditional probabilities \( P(E \land H) \) and \( P(H) \) is not required; moreover, the conditioning event \( H \)—which must be a possible event—may have zero probability. So, conditioning in a coherent setting gives rise to a general scenario that makes the classic Radon–Nikodym procedure—and the relevant concept of regularity—neither necessary nor significant. This has been repeatedly discussed elsewhere: see, e.g., Coletti and Scozzafava (2005).

Let us also mention, among the problems that a non-traditional PT is able to manage, those situations constituting a broader class than that described by a numerical probability: we refer to the so-called comparative probability, which is an ordinal relation “less than or equal” on a set \( \mathcal{E} \) of conditional events: \( A \land H \preceq B \land K \) can be read as \( A \land H \) “is no more probable than” \( B \land K \). In this context one is interested to find relations compatible with a numerical probability: we define coherent a comparative probability if there exists a coherent numerical conditional probability \( P: \mathcal{E} \rightarrow [0, 1] \) representing it, i.e. such that

\[ A \land H \preceq B \land K \quad \text{if and only if} \quad P(A|H) \leq P(B|K). \]

Since the procedure to test coherence—see, e.g., Coletti (1994)—is the same as in the numerical case, i.e. finding a set of probability distributions verifying some equalities or inequalities, we can consider also “combined” comparative and numerical evaluations. So in our framework it is possible to put together different kinds of partial knowledge, as shown in the paradigmatic example reported in the last section.

Concerning more specifically fuzzy theory, we refer to the state of information—at a given moment—of a real (or fictitious) person that will be denoted by “You”. If \( X \) is a—not necessarily numerical—random quantity with range \( CX \), let \( A_x \), for any \( x \in CX \), be the event \( \{X = x\} \). The family \( \{A_x\}_{x \in CX} \) is obviously a partition of the certain event \( \Omega = CX \). Now, let \( \varphi_X \) be any property related to the random quantity \( X \): from a pragmatic point of view, it is natural to think that You have some information about possible values of \( X \), which allows You to refer to a suitable membership function of the fuzzy subset of “elements of \( CX \) with the property \( \varphi_X \)”.

For example, if \( X \) is a numerical quantity and \( \varphi_X \) is the property “small”, for You the membership function \( \mu(x) \) may be put equal to 1 for values \( x \) of \( X \) less than a given \( x_1 \), while it is put equal to 0 for values greater than a suitable \( x_2 > x_1 \); then it is taken as decreasing from 1 to 0 in the interval from \( x_1 \) to \( x_2 \): this choice of the membership function implies that, for You, elements of \( CX \) less than \( x_1 \) have the property \( \varphi_X \), while those greater than \( x_2 \) do not.
So the real problem is that you are doubtful on having or not the property \( \varphi_X \) those elements of \( C_X \) between \( x_1 \) and \( x_2 \). Then the interest is in fact directed toward conditional events such as \( E|A_x \), where \( x \) ranges over the interval from \( x_1 \) to \( x_2 \), with

\[
E = \{ \text{You claim (that X has) the property } \varphi_X \}, \quad A_x = \{ \text{the value of X is } x \}.
\]

It follows that you may assign a subjective probability \( P(E|A_x) \) equal, e.g., to 0.2 without assigning a degree of belief of 0.8 to the event \( E \) under the assumption \( A_x^c \)—i.e., the value of \( X \) is not \( x \)—since an additivity rule with respect to the conditioning events does not hold. In other words, it seems sensible to identify the values of the membership function \( \mu(x) \) with suitable conditional probabilities. In particular, putting

\[
H_0 = \{ X \text{ is greater than } x_2 \}, \quad H_1 = \{ X \text{ is less than } x_1 \},
\]

one has that \( E \) and \( H_0 \) are incompatible and that \( H_1 \) implies \( E \), so that, by the rules of a conditional probability, \( P(E|H_0) = 0 \) and \( P(E|H_1) = 1 \).

Notice that this conditional probability \( P(E|A_x) \) is directly introduced as a function on the set of conditional events, and without assuming any given algebraic structure. Is that possible? In the usual (Kolmorogovian) approach to conditional probability the answer is NO, since the introduction of \( P(E|A_x) \) would require the consideration and the assessment of \( P(E \land A_x) \) and \( P(A_x) \), assuming positivity of the latter. But this could not be a simple task: in fact in this context the only sensible procedure is to assign directly \( P(E|A_x) \). For example, it is possible to assign the conditional probability that “You claim this number is small” knowing its value \( x \), but not necessarily the probability that “The number has the value \( x \)”; not to mention that, for many choices of the random quantity \( X \), the corresponding probability can be zero. These problems are easily by-passed in our framework (cf. Appendix).

For the formal definitions concerning fuzzy sets, see Section 2, where we recall our results expounded in a series of papers: see, e.g., Coletti and Scozzafava (1999a, 2001a, 2004); a preliminary sketch of the interpretation of the membership function in terms of conditional probability is in Scozzafava (1993). We show not only how to define fuzzy subsets, but we also introduce in a very natural way the counterparts of the basic continuous \( T \)-norms and the corresponding dual \( T \)-conorms, bound to the former by coherence.

In conclusion, our theory is not a probabilistic—in the usual traditional sense—interpretation of fuzziness, since a conditional probability is not a probability, except in the trivial case in which the conditioning event is fixed and we let the first one vary: notice that we are proceeding the other way round!.

In Section 3 we discuss, from our point of view, the examples introduced by Zadeh to challenge standard PT.

As an interesting and fundamental by-product of our approach, it comes out (see Section 4) a natural interpretation of possibility theory, both from a semantic and a syntactic point of view. Deepening the study of the properties of a coherent conditional probability looked on as a general non-additive uncertainty measure of the conditioning events, we proved—see Coletti and Scozzafava (2003a)—that this measure is a capacity if and only if it is a possibility. Furthermore—see Coletti and Scozzafava (2003b)—we have been able to look on a coherent conditional probability as a suitable information measure.

Roughly speaking, looking on a coherent conditional probability as a general non-additive “uncertainty” measure \( m(\cdot) = P(E|\cdot) \) of the conditioning events amounts to entering what in the statistical jargon could be called the “likelihood world”, with its various “ad hoc” extensions from a point function to a set function. The question is: without a clear, precise and rigorous mathematical frame, is the likelihood “per se” a proper tool to deal with statistical inference and to manage partial and vague information? These aspects are discussed in Section 4.3.

In the relevant literature there are several papers containing approaches which are seemingly similar to that expounded here. So in the introduction to our paper Coletti and Scozzafava (2004) we mentioned some related references where a frame in which “tout se tient” is lacking. In a recent paper by Booker and Singpurwalla (2004)—in which an event is, as usual, a subset of the sample space and not a proposition, so that there is no room for a distinction between a property \( \varphi \) related to a fuzzy subset and the relevant event “You claim \( \varphi \)”—the likelihood has its usual intuitive and empirical meaning. It is not looked on as the restriction of a suitable—coherent—conditional probability: it turns out that such an approach does not allow a complete re-reading of fuzzy theory and its embedding in a more general framework. This embedding makes it possible to introduce operations through norms and conorms, to deal with possibility and information measures, to manage perception-based information, etc.
2. Membership function as coherent conditional probability

Let \( \varphi_X \) be any property—in the sequel, to simplify notation we will write simply \( \varphi \) in place of \( \varphi_X \)—related to the random quantity \( X \); notice that a property, even if expressed by a statement, does not single-out an event, since the latter needs to be expressed by a non-ambiguous proposition that can be either true or false. Consider now the event \( E_\varphi = \{ \text{You claim } \varphi \} \) and a coherent conditional probability \( P ( E_\varphi | A_X ) \), looked on as a real function \( \mu_\varphi (x) = P ( E_\varphi | A_X ) \) defined on \( C_X \).

Since the events \( A_x \) are incompatible, then—by Corollary A1, see the Appendix—every \( \mu_\varphi (x) \) with values in \([0, 1]\) is a coherent conditional probability. So we can define a fuzzy subset in the following way.

**Definition 1.** Given a random quantity \( X \) with range \( C_X \) and a related property \( \varphi \), a fuzzy subset \( E_\varphi^* \) of \( C_X \) is the pair

\[
E_\varphi^* = \{ E_\varphi, \mu_\varphi \},
\]

with \( \mu_\varphi (x) = P ( E_\varphi | A_X ) \) for every \( x \in C_X \).

So a coherent conditional probability \( P ( E_\varphi | A_X ) \) is clearly a measure of how much You, given the event \( A_x = \{ X = x \} \), are willing to claim the property \( \varphi \), and it plays the role of the membership function of the fuzzy subset \( E_\varphi^* \).

Notice also that the significance of the conditional event \( E_\varphi | A_X \) is reinforced by looking on it as “a whole”, avoiding a separate consideration of the two propositions \( E_\varphi \) and \( A_x \).

A fuzzy subset \( E_\varphi^* \) is a crisp set when the only coherent assessment \( \mu_\varphi (x) = P ( E_\varphi | A_X ) \) has range \([0, 1]\). Then, by Corollary A2 (see Appendix), a fuzzy subset \( E_\varphi^* \) is a crisp set when the property \( \varphi \) is such that, for every \( x \in C_X \), either \( E_\varphi \land A_x = \emptyset \) or \( A_x \subseteq E_\varphi \).

**Remark 1.** Let us emphasize that in our context the concept of fuzzy event, as introduced by Zadeh (1968), is nothing else than a proposition, i.e., an ordinary event, of the kind “You claim the property \( \varphi \)”. So, according to the rules of conditional probability—in particular, the “disintegration” formula, often called in the relevant literature “theorem of total probability”—we can easily compute its probability as

\[
P ( E_\varphi ) = \sum_x P ( A_x ) P ( E_\varphi | A_X ) = \sum_x P ( A_x ) \mu_\varphi (x),
\]

which coincides with Zadeh’s definition of the probability of (what he calls) a “fuzzy” event.

Notice that this result is not only a trivial consequence of probability rules and not a definition, but puts also under the right perspective the subjective nature of a membership function, showing once again that our approach to probability goes beyond—both syntactically and semantically—the traditional one, denoted PT by Zadeh.

Consider now two fuzzy subsets \( E_\varphi^*, E_\psi^* \), corresponding to the random quantities \( X \) and \( Y \) (possibly \( X = Y \)), and assume that, for every \( x \in C_X \) and \( y \in C_Y \), both the following equalities hold:

\[
P ( E_\varphi | A_X \land A_Y ) = P ( E_\varphi | A_X ), \quad P ( E_\psi | A_X \land A_Y ) = P ( E_\psi | A_Y ),
\]

with \( A_y = \{ Y = y \} \).

**Remark 2.** We refer to fuzzy subsets of two different spaces \( X \) and \( Y \). This will allow to introduce the operations of union and intersection in the product space \( X \times Y \), so that we can deal also with those situations in which we need to consider jointly (for a given population) properties such as “tall” and “blond”.

The binary operations of union and intersection and that of complementation are defined as follows:

\[
E_\varphi^* \cup E_\psi^* = \{ E_{\varphi \lor \psi}, \mu_{\varphi \lor \psi} \}, \quad E_\varphi^* \cap E_\psi^* = \{ E_{\varphi \land \psi}, \mu_{\varphi \land \psi} \}, \quad ( E_\varphi^* )^c = \{ E_{\neg \varphi}, \mu_{\neg \varphi} \}.
\]
where—by a fairly improper notation—\( \varphi \lor \psi \), \( \varphi \land \psi \) denote, respectively, the properties “\( \varphi \) or \( \psi \)”, “\( \varphi \) and \( \psi \)”, and \( E_{\varphi \lor \psi} = E_\varphi \lor E_\psi \), \( E_{\varphi \land \psi} = E_\varphi \land E_\psi \), while \( \mu_{\varphi \lor \psi} \) and \( \mu_{\varphi \land \psi} \) are defined on \( C_{XY} \subseteq C_X \times C_Y \) by putting

\[
\mu_{\varphi \lor \psi}(x, y) = P \left( E_\varphi \lor E_\psi \mid A_x \land A_y \right), \quad \mu_{\varphi \land \psi}(x, y) = P \left( E_\varphi \land E_\psi \mid A_x \land A_y \right).
\]

The conditional event \( \left( E_\varphi \lor E_\psi \right) \mid \left( A_x \land A_y \right) \) is true iff \( A_x \land A_y \) and \( E_\varphi \lor E_\psi \) are both true: and the latter event is true, by definition of union, when at least one of the two events is true, that is when “You claim \( \varphi \)” or when “You claim \( \psi \)”. On the other hand, \( E_{\varphi \lor \psi} \) is true when “You claim \( \varphi \) or \( \psi \)”, and this requires to put \( E_{\varphi \lor \psi} = E_\varphi \lor E_\psi \). Similar considerations apply to the events \( E_{\varphi \land \psi} \) and \( E_\varphi \land E_\psi \).

Notice also the following relation: \( E_{\neg \varphi} \neq \left( E_\varphi \right)^c \) —where \( \left( E_\varphi \right)^c \) denotes the contrary of the event \( E_\varphi \)—while the equality holds only for a crisp set; for example, the propositions “You claim not young” and “You do not claim young” are logically independent. Then, while \( E_\varphi \lor \left( E_\varphi \right)^c = C_X \), we have instead \( E_\varphi \lor E_{\neg \varphi} \subseteq C_X \).

We could also introduce the tautological property \( T = \varphi \lor \neg \varphi \)—for any \( \varphi \)—which satisfies (trivially) the relation \( E_T \subseteq \Omega \), and the void property \( V = \varphi \land \neg \varphi \)—for any \( \varphi \)—which satisfies the relation \( E_V \neq \emptyset \). Therefore, if we consider the union of a fuzzy subset and its complement

\[
E_\varphi^* \cup \left( E_\varphi \right)^c = \left\{ E_\varphi \lor \neg \varphi, \ \mu_{\varphi \lor \neg \varphi} \right\}
\]

we obtain in general a fuzzy subset of (the universe) \( C_X \). On the other hand, it is easy to check that the complement of a crisp set is also a crisp set: in fact, from \( E_\varphi \land A_x = \emptyset \) it follows \( A_x \subseteq \left( E_\varphi \right)^c = E_{\neg \varphi} \), and from \( A_x \subseteq E_\varphi \) it follows \( \left( E_\varphi \right)^c \land A_x = \emptyset \), that is \( E_{\neg \varphi} \cap A_x = \emptyset \).

Given two fuzzy subsets \( E_\varphi^* \) and \( E_\varphi \), the rules of conditional probability give, taking into account (1),

\[
P \left( E_\varphi \lor E_\psi \mid A_x \land A_y \right) = P \left( E_\varphi \mid A_x \right) + P \left( E_\psi \mid A_y \right) - P \left( E_\varphi \land E_\psi \mid A_x \land A_y \right).
\]

Therefore, to evaluate \( u = P \left( E_\varphi \lor E_\psi \mid A_x \land A_y \right) \) it is necessary—and sufficient—to know also the value of the conditional probability \( v = P \left( E_\varphi \lor E_\psi \mid A_x \land A_y \right) \), and vice versa.

**Remark 3.** Putting \( E_\varphi \mid A_x \land A_y = A \), \( E_\psi \mid A_x \land A_y = B \),

\[
P(A) = P(E_\varphi \mid A_x) = a, \quad P(B) = P(E_\psi \mid A_y) = b,
\]

and \( A \lor B = U \), \( A \land B = V \), (2) can be written as a functional equation

\[
S(a, b) = a + b - T(a, b),
\]

since, by resorting to the theorem characterizing coherent probability assessments—a particular case of Theorem A1—it is not difficult to prove that among all possible extensions of \( P \) from \( A \) and \( B \) to the “new” event \( A \land B = V \), any chosen coherent value \( P(V) \) belongs to an interval whose end points depend only on \( P(A) \) and \( P(B) \), that is

\[
\max\{a + b - 1, 0\} \leq P(V) \leq \min\{a, b\},
\]

so that it is a suitable function \( T(a, b) \) of \( a \) and \( b \); similarly, the coherent interval for \( P(A \lor B) = P(U) \) is

\[
\max\{a, b\} \leq P(U) \leq \min\{a + b, 1\},
\]

and so \( P(U) \) can be denoted as \( S(a, b) \). Notice that the above functional equation is the same discussed by Frank (1979), restricted to the pairs \((a, b)\) of kind (3).

In conclusion, the only constraint for the value of \( v \) is

\[
\max \left\{ P \left( E_\varphi \mid A_x \right) + P \left( E_\psi \mid A_y \right) - 1, 0 \right\} \leq v \leq \min \left\{ P \left( E_\varphi \mid A_x \right), P \left( E_\psi \mid A_y \right) \right\},
\]

(4)
Three possible choices for the value of the conditional probability \( v \) give rise to different well-known—see, e.g., Klement et al. (2000)—\( t \)-norms and \( t \)-conorms:

(a) give \( v \) the maximum possible value, that is \( v = \min \left\{ P \left( E_\varphi | A_x \right), P \left( E_\psi | A_y \right) \right\} \); then in this case we necessarily obtain, by (2), that

\[
P \left( E_\varphi \vee E_\psi | A_x \wedge A_y \right) = \max \left\{ P \left( E_\varphi | A_x \right), P \left( E_\psi | A_y \right) \right\}.
\]

This assignment corresponds to the choice of the so-called \( T_M \) and \( S_M \) as \( T \)-norm and \( T \)-conorm.

(b) give \( v \) the minimum value, that is \( \max \left\{ P \left( E_\varphi | A_x \right) + P \left( E_\psi | A_y \right) - 1, 0 \right\} \), i.e., the \( \text{Łukasiewicz} \) \( T \)-norm. In this case we necessarily obtain, again by (2), that

\[
P \left( E_\varphi \vee E_\psi | A_x \wedge A_y \right) = \min \left\{ P \left( E_\varphi | A_x \right) + P \left( E_\psi | A_y \right), 1 \right\}
\]

i.e. the \( \text{Łukasiewicz} \) \( T \)-conorm.

(c) give \( v \) the value \( P \left( E_\varphi | A_x \right) P \left( E_\psi | A_y \right) \), that is assume that \( E_\varphi \) is stochastically independent of \( E_\psi \) given \( A_x \wedge A_y \). In this case we necessarily obtain

\[
P \left( E_\varphi \vee E_\psi | A_x \wedge A_y \right) = P \left( E_\varphi | A_x \right) + P \left( E_\psi | A_y \right) - P \left( E_\varphi | A_x \right) P \left( E_\psi | A_y \right),
\]

i.e. the so-called probabilistic sum \( S_P \) and product \( T_P \).

We recall that the three coherent choices discussed above correspond to the particular values \( \lambda = 0, \lambda = 1, \lambda = \infty \), respectively, of the fundamental (archimedean) Frank \( t \)-norms \( T_\lambda \) and \( t \)-conorms \( S_\lambda \)—see Frank (1979)—with \( \lambda \in [0, \infty] \), that is (for \( \lambda \) different from the above three values)

\[
T_\lambda(x, y) = \log_\lambda \left( 1 + \frac{(\lambda^x - 1)(\lambda^y - 1)}{\lambda - 1} \right),
\]

\[
S_\lambda(x, y) = 1 - \log_\lambda \left( 1 + \frac{(\lambda^{1-x} - 1)(\lambda^{1-y} - 1)}{\lambda - 1} \right).
\]

In our framework (where, given the value \( P \left( E_\varphi \wedge E_\psi | A_x \wedge A_y \right) \) relative to the intersection, singling-out a \( t \)-norm, then the corresponding choice of the \( t \)-conorm, which determines the value \( P \left( E_\varphi \vee E_\psi | A_x \wedge A_y \right) \) relative to the union, is uniquely driven by the coherence of the relevant conditional probability) we are able to capture also Frank \( t \)-norms and \( t \)-conorms, for any \( \lambda \in [0, \infty] \).

3. Perception-based probabilistic reasoning

The radical thesis placed by Zadeh on the table is that PT has serious limitations that cannot be overcome within the conceptual structure of bivalent logic, and so it cannot provide tools for operating on perception-based information.

Our thesis is that we are able to manage situations of this kind, since PT is just a trivial particular case of our general approach to probability—and, mainly, to conditional probability—through coherence, and this approach encompasses all the tools that are necessary to deal with the kind of problems raised by Zadeh. On the other hand, for some of them—namely those numbered 1, 2, 3 in Zadeh (2002)—it is unclear why they have been put in the list; possibly it depends on a restrictive view of probability only in terms of observed—or observable—frequency, which is instead only a way—among many others—for computing a probability:

1. What is the probability that my tax return will be audited?
2. What is the probability that my car may be stolen?
3. How long does it take to get from the hotel to the airport by taxi?

So we do not consider these situations as pertaining to our discussion. Concerning other problems in which a role is played by a perception-based information (to use Zadeh’s terminology), we are not going to discuss any single example,
but we refer to the *model constituted by some balls in a box*, in different situations concerning their number, size, color, and any other feature of interest.

We start with the following very simple situation:

A box contains \( n \) balls of various sizes \( s_1, \ldots, s_k \) (\( k \leq n \)), with respective fractions \( f_1, \ldots, f_k \). Consider an experiment consisting in drawing a ball from the box, and let \( E_L \) be the event (referred to the drawn ball) “You claim (the size is) large”. The question is: what is the probability of \( E_L \)?

In our context the problem is trivial, since it amounts to the computation of the probability of the (“fuzzy”) event \( E_L \). Introduce the random variable \( \mathcal{S} \), with range \{ \( s_1, \ldots, s_k \) \}, of the sizes of the balls in the box, and consider the conditional events \( E_L | S_i \), with \( S_i = \{ \mathcal{S} = s_i \} \). So by a trivial computation we have

\[
P (E_L) = \sum_i P (E_L | S_i) P (S_i) = \sum_i \mu_L (s_i) f_i,
\]

where \( \mu_L \) is the membership function of the fuzzy set “large”, which is context dependent, since it refers to the sizes of the balls in the given box.

The same procedure can be obviously followed for the properties “medium” and “small”.

**Notation.** From now on, to simplify the exposition, we will sometimes hallmark a “fuzzy” event by the symbol \( F \) followed by the relevant property. Recall that in our context—as discussed in Section 2, Remark 1—a fuzzy event is a . . . “normal” event, i.e., a particular proposition beginning with “You claim . . .”: for example, the event \( E_L = “\text{You claim large}” \) can be simply written as

\[
E_L = F\text{-large},
\]

and the range of the property “large” should be clear from the context (e.g., in this case, the sizes of the balls).

Consider now the case that the balls have different colors—with known fractions—and introduce the corresponding random variable \( \mathcal{C} \), with range \( c_1, \ldots, c_m \). We can consider the fuzzy subset of \( \mathcal{C} \) singled-out by the membership function

\[
\mu_D (c_r) = P (E_D | C_r),
\]

where \( E_D \) is a . . . “normal” event, i.e., a particular proposition beginning with “You claim . . .”: for example, the event \( E_D = “\text{You claim large and dark}” \) can be simply written as

\[
E_D = F\text{-dark}.
\]

By the same procedure followed above for \( P (E_L) \), getting

\[
P (E_L \land E_D) = \sum_{i,j} P ((E_L \land E_D) \land (S_i \land C_j)) p_{ij} = \sum_{i,j} \mu_{LD} (s_i, c_j) p_{ij},
\]

where \( p_{ij} \) denotes the probability of a size \( s_i \) and a color \( c_j \); the values \( p_{ij} \) can be computed by disintegration with respect to the possible compositions of the box, possibly taking into account suitable conditional independence assumptions. Concerning the membership \( \mu_{LD} (s_i, c_j) \), any value belonging to the interval singled-out by formula (4) of the previous section does the job, and also any Frank \( t \)-norm: in particular, a possible choice is the usual “minimum”, as done by Zadeh.

Notice that this general approach allows to assume—possibly—dependence between size and color, taking in this case the product as \( t \)-norm.

Suppose now that we ignore the color of the balls and that we do not know the fractions \( f_i \) of balls having the various sizes \( s_i \), i.e. the box is of unknown composition. Consider the event, referred to the balls in the box

\[
E_{ML} = “\text{You claim that most are large}” = F\text{-}(most\text{ are large}).
\]

The relevant claim may be seen as the assignment of a high probability of being claimed—referred to a ball to be drawn—large. Clearly, the possible values of this probability are identified with the possible values of the fractions \( p_i \) of large balls, and these values constitute the range of a random variable \( \Pi \), related to the events \( \Pi_i = \{ \Pi = p_i \} \). Then
we can introduce the membership function \( \mu_M(p_i) = P(E_M | P_i) \), with \( E_M = F \)-high, referred to the probability of drawing a large ball.

Going back to \( E_{M,L} \), one of the questions posed by Zadeh is: assuming \( E_{M,L} \), i.e. that most balls are large, what is the probability of drawing a large ball?

In our setting and notation, this amounts to the evaluation of the conditional probability \( P(E_L | E_{M,L}) \). Introducing events \( H_k \) \( (k = 1, 2, \ldots, r) \) denoting the possible compositions of the box with respect to the possible sizes of the balls, we may resort to the disintegration formula

\[
P(E_L | E_{M,L}) = \sum_k P(E_L | H_k \land E_{M,L}) P(H_k | E_{M,L}) = \sum_k P(E_L | H_k) P(H_k | E_{M,L}),
\]

the latter equality coming from \( E_L \) being conditionally independent of \( E_{M,L} \) given \( H_k \); in fact, for any known composition of the box, the probability of drawing a large ball does not depend on the knowledge that most balls are large.

Recalling the random variable \( H \) introduced above, we can represent the probability \( P(E_L | H_k) \) appearing in the previous formula as

\[
P(E_L | H_k) = \sum_i P(E_L | H_k \land H_i) P(H_i | H_k) = \sum_i P(E_L | H_k) P(H_i | H_k),
\]

while the probability \( P(H_k | E_{M,L}) \) can be easily computed via Bayes’ theorem and \( P(E_{M,L} | H_k) \).

4. Likelihood and its (coherent) extensions

Taking as starting point a membership function, looked on as a pointwise distribution, in our setting it is completely natural to consider the so-called fuzzy measures. In fact this requires only to extend (Theorem A3, see Appendix) a coherent conditional probability assessment given on the family \( \{E_\varphi | A_x\} \) to the larger family of conditional events \( \{E_\varphi | \mathcal{A}\} \), with \( A \) element of the algebra \( \mathcal{A} \) spanned by events \( \{A_x\} \). Intuitively, \( P(E_\varphi | A) \) is the probability that “You claim \( \varphi \)” in the hypothesis that the value of the variable \( X \) belongs to \( A \).

4.1. Possibility functions and measures

We start with the following (“natural”) definition.

**Definition 2.** Let \( E \) be an arbitrary event and \( P \) any coherent conditional probability on the family \( \mathcal{G} = \{E \times \{A_x\}_{x \in C_X}\} \), admitting \( P(E|\Omega) = 1 \) as (coherent) extension. A distribution of possibility on \( C_X \) is the real function \( \pi(x) = P(E | A_x) \).

When \( C_X \) is finite, since every extension of \( P(E|\cdot) \) must satisfy the axioms of a conditional probability (see Appendix), condition \( P(E|\Omega) = 1 \) gives

\[
P(E|\Omega) = \sum_{x \in C_X} P(A_x | \Omega) P(E | A_x) \quad \text{and} \quad \sum_{x \in C_X} P(A_x | \Omega) = 1.
\]

Then \( 1 = P(E|\Omega) \leq \max_{x \in C_X} P(E | A_x) \); therefore \( P(E|A_x) = 1 \) for at least one event \( A_x \).

On the other hand, we notice that in our framework—where null probabilities for possible conditioning events are allowed—from \( P(E) = 1 \) it does not necessarily follow that \( P(E | A_x) = 1 \) for every \( x \); in fact we may well have \( P(E | A_y) = 0 \)—or else equal to any other number between 0 and 1—for some \( y \in C_X \). Obviously, the constraint \( P(E | A_x) = 1 \) for some \( x \) is not necessary when the cardinality of \( C_X \) is infinite.

From now on, given an arbitrary event \( E \), let \( \mathcal{G} \) be a family of conditional events \( \{E | H_i\}_{i \in I} \), where \( \text{card}(I) \) is arbitrary and events \( H_i \)'s are a partition of \( \Omega \); \( P(E|\cdot) \) is an arbitrary—coherent—conditional probability on \( \mathcal{G} \); \( \mathcal{K} \) is the algebra spanned by the \( H_i \)'s, and \( \mathcal{K}^0 = \mathcal{K} \setminus \{\emptyset\} \).

Here we list some of the main relevant results, taken from Coletti and Scozzafava (2003a,b).
By Theorem A1 (see Appendix), $P$ can be extended to a coherent conditional probability on $\mathcal{E}' = \{ E | H : H \in \mathcal{H}' \}$, and the latter in turn can be extended to a coherent conditional probability on $\mathcal{E}' = \mathcal{E}' \cup \{ H | K : H, K \in \mathcal{H} \}$. This satisfies, by axiom (iii) of a conditional probability,

$$P(E|H \vee K) = P(E|H)P(H|H \vee K) + P(E|K)P(K|H \vee K),$$

for every $H \wedge K = \emptyset$. Since we have—by axioms (i) and (ii) of a conditional probability—$P(H|H \vee K) + P(K|H \vee K) = 1$, we get the following

**Theorem 1.** Any coherent extension of $P$ to $\mathcal{E}'$ is such that, for every $H, K \in \mathcal{H}$, with $H \wedge K = \emptyset$,

$$\min\{P(E|H), P(E|K)\} \leq P(E|H \vee K) \leq \max\{P(E|H), P(E|K)\}. \quad (5)$$

It follows that any coherent extension of $P$ to $\mathcal{E}'$ is such that, for every $H, K \in \mathcal{H}$, with $H \wedge K = \emptyset$,

$$P(E|H \vee K) \leq P(E|H) + P(E|K).$$

**Remark 4.** The previous inequality can be easily extended to a partition of $\Omega$. Notice that, except in the trivial case that every partition has an event $H_J$ with $P(E|H_J) = 1$ while for all others $P(E|H_I) = 0$, the function $L(\cdot) = P(E|\cdot)$ is not additive. For a further discussion of this aspect, see the following Section 4.3.

**Theorem 2.** Any real function $f$ defined on $\mathcal{H}$ such that, if $H \wedge K = \emptyset$,

$$\min\{f(H), f(K)\} \leq f(H \vee K) \leq \max\{f(H), f(K)\},$$

is a capacity—i.e., it is monotone with respect to the implication $\subseteq$—if and only if, for every $H, K \in \mathcal{H}$,

$$f(H \vee K) = \max\{f(H), f(K)\}.$$ 

So the function $f(\cdot) = P(E|\cdot)$, with $P$ a coherent conditional probability, in general is not a capacity.

The question now is: are there coherent conditional probabilities $P(E|\cdot)$ which are capacities? We reached a positive answer by means of the following result—given in Coletti and Scozzafava (2004)—which represents the main tool to introduce possibility measures in our context referring to coherent conditional probabilities.

**Theorem 3.** Let $f : \mathcal{E} \to [0, 1]$ be any function such that

$$f(E|H_I) = 0 \text{ if } E \wedge H_I = \emptyset \text{ and } f(E|H_I) = 1 \text{ if } H_I \subseteq E \quad (6)$$

holds. Then

(i) $f$ is a coherent conditional probability;

(ii) any $P$ extending $f$ on $\mathcal{H} = \{ E \} \times \mathcal{H}'$ and such that

$$P(E|H \vee K) = \max\{P(E|H), P(E|K)\} \text{ for every } H, K \in \mathcal{H}' \quad (7)$$

is a coherent conditional probability.

**Definition 3.** Let $\mathcal{H}$ be an algebra of subsets of $C_X$ and $E$ an arbitrary event. If $P$ is any coherent conditional probability on $\mathcal{H} = \{ E \} \times \mathcal{H}'$, with $P(E|\Omega) = 1$ and such that

$$P(E|H \vee K) = \max\{P(E|H), P(E|K)\} \text{ for every } H, K \in \mathcal{H}'$$

then a possibility measure on $\mathcal{H}$ is the real function $\Pi$ defined by $\Pi(H) = P(E|H)$ for $H \in \mathcal{H}'$ and $\Pi(\emptyset) = 0$.

In our context, Theorem 3 assures that any possibility measure can be obtained as coherent extension—unique, in the finite case—of a possibility distribution. Vice versa, given any possibility measure $\Pi$ on an algebra $\mathcal{H}$, there exists an
event $E$ and a coherent conditional probability $P$ on $\mathcal{H} = \{E\} \times \mathcal{H}^0$ agreeing with $\Pi$, i.e. whose extension to $\{E\} \times \mathcal{H}$ (with $P(E|\emptyset) = 0$) coincides with $\Pi$.

An immediate consequence of Theorems 2 and 3 is the following:

**Corollary 1.** Any coherent $P$ extending $f$ on $\mathcal{H} = \{E\} \times \mathcal{H}^0$ is a capacity if and only if it is a possibility.

Going back to our interpretation of a membership function $\mu(x)$ through a suitable coherent conditional probability, and putting

$$H_0 = \{ x \in \mathcal{C}_X : \mu(x) = 0 \}, \quad H_1 = \{ x \in \mathcal{C}_X : \mu(x) = 1 \},$$

the conditional probability $P(E|H^c)$, with $H = H_0 \cup H_1$, is a measure of how much You are willing to claim property $\varphi$ if the only fact you know is that $x \in H^c$. On the other hand, every membership function with $H_1 \neq \emptyset$ can be regarded as a possibility distribution.

**Remark 5.** If $\mathcal{A}$ is an algebra of subsets of $\mathcal{C}_X$, the ensuing possibility measure can be interpreted in the following way: it is a sort of “aggregated” membership—relative to each finite $A \in \mathcal{A}$—which takes, among all the possible choices for its value on $A$, i.e. among all possible extensions satisfying (6), the maximum of the membership in $A$. Then some of the usual arguments appear counterintuitive: in fact, the “aggregated” membership should possibly decrease when the information is not concentrated on a given $x$ but is “spread” over a larger set. For example, considering the statement “the natural number $x$ is small”, you may be willing, if you know that $x = 39$, to put $\mu(x) = 0.2$, while if you know that $x = 2$, you may be willing to put $\mu(x) = 0.9$; on the other hand, knowing that $x$ is between 2 and 39, the corresponding possibility is still 0.9. So our results may suggest to take as such aggregated measure a function which is not a capacity, yet satisfying condition (6).

### 4.2. Information measures

With the aim of studying information measures in the framework of coherent conditional probabilities, we gave also the following definition, which parallels, in a sense, Definition 2 for possibility measures.

**Definition 4.** Let $F$ be an arbitrary event and $P$ any coherent conditional probability on the family $\mathcal{G} = \{F\} \times \{A_x\}_{x \in C_X}$, admitting $P(F|\Omega) = 0$ as (coherent) extension. We define pointwise information measure on $C_X$ the real function $\psi(x) = P(F|A_x)$.

When $C_X$ is finite, since every extension of $P(F|\cdot)$ must satisfy the axioms of a conditional probability, considering the condition $P(F|\Omega) = 0$, we necessarily have

$$P(F|\Omega) = \sum_{x \in C_X} P(A_x|\Omega) P(F|A_x) \quad \text{and} \quad \sum_{x \in C_X} P(A_x|\Omega) = 1.$$ 

Then $0 = P(F|\Omega) \leq \min_{x \in C_X} P(F|A_x)$, so $P(F|A_x) = 0$ for at least one event $A_x$. On the other hand, we notice that in our framework it does not necessarily follow, from $P(F) = 0$, that $P(F|A_x) = 0$ for every $x$; in fact we may well have $P(F|A_y) = 1$—or to any other number between 0 and 1—for some $y \in C_X$. Obviously, the constraint $P(F|A_x) = 0$ for some $x$ is not necessary when the cardinality of $C_X$ is infinite; recall that we are in a finitely additive setting, and so all the events of a partition may have zero probability.

Under the same conditions mentioned before Theorem 1, we get an immediate consequence of Theorem 1 itself:

**Theorem 4.** Any real function $f$ defined on $\mathcal{H}$ such that, if $H \land K = \emptyset$,

$$\min\{ f(H), f(K) \} \leq f(H \lor K) \leq \max\{ f(H), f(K) \},$$

is antimonotone with respect to $\subseteq$ if and only if, for every $H, K \in \mathcal{H}$,

$$f(H \lor K) = \min\{ f(H), f(K) \}.$$
The following result proves the existence of coherent conditional probabilities $P(F|\cdot)$ antimonotone with respect to $\subseteq$. It represents also the main tool to introduce information measures in our context referring to coherent conditional probabilities.

**Theorem 5.** Let $f : \mathcal{E} \to [0, 1]$ be any function such that

$$f(F | H_i) = 0 \text{ if } F \cap H_i = \emptyset \quad \text{and} \quad f(F | H_i) = 1 \text{ if } H_i \subseteq F$$

holds. Then any $P$ extending $f$ on $\mathcal{E} = \{F\} \times \mathcal{H}^0$ and such that

$$P(F | H \lor K) = \min\{P(F | H), P(F | K)\} \quad \text{for every } H, K \in \mathcal{H}^0,$$

is a coherent conditional probability.

In the case that the assessment $P(F | H_i)$ admits $P(F | \emptyset) = 0$ as coherent extension, we obtain as well a coherent extension by requiring both $P(F | \emptyset) = 0$ and choosing—as in the above theorem—“min” as combination rule to make the extension of $P$.

We recall now the definition of information measure as given by Kampé de Feriet and Forte (1967).

**Definition 5.** A function $I$ from an arbitrary set of events $\mathcal{E}$, and with values on $R^* = [0, +\infty]$, is an information measure if it is antimonotone, i.e. the following condition holds: for every $A, B \in \mathcal{E}$

$$A \subseteq B \implies I(B) \leq I(A).$$

So, if both $\emptyset$ and $\Omega$ belong to $\mathcal{E}$, it follows that

$$0 \leq I(\Omega) = \inf_{A \in \mathcal{E}} I(A) \leq \sup_{A \in \mathcal{E}} I(A) = I(\emptyset).$$

It is claimed that the above inequality is necessary and sufficient to build up an information theory; nevertheless, with the purpose of attributing an universal value to $I(\emptyset)$ and $I(\emptyset)$, the further conditions $I(\emptyset) = +\infty$ and $I(\Omega) = 0$ are given. The choice of these two values is obviously aimed at reconciling with the Wiener–Shannon theory. In general, the above definition implies only that

$$I(A \lor B) \leq \min\{I(A), I(B)\};$$

we can specify the rule of composition by introducing a binary operation $\odot$ to compute $I(A \lor B)$ by means of $I(A)$ and $I(B)$.

**Definition 6.** An information measure defined on an additive set of events $\mathcal{A}$ is $\odot$-decomposable if there exists a binary operation $\odot$ from $[0, +\infty] \times [0, +\infty]$ to $[0, +\infty]$ such that, for every $A, B \in \mathcal{A}$ with $A \land B = \emptyset$, we have

$$I(A \lor B) = I(A) \odot I(B).$$

So “min” is just one of the possible choices of $\odot$. Obviously, every bounded non-negative function $I$ defined on an algebra of events, with $I(\Omega) = 0$ and satisfying $I(A \lor B) = \min\{I(A), I(B)\}$, can be regarded as an increasing transformation $g$ of an information measure.

On the other hand, our results make us able to introduce—from our point of view—a “convincing” definition of information measure.

**Definition 7.** Let $\mathcal{H}$ be an algebra of subsets of $C_X$ (the range of a random quantity $X$) and $F$ an arbitrary event. If $P$ is any coherent conditional probability on $\mathcal{H} = \{F\} \times \mathcal{H}^0$, with $P(F | \emptyset) = 0$ and such that

$$P(F | H) \leq P(F | K) \quad \text{for } H, K \in \mathcal{H}^0 \text{ with } K \subseteq H,$$

then an information measure on $\mathcal{H}$ is a real function $I$ defined, for $H \in \mathcal{H}^0$, by $I(H) = g(P(F | H))$, with $g$ any increasing function from $[0, 1]$ to $[0, +\infty]$ and $I(\emptyset) = +\infty$. 


Definition 8. Under the same conditions as above, if $P$ is any coherent conditional probability on $\mathcal{X} = \{F\} \times \mathcal{H}^0$, with $P(F|\Omega) = 0$ and such that

$$P(F|H \vee K) = \min\{P(F|H), P(F|K)\} \quad \text{for} \quad H, K \in \mathcal{H}^0,$$

then a min-decomposable information measure on $\mathcal{X}$ is a real function $I$ defined, for $H \in \mathcal{H}^0$, by $I(H) = g(P(F|H))$, with $g$ any increasing function from $[0, 1]$ to $[0, +\infty]$ and $I(\emptyset) = +\infty$.

In conclusion, we have proved the following:

Theorem 6. Under the same conditions of Theorem 5, any $P$ extending $f$ on $\mathcal{X} = \{F\} \times \mathcal{H}^0$ is an increasing transformation of an information measure if and only if it is an increasing transformation of a min-decomposable information measure.

4.3. Likelihood and statistical inference

Among statisticians there seems to be lack of consensus about the meaning of “statistical information” and how such an important notion should be meaningfully formalized: see, e.g., Basu (1988). On the contrary, if we agree to the stipulation that our search for the “whole of the relevant information in the data” should be limited within the framework of a given statistical model, then most statisticians could not find any cogent reason for not identifying the “information in the data” with the likelihood function generated by it.

An ensuing and long debated problem concerns the following question: is the likelihood just a point function or can it be also seen as a measure? Why cannot we talk of the likelihood of a composite hypothesis in the same way as we talk about the probability of a composite event? Statisticians are usually inclined to accept the following “law of likelihood”: of two (simple) hypotheses that are consistent with given data, the better supported by the data is the one that has greater likelihood $L(\omega)$, where $\omega$ ranges in the parameter space $\Omega$. An immediate consequence of this is the controversial inferential method based on the choice of the $\omega$ which corresponds to the “maximum likelihood”.

On the other hand, not all statisticians are willing to support also the so-called “strong law of likelihood”, that can be expressed as follows: for any two subsets $A$ and $B$ of the parameter space $\Omega$, the data support the hypothesis $\omega \in A$ better than the hypothesis $\omega \in B$ if

$$\sum_{\omega \in A} L(\omega) > \sum_{\omega \in B} L(\omega).$$

This amounts—essentially—to extend the domain of the likelihood function, since these sums can in fact be interpreted as if they were the “aggregated” likelihoods of the sets $A$ and $B$.

In Sections 4.1 and 4.2 we have shown different ways of extending coherently a conditional probability $P(E|\cdot)$, and a very particular case of it—referring to discrete distributions—is the likelihood $L(\omega) = P(E|\omega)$. As already noticed in Remark 4, $L(\omega)$ is not additive, and this is true even when it is interpreted (in a Bayesian context) as the posterior corresponding to a uniform prior.

Example. Consider a parameter space $\{\omega_1, \omega_2, \ldots, \omega_n\}$, with $P(\omega_i) = 1/n$; we have, assuming $P(E) > 0$,

$$P((\omega_1 \vee \omega_2)|E) = \frac{P(\omega_1 \vee \omega_2) P(E|(\omega_1 \vee \omega_2))}{P(E)}, \quad (8)$$

$$P(\omega_i|E) = \frac{P(\omega_i) P(E|\omega_i)}{P(E)} = \frac{P(E|\omega_i)}{nP(E)}, \quad i = 1, 2, \quad (9)$$

and so, by adding the two Eqs. (9), we get

$$P(\omega_1|E) + P(\omega_2|E) = P((\omega_1 \vee \omega_2)|E) = \frac{P(E|\omega_1) + P(E|\omega_2)}{nP(E)}.$$
Then, since $P(\omega_1 \lor \omega_2) = 2/n$, from (8) it follows:

$$2P(E | (\omega_1 \lor \omega_2)) = P(E | \omega_1) + P(E | \omega_2),$$

i.e. $P(E | (\omega_1 \lor \omega_2))$ is a convex combination (with equal weights) of $P(E | \omega_1)$ and $P(E | \omega_2)$.

Going back to the results of Sections 4.1 and 4.2, the question is whether the two extreme cases obtained extending $P(E|A)$ to the union of conditioning events by taking the maximum or by taking the minimum—i.e. possibility measure or antimonotone measure—are the most natural ways to extend membership—or likelihood—functions.

We recall that, given a finite partition $\mathcal{H}_0 = \{H_i\}$ of $\Omega$, coherence implies

$$\min_i \{P(E | H_i)\} \leq P\left(E \bigg| \bigvee_i H_i \right) \leq \max_i \{P(E | H_i)\}$$

but the converse is not true. So, in general, a value between the two extremes is not necessarily a coherent choice for the conditional probability $P(E | \bigvee_i H_i)$, which can be looked on as a sort of “aggregated” membership or of likelihood “measure”.

Coherent choices have been characterized—for a finite family—in Ceccacci et al. (2003): they are weighted means of the $P(E | H_i)$’s, where weights equal to zero or one are allowed. Here is the relevant theorem.

**Theorem 7.** Let $E$ be an arbitrary event and $\mathcal{A}$ be a finite family of conditional events $\{E | H_i\} (i = 1, 2, \ldots, n)$, where $\mathcal{H}_0 = \{H_i\}$ is a partition of $\Omega$. Let $\mathcal{A}$ be the algebra spanned by the $H_i$’s, and put $\mathcal{A}^0 = \mathcal{A} \setminus \{\emptyset\}$. If $P : \mathcal{A} \to [0, 1]$ is a coherent conditional probability, i.e. any function such that

$$P(E | H_i) = 0 \text{ if } E \cap H_i = \emptyset, \quad P(E | H_i) = 1 \text{ if } H_i \subseteq E,$$

the following two statements are equivalent:

(i) an extension of $P$ to $\mathcal{H} = \{E\} \times \mathcal{A}^0$ is a coherent conditional probability;

(ii) there exist subfamilies $\mathcal{H}_0 \supset \mathcal{H}_1 \supset \cdots \supset \mathcal{H}_2 \supset \cdots \supset \mathcal{H}_k$ and relevant sets of coefficients $\lambda_i^2 \geq 0 (i = 1, \ldots, i_2)$, where $i_2$ is the number of events $H_i \in \mathcal{H}_2$, with $H_i \in \mathcal{H}_2$ if and only if $\lambda_i^2 - 1 = 0$ and $\lambda_i^1 - 1 = 0$ for any $i$, with

$$\sum_i \lambda_i^2 = 1,$$

for all $\mathcal{H}_2$, and if at least one $H_i$ belongs to $\mathcal{H}_2 \setminus \mathcal{H}_{i+1}$, then $P(E | H)$ is the only solution of (10).

We sketch the procedure to search for the $\lambda_i^2$’s, given an extension of the assessment $\{P(E | H_i) \cdot i = 1, 2, \ldots, n\}$, and so to prove that the extension is coherent. Since all the $H_i$’s belong to $\mathcal{H}_0$, then (10) holds (with $x = 0$) for all (possible) events belonging to the algebra $\mathcal{A}$. So the (first) set of coefficients $\lambda_i^0 \cdot (i = 1, 2, \ldots, n)$ satisfy all equations of kind (10), with $x = P(E | H)$ for every $H \in \mathcal{A}^0$. Given now an $H \in \mathcal{A}^0$, if at least one $\lambda_i^0$ (for $i$ such that $H_i \subseteq H$) is positive, then $P(E | H)$ is the only solution of (10), and we have

$$P(E | H) = \frac{1}{\lambda(H)} \sum_{H_i \subseteq H} \lambda_i^0 P(E | H_i), \quad (11)$$

where

$$\lambda(H) = \sum_{H_i \subseteq H} \lambda_i^0.$$

Moreover, if on the contrary $\lambda_i^0 = 0$ for all $i$ such that $H_i \subseteq H$, then (10) is trivially satisfied for all value of $x$, and all $H_i \subseteq H$ belong to $\mathcal{H}_1$. In this case we must find coefficients $\lambda_i^1$ satisfying all equations (10) related (only) to the events $H$ obtained as unions of the $H_i$’s such that $\lambda_i^0 = 0$; and so on.
Appendix. It has in this case the solution, for the probabilities of atoms generated by the given events. A further information comes from his database, that is the “partial likelihood” which the doctor makes the probability assessments. Moreover, he believes that \(H\) implies both \(E\) and \(H\), so—by our approach to default reasoning, see Coletti and Scozzafava (2002a)—we have the insufficiency, so that there is no need to resort from time to time to different “ad hoc” procedures: see also Coletti et al. (2001).

A patient arrives at a hospital showing symptoms of choking, and the doctor considers the following hypotheses concerning the patient situation:

\[ H_1 = \text{“cardiac insufficiency”}, \quad H_2 = \text{“asthma attack”}, \quad H_3 = H_2 \land H, \]

where \(H = \text{“cardiac lesion”}\).

The doctor does not regard them as mutually exclusive; moreover, he assumes the following natural logical relation:

\[ H_3 \subset H_1 \land H_2. \]

The doctor makes the probability assessments

\[ P(H_2) = \frac{1}{3}, \quad P(H_3) = \frac{1}{5}, \quad P(H_1 \lor H_2) = \frac{3}{5} \]

and he expresses the comparative judgement

\[ H_2 \prec H_1, \]

i.e. the asthma attack is “less probable” than cardiac insufficiency. It means that \(P(H_1)\) must be a value strictly greater than \(\frac{1}{3}\), so that we get—in checking coherence—an inequality in the relevant system: see the theory recalled in the Appendix. It has in this case the solution, for the probabilities of atoms generated by the given events

\[ P(H_1 \land H_2 \land H_3) = p - \frac{7}{15}, \quad P(H_1^c \land H_2 \land H_3^c) = \frac{3}{5} - p, \]

\[ P(H_1 \land H_2 \land H_3) = \frac{1}{5}, \quad P(H_1 \land H_2^c \land H_3^c) = \frac{4}{15}, \quad P(H_1^c \land H_2^c \land H_3^c) = \frac{2}{5} \]

under the constraint \(\frac{7}{15} \leq p \leq \frac{3}{5}\), where the parameter \(p\) represents the probability of \(H_1\).

Put now \(E = \text{“taking the medicine M against asthma does not reduce choking symptoms”}\). Since the fact \(E\) is incompatible with having asthma attack \(H_2\), unless the patient has cardiac insufficiency or lesion—recall that \(H_3\) implies both \(H_1\) and \(H_2\)—then for the doctor

\[ H_2 \land H_1^c \land E = \emptyset. \]

Moreover, he believes that usually the medicine \(M\) does not reduce symptoms if the patient suffers from cardiac insufficiency, so—by our approach to default reasoning, see Coletti and Scozzafava (2002a)—we have \(P(E \mid H_1) = 1\). A further information comes from his database, that is the “partial likelihood” \(P(E \mid H_1^c) = \frac{1}{4}\).

The whole “knowledge” is coherent, as it can be proved by solving the relevant system, and we get, for the probabilities of the atoms,

\[ P(E^c \land H_1 \land H_2 \land H_3) = 0, \quad P(E^c \land H_1 \land H_2 \land H_3^c) = 0, \quad P(E^c \land H_1^c \land H_2 \land H_3^c) = \frac{2}{15}, \]
The assessment database, we may put, e.g.,

\[ \text{if } G \subseteq \mathcal{B} \text{ and } H \subseteq \mathcal{B} \text{ then } \mathbb{P}(E) = \frac{2}{3}, \]

\[ \mathbb{P}(E \cap H_1 \cap H_2 \cap H_3) = 0, \quad \mathbb{P}(E \cap H_1 \cap H_2) = \frac{1}{3}, \quad \mathbb{P}(E \cap H_1) = \frac{2}{3}, \quad \mathbb{P}(E) = \frac{1}{3}, \]

and \( p = \frac{7}{15} \). Then the updating process allows to compute (for \( i = 1, 2, 3 \)) \( \mathbb{P}(H_i \mid E) \), for example \( \mathbb{P}(H_2 \mid E) = \frac{3}{7} \).

Now, if someone gives the doctor the vague information that “the patient suffers from hypertension”, then we may resort to our approach to fuzzy theory by introducing the conditional event \( I \mid A_x \), with \( I = “\text{the doctor claims that the patient suffers from hypertension}” \) and \( A_x = “\text{the blood-pressure of the patient is } x” \); then, according to the doctor’s database, we may put, e.g.,

\[
P(I \mid A_x) = \begin{cases} 
0, & x \leq 160 \\
\frac{1}{3}, & 160 < x \leq 180 \\
\frac{3}{5}, & 180 < x \leq 200 \\
1, & x > 200
\end{cases}
\]

as the membership function of the fuzzy set of people “suffering from hypertension”. Since it is expressed in terms of a conditional probability, it is possible to go on in the updating process by assessing—coherently—\( \mathbb{P}(H_i \mid E \land I) \) by the same procedure.

We omit the details: this example extends Example 3 from Coletti and Scozzafava (2000) and has been discussed in Coletti and Scozzafava (2002b).

### Appendix A. Coherent conditional probability

We generalize the idea of de Finetti of looking at a conditional event \( E \mid H \), with \( H \neq \emptyset \), as a 3-valued logical entity, which is true when both \( E \) and \( H \) are true, false when \( H \) is true and \( E \) is false, "undetermined" when \( H \) is false, by letting the third value suitably depend on the given ordered pair \((E, H)\) and not being just an undetermined common value for all pairs: it turns out that this function is a measure of the degree of belief in the conditional event \( E \mid H \), which under suitable—and natural—conditions is a decomposable conditional measure, as explained in detail in Coletti and Scozzafava (1999b, 2001b); we can get, in particular, a conditional probability \( \mathbb{P}(E \mid H) \), in its most general sense related to the concept of coherence, satisfying the classic axioms as given by de Finetti (1949) and Popper (1959); see also Dubins (1975) and Rényi (1956), where condition (ii) is replaced by the stronger one of countable additivity.

These classic axioms for a conditional probability \( P \) are: given a set \( \mathcal{C} = \mathcal{B} \times \mathcal{B}^0 \) of conditional events \( E \mid H \) such that \( \mathcal{B} \) is a Boolean algebra and \( \mathcal{B} \subseteq \mathcal{B} \) is closed with respect to (finite) logical sums, and putting \( \mathcal{B}^0 = \mathcal{B} \setminus \{\emptyset\} \), then

\[
P : \mathcal{C} \rightarrow [0, 1]
\]

is such that

(i) \( \mathbb{P}(H \mid H) = 1 \), for every \( H \in \mathcal{B}^0 \),
(ii) \( \mathbb{P}(\cdot \mid H) \) is a (finitely additive) probability on \( \mathcal{B} \) for any given \( H \in \mathcal{B}^0 \),
(iii) \( \mathbb{P}((E \land A) \mid H) = \mathbb{P}(E \mid H) \cdot \mathbb{P}(A \mid (E \land H)) \), for every \( E, A \in \mathcal{B} \) and \( E, E \land H \in \mathcal{B}^0 \).

A peculiarity—which entails a large flexibility in the management of any kind of uncertainty—of this approach to conditional probability is that, due to its direct assignment as a whole, the knowledge—or the assessment—of the "joint" and "marginal" unconditional probabilities \( \mathbb{P}(E \land H) \) and \( \mathbb{P}(H) \) is not required; moreover, the conditioning event \( H \)—which must be a possible event—may have zero probability.

A conditional probability \( P \) is defined on \( \mathcal{B} \times \mathcal{B}^0 \): however, it is possible, through the concept of coherence, to handle also those situations where we need to assess \( P \) on an arbitrary set of conditional events

\[
\mathcal{C} = \{ E_i \mid H_1, \ldots, E_n \mid H_n \}.
\]

**Definition A1.** The assessment \( P(\cdot \mid \cdot) \) on \( \mathcal{C} \) is coherent if there exists \( \mathcal{C}' \supseteq \mathcal{C} \), with \( \mathcal{C}' = \mathcal{B} \times \mathcal{B}^0 \) (\( \mathcal{B} \) Boolean algebra and \( \mathcal{B} \subseteq \mathcal{B} \) closed with respect to logical sums) such that \( P \) can be extended from \( \mathcal{C} \) to \( \mathcal{C}' \) as a conditional probability.
We need also to recall the following:

**Definition A2.** Given an arbitrary finite family \( \{E_1, \ldots, E_n\} \), of events, all intersections

\[
E_1^* \land E_2^* \land \cdots \land E_n^*,
\]

different from the impossible event \( \emptyset \), obtained by putting—in all possible ways—in place of each \( E_i^* \), for \( i = 1, 2, \ldots, n \), the event \( E_i \) or its contrary \( E_i^c \), are called atoms generated by the given events. The events \( E_1, \ldots, E_n \) are called logically independent when the number of atoms equals \( 2^n \).

A characterization of coherence is given by the following theorem: see, e.g., Coletti (1994), Coletti and Scozzafava (1996, 1999b) and the book Coletti and Scozzafava (2002a).

**Theorem A1.** Let \( \mathcal{C} \) be an arbitrary finite family of conditional events and \( \mathcal{A}_0 \) denote the set of atoms \( A_r \) generated by the events \( E_1, H_1, \ldots, E_n, H_n \). For a real function \( P \) on \( \mathcal{C} \) the following two statements are equivalent:

(a) \( P \) is a coherent conditional probability on \( \mathcal{C} \);

(b) there exists (at least) a class of probabilities \( \{P_0, P_1, \ldots, P_k\} \), each probability \( P_\beta \) being defined on a suitable subset \( \mathcal{A}_x \subseteq \mathcal{A}_0 \), such that for any \( E_i \mid H_i \in \mathcal{C} \) there is a unique \( P_\beta \) with

\[
\sum_{A_r \subseteq H_i} P_\beta (A_r) > 0, \quad P (E_i \mid H_i) = \frac{\sum_{A_r \subseteq E_i \land H_i} P_\beta (A_r)}{\sum_{A_r \subseteq H_i} P_\beta (A_r)};
\]

moreover \( \mathcal{A}_x' \subseteq \mathcal{A}_x'' \) for \( \mathcal{A} > \mathcal{A}'' \) and \( P_\beta (A_r) = 0 \) if \( A_r \in \mathcal{A}_x' \).

Any class \( \{P_\beta\} \) singled-out by condition (b) is said to agree with the conditional probability \( P \).

The proof of the equivalence between conditions (a) and (b) gives rise to an algorithm to test the coherence of the assessment \( P \), based on the equivalence between condition (b) and the compatibility of a sequence of systems \( (\mathcal{S}_x) \): precisely, all systems of the following sequence, with unknowns \( x_r^\beta = P_\beta (A_r) \geq 0, A_r \in \mathcal{A}_\beta \), and \( \beta = 0, 1, 2, \ldots, k \leq n \), are compatible:

\[
(S_\beta)
\begin{cases}
\sum_{A_r \subseteq E_i \land H_i} x_r^\beta = P (E_i \mid H_i) \sum_{A_r \subseteq H_i} x_r^\beta, \\
\text{for all } E_i \mid H_i \in \mathcal{C} \text{ such that } \sum_{A_r \subseteq H_i} x_r^{\beta-1} = 0 \\
\sum_{A_r \subseteq H_0} x_r^\beta = 1
\end{cases}
\]

(put, for all \( H_i \)'s, \( \sum_{A_r \subseteq H_i} x_r^{-1} = 0 \) when \( \beta = 0 \)), where \( H_0^\beta = H_0 = H_1 \lor \cdots \lor H_n \), while \( x_r^{\beta-1} \) denotes a solution of \( (S_{\beta-1}) \) and \( H_0^\beta \) is, for \( \beta \geq 1 \), the union of the \( H_i \)'s such that \( \sum_{A_r \subseteq H_i} x_r^{\beta-1} = 0 \).

As proved in the aforementioned papers, conditions (a) and (b) are equivalent also to the following de Finetti’s coherence—as expressed, for example, in Lehman (1955)—where \( p_i = P (E_i \mid H_i) \):

(c) for any choice of the real numbers \( \lambda_1, \ldots, \lambda_n \) we must have

\[
\sup_{A_r \land H_0} \sum_{i=1}^n \lambda_i H_i (E_i - p_i) \geq 0,
\]

where \( H_0 = \sqrt[n]{1} H_i \). The random quantity

\[ G = \sum_{i=1}^n \lambda_i H_i (E_i - p_i) \]
can be interpreted as the gain corresponding to a combination of \( n \) bets of amounts \( \lambda_1 p_1, \ldots, \lambda_n p_n \) on \( E_1|H_1, \ldots, E_n|H_n \), with arbitrary stakes \( \lambda_1, \ldots, \lambda_n \).

We recall also this noteworthy result, discussed at length in Coletti and Scozzafava (2002a, Section 11.3):

**Theorem A2.** Coherence of an assessment \( P(\cdot|\cdot) \) on an infinite set \( \mathcal{E} \) of conditional events is equivalent to coherence on any finite subset \( \mathcal{C} \) of \( \mathcal{E} \).

It follows that Theorem A1 can be applied also if the given family of conditional events is not finite.

Given a family \( \mathcal{C} \) of conditional events \( \{E_i|H_i\}_{i \in I} \), where \( \text{card}(I) \) is arbitrary and the events \( H_i \)'s are a partition of \( \Omega \), we recall the following two corollaries of the characterization Theorem A1:

**Corollary A1.** Any function \( f : \mathcal{C} \rightarrow [0, 1] \) such that \( f(E_i|H_i) = 0 \) if \( E_i \wedge H_i = \emptyset \) and \( f(E_i|H_i) = 1 \) if \( H_i \subseteq E_i \) is a coherent conditional probability.

**Corollary A2.** If \( P(\cdot|\cdot) \) is a coherent conditional probability such that \( P(E|H_i) \in \{0,1\} \), then the following two statements are equivalent:

(i) \( P(\cdot|\cdot) \) is the only coherent assessment on \( \mathcal{C} \);

(ii) it is \( H_i \wedge E = \emptyset \) for every \( H_i \in \mathcal{H}_0 \) and \( H_i \subseteq E \) for every \( H_i \in \mathcal{H}_1 \), where \( \mathcal{H}_r = \{H_i : P(E|H_i) = r\} \), \( r = 0, 1 \).

Concerning coherence, another fundamental result is the following, essentially due—for unconditional events, and referring to an equivalent form of coherence in terms of betting scheme—to de Finetti (1949).

**Theorem A3.** Let \( \mathcal{X} \) be any family of conditional events, and take an arbitrary family \( \mathcal{C} \subseteq \mathcal{X} \). Let \( P \) be an assessment on \( \mathcal{C} \); then there exists a (possibly not unique) coherent extension of \( P \) to \( \mathcal{X} \) if and only if \( P \) is coherent on \( \mathcal{C} \).

Just as the corresponding theorem for unconditional assessments is a generalization—to a coherent \( P \)—of the well-known theorem—see Horn and Tarski (1948)—that every finitely additive probability on a subalgebra of a Boolean algebra can be extended to the whole algebra, Theorem A3 generalizes—to a coherent conditional \( P \)—the full analogue of Horn–Tarski theorem established by Krauss (1968, Theorem 3.6), concerning the possibility of extending a finitely additive conditional probability from \( \mathcal{A} \times \mathcal{A}^0 \)—where \( \mathcal{A} \) is an algebra—to \( \mathcal{B} \times \mathcal{B}^0 \), with \( \mathcal{B} \) any algebra such that \( \mathcal{B} \supseteq \mathcal{A} \).

In other words, Theorem A3 points out a further relaxing of the framework—besides that obtained by a finitely additive setting with respect to a countably additive one—since it refers to the weaker notion of coherent (conditional) probability.

**References**


Frank, M.J., 1979. On the simultaneous associativity of $F(x,y)$ and $x + y - F(x,y)$. Aequationes Math. 19, 194–226.