From Information to Probability
An Axiomatic Approach

Inference is Information Processing

Wilhelm Rödder[1] and Gabriele Kern-Isberner[2]

[1, 2] FernUniversität in Hagen, Germany
[1] Lehrstuhl für BWL, insbesondere Operations Research, FB Wirtschaftswissenschaft
[2] Lehrstuhl für Praktische Informatik VIII – Wissensbasierte Systeme, FB Informatik

Abstract: We define the very rich language of composed conditionals on a three-valued logic and use this language as the communication tool between man and machine. Communication takes place for three reasons: Knowledge acquisition, query and response. Learning, thinking and answering questions are of pure information theoretical nature. The pivot of this knowledge processing concept is the amount of information in [bit] which we receive, if we learn a conditional to become true. We follow an axiomatic approach to information theory rather than the classical probabilistic approach of Shannon; information comes first, then comes probability. In the light of this philosophy query and response experience new interpretations. Both, acquisition and response are realized by maximizing entropy and minimizing relative entropy, respectively. The iterative solution of these mathematical optimization problems gives new insights into the adaptation of prior knowledge to new information. The expert system shell SPIRIT supports this kind of knowledge processing which will be demonstrated by suitable examples.

Key words: AI, conditionals, entropy, inference, information, knowledge, SPIRIT.

1. Introduction

Probability was first, then came information. Humans like to gamble and feel a strong desire to predict the outcomes of games. History of probability theory is the history of calculating chances in hazards. The Italian scientists Cardano and Galileo already in the 16-th century studied playing at dice, the French mathematicians Pascal and Fermat founded combinatorics and the Dutchman Huygens in 1657 published his work "Tractatus de ratiociniis in ludo alea". Leibniz (De incerti aestimatione), Bernoulli (Ars conjectandis) and Montmort (Essay d'ana-

lyse sure les jeux de hazard) continued the analysis of chances and began to anticipate future developments of strategies in games.

Perhaps the English clergyman Bayes (1702 – 1761) was the first to systematically improve the estimation of probabilities by new information. In the 18-th and the first decade of the 19-th century de Laplace, Gauß, Legendre and Poisson smoothed the way for Francis Galton, Karl Pearson, William S. Gosset and Sir Ronald A. Fisher. These scientists founded the English school of statistics and their influence on modern probability theory is enormous, just as that of the Russian scientists Chebychew, Markov, Ljapunov and Kolmogorov. The axiomatic development of Kolmogorov joined probability theory and measure/ integration-theory and so made it "presentable at court" for pure mathematics.

Probability during 400 years was always prediction: What is the degree of belief that an event will happen? Probability theory was always oriented towards future and never looked back after the game was over. Seldom or never the question was "What will I have learned if this specific event realized?" It was the merit of Hartley [HAR, 1928] and Shannon [SHA, 1948] to focus things this way. Their intension and especially that of Shannon was the maximization of the information rate of messages transmitted from a source to a receiver via a channel. To maximize this rate, a suitable codification of the messages was necessary and then the average amount of information "we would have learned if we received these messages" is maximum. For an ergodic source, entropy measures the average information rate which the receiver "learns".

The entropy ($0 \cdot ld\, 0 = 0$ and ld the dual logarithm)

$$H = -\sum_{v} P(v) \cdot ld\, P(v) \qquad (1)$$

is a function of the signals' or messages' v probabilities and as such a function of a probability distribution.

Probability theory was first, then came information theory!

History could have happened the other way around. What if people already in the 17-th century would have asked "How much will I have learned?".

A part of (human) knowledge and intelligence is the ability to condition things to each other. If in our memory event A conditions event B, we can conclude B from learning that A evidently is true. The more such conditions we memorize the more we know. Highly connected conditioning chains are part of our intelligence.

If we learn an event to be true which we very likely believed to be true beforehand, little we learn. If it was un-

expected, much information we get. Learning and concluding things from memorized conditionals and from evident facts is information processing rather than a probabilistic approach!

In this paper information is first, then comes probability.

To develop information processing in a precise syntactical and semantical context we first justify the use of composed conditional events in section 2.1 and then give the mathematical prerequisites for the three-valued conditional logic on these conditional events in 2.2. In section 3.1 we axiomatically introduce information as a real-valued measure on the set of all conditionals and in section 3.2 we derive probability from information. We dedicate section 4 the information theoretic process of knowledge acquisition, query and response. Subsection 4.1 prepares notations, subsection 4.2 developes knowledge processing as a transformation process of information measures on the set of conditionals and gives a small example, which was calculated by the expert system shell SPIRIT. In sections 4.3 and 4.4 knowledge adaptation undergoes sophisticated information theoretical interpretations. Section 5 concludes this paper and refers to the applicability of the new knowledge processing to large decision problems.

2. Semantics, Syntax and Logic of Conditionals

2.1 Semantical Preliminaries

In this paper we study propositions and conditioned propositions about individuals or objects from a vague population of which we have incomplete information, only. A vague population is something like "all humans" or "all animals" or even worse "all that is describable by certain properties". It corresponds to our perception or even contemplation and is never countable piece by piece. About such vague populations we collect information, by induction and analogies and by instructions from authorities. Such information usually is of the "if-then" type, it is conditioned information about the objects of the vague population.

Very often we learn provisional things (default rules) and only later, when contraditions arise or when wrong estimations cause dissatisfaction, we specify further by conditioning. The default that birds fly, for the first time may be abandoned in the zoological garden with a flock of ostriches. A hierarchy of conditioned conditionals specify and partionate our knowledge more and more. New characterizing attributes arise, the vague population

augments each day when our horizon widens. And all this works very well.

The focus of our paper is quite different from the one realized in Bayes nets, where a population's distribution is fully determined by conditional structure and conditional probabilities, c.f. [JEN, 1994]. We feel that in this case there is not enough room for inference of propositions from other propositions. There is only a technically brilliant but unpretentious calculation of conditional probabilities in a fully determined contingency table.

Inference is more. Inference is the result of presumption and logical entailment about the vague population of our perception or even contempletation.

Inference takes place in spite of incomplete information about this population. Knowledge adaptation is inference and answering questions is based on inference – if it is not the mere repetition of something we have learned by heart, earlier.

So the inference process developed here is of a highly subjective nature and reflects our "as for as I know". Knowledge includes the ability to predict propositions. Knowledge is mutual determination of propositions by propositions. The more determination, the less surprise if something really happens. And these mutual connections, which can most adequately be expressed by conditionals, are in continuous transition, as new information improves our knowledge. (cf. also [KIS, 2001])

The next subsection provides the conditional logic's syntax which enables a semantically sound knowledge adaptation and inference process.

2.2 Syntax and Conditional Logic

The syntax consists of a set of finite valued variables $V=\{V_1,\ldots,V_L\}$ and their respective values v_l of V_l. It includes the connectives $\bar{}$ (not), $\wedge$ (and), $\vee$ (or) and also the conditional operator | (given), as well as brackets and suitable syntactical rules. With elementary formulas VARIABLE = value and the above connectives we describe events of the event field $(\Omega, \emptyset, \bar{}, \wedge, \vee)$ on elementary events $v=v_1\ldots v_L$, being Ω the all – and $\emptyset$ the empty event. We do not distinguish between formulas and events, further on.

Events in Ω (or the corresponding formulas) are denoted by capital letters, indexed if necessary: $A,B,\ldots,A_i,B_i,\ldots,E_j,F_j$. They build the language L.

Events from L can be conditioned by the conditional operator |: $B|A$. $B|A$ is called a conditional. The set of all con-

ditionals build the language $L|L$.

From a logical point of view, the formula $V_l=v_l$ is a literal. A literal is an atomic proposition which is true (t) or false (f) for an object or individual of the population. The same holds for any formula $A,B,\ldots,A_i,B_i,\ldots,E_j,F_j$. Canonical conjuncts such as $V_1=v_1 \wedge \ldots \wedge V_L=v_L$ are written as $v=v_1\ldots v_L$, for short. They are the finest pattern to identify objects and therefore often are called worlds. Negation is indicated by barring the corresponding proposition.

If a world v is implicant of $A, v \subset A$, we write $A(v)=t$.

If a world v is implicant of $\overline{A}, v \subset \overline{A}$, we write $A(v)=f$.

If $A(v)=t$ ($A(v)=f$) for all v, we say $A=t (A=f)$.

Following Calabrese [CAL, 1991] and earlier de Finetti [FIN, 1972] we define for a conditional $B|A$:

$$B|A(v)=\begin{cases} t & v \subset BA \left(= B \wedge A\right) \\ f & v \subset \overline{B}A \left(= \overline{B} \wedge A\right) \\ u & v \subset \overline{A}. \end{cases} \tag{2}$$

A conditional might be true or false for a true premise, and otherwise it is undefined. We say that $B|A=t$ ($B|A=f$), if $B|A(v)$ is true (false) in all worlds.

If $A=t$ we write $B|t$ instead of $B|A$. So the set of all conditionals $B|A$ includes the set of all (unconditioned) conditionals $B|t$ which is isomorphic to the original event field $(\Omega, \phi, \bar{}, \wedge, \vee)$. We often call such $B|t$ facts rather than conditionals. A conditional $B|A$ is called contradictory, if $AB=f$.

We shall now define the negation, conjunction, disjunction and the conditioning of conditionals. Doing so, we make use of a boolean extension as it was suggested by Calabrese, following the three-valued logic in Rescher [RES, 1969]. Other extensions are possible, cf. [DUP, 1990], [DUP, 1991]. Mind the fact that for conditioning of conditionals we differ slightly from Calabrese.

Let $\bar{},\wedge,\vee$ behave on $\{t,f\}$ as usual and let

$$\overline{u}=u, t \wedge u = t \vee u = t,\ f \wedge u = f \vee u = f,\ u \wedge u = u \vee u = u\ . \tag{3}$$

Let the conditional operator behave like follows

$$t|t = t,\ f|t = f,\ \ t|u = f|u = u|u = u|f = f|f = t|f = u|t = u. \tag{4}$$

With these conventions and with the respective syntax rules a highly complex hierarchy of composed condition-

als is available.

For that purpose define "pointwise"

- $\overline{B|A}(v) = \overline{B|A(v)}$
- $[(B|A) \wedge (D|C)](v) = (B|A)(v) \wedge (D|C)(v)$ (5)
- $[(B|A) \vee (D|C)](v) = (B|A)(v) \vee (D|C)(v)$
- $[(B|A)(D|C)](v) = (B|A)(v)(D|C)(v)$.

Mind the fact the right hand sides can be evaluated because of (3), (4). The pointwise assignment in (5) then defines composed conditionals on $L|L$. They might be generalized to any hierarchical level, respecting certain syntax rules.

Such syntactical formulas allow a very rich linguistic semantics on the population of all individuals. The reader is invited to construct examples of such connected propositional expressions or study [CAL, 1991], [ROD, 2000].

Despite the linguistic richnes of $L|L$ it is often convenient to reduce complex propositional conditional formulas to simple ones, i. e. to conditionals of the form $B|A$ and $A,B \in L$. This reduction always is possible because of the equalities (6), cf. [CAL, 1991] and easy calculations. In detail, we have

- $(B|A) \wedge (D|C) = [(B \vee \overline{A})(D \vee \overline{C})] \mid A \vee C$
- $(B|A) \vee (D|C) = (AB \vee CD) \mid A \vee C$ (6)
- $(B|A) \mid (D|C) = B \mid ACD.$

Keep in mind, however that the formulas under such reduction loose their intelligibility and linguistic clearness. Conditional Basic Systems CBSs are sets of conditionals which "cover" $L|L$ and represent its overall conditional structure. A CBS is a basis of all conditionals similar as is a basis of a vector space. For this reason a CBS must "span" $L|L$.

Definition 1 (c-independent and disjoint conditionals)

i) $B|A$ is conditionally or c-independent of $D|C$ iff $(B|A) \mid (D|C) = B|A$.

ii) $B|A$ and $D|C$ are disjoint iff $ABCD=f$.

It is straightforward to prove that $B|A$ is c-independent of $D|C$ iff $A\subset CD$.

Definition 2 (Conditional Basic System)

A set S of noncontradictory conditionals is a conditional basic system CBS, iff for any pair $B|A$, $D|C$ from S one of the following statements holds:

- $B|A$ is c-independent of $D|C$
- $D|C$ is c-independent of $B|A$
- $B|A$ and $D|C$ are disjoint,

and if each $v|t$ is a conjunction of c-independent elements from S.

Observation 1

Let S be a CBS. Then for each world v, $v|t$ can be written as $v|t=\bigwedge_{\substack{f|e\in S\\ v\subset fe}}$.

Proof: By definition 2, $v|t=\bigwedge_{i=1}^{m} f_i|e_i$ where the $f_i|e_i$ are noncontradictory elements from S, a CBS, and for $i\neq j$ $f_i|e_i$ is c-independent of $f_j|e_j$ or vice versa. Let w.l.o.g. the indices be so that $f_i|e_i$ is c-independent of $f_{i+1}|e_{i+1}$ which implies $e_i\subset e_{i+1}f_{i+1}$, cf. the last equation in (6). Because of this inclusion we get immediately

$$\begin{aligned} v|t &= \bigwedge_{i=1}^{m} f_i|e_i \\ &= \left(\bar{e}_1e_2f_2 \vee \bar{e}_2e_3f_3 \vee \ldots \vee \bar{e}_{m-1}e_mf_m \vee e_1f_1e_2f_2\ldots e_mf_m\right)|e_m \\ &= \left(\bar{e}_1e_2f_2 \dot{\vee} \bar{e}_2e_3f_3 \dot{\vee} \ldots \dot{\vee} \bar{e}_{m-1}e_mf_m \vee e_1f_1\right)|e_m. \end{aligned}$$

This can hold only if $e_m=t$ and if $v=\bar{e}_1e_2f_2 \dot{\vee} \ldots \dot{\vee} \bar{e}_{m-1}e_mf_m \vee e_1f_1$. Since $e_1f_1\neq f$, cf. definition 2, we have $\bar{e}_1e_2f_2=\ldots=\bar{e}_{m-1}e_mf_m=f$ and $v=e_1f_1\subset e_2f_2\subset\ldots\subset e_mf_m$.

In particular, $v\subset e_if_i$ for all i, $1\ldots m$.

Example 1 shows different CBSs. Either one will be a suitable basis to measure information and uncertainty inherent in $L|L$'s conditional structure.

Example 1 (Conditional Basic Systems)

i) $\{v = v_1 \ldots v_L\}$ is a CBS.

ii) $\{v_1\} \cup \{v_2|v_1\} \cup \ldots \cup \{v_L|v_{L-1} \ldots v_1\}$ is a CBS.

iii) $\{v_l \ldots v_1\} \cup \{v_L \ldots v_{l+1}|v_l \ldots v_1\}$ is a CBS.

iv) ii) and iii) for any permutation $l_1 \ldots l_L$ of $1 \ldots L$ is a CBS.

To verify e.g. ii) in example 1 use definition 1, definition 2, the formula for the conditioned conditional in (6) and prove the resolution into factors $v = v_1 \wedge (v_2|v_1) \wedge \ldots \wedge (v_L|v_{L-1}, \ldots, v_1)$.

Every conditional $B|A$ is equal to $BA|A$, see (2). Both propositions, BA and A, obviously can be expressed by their canonical disjuncts v and those, in turn, as conjuncts of elements from any CBS S, cf. definition 2.

This implies the equation

$$B|A = \left(\bigvee_{v \subset BA} \bigwedge_{\substack{f|e \in S \\ \text{independent} \\ v \subset fe}} f|e \right) \Big| \left(\bigvee_{v \subset A} \bigwedge_{\substack{f|e \in S \\ \text{independent} \\ v \subset fe}} f|e \right) \tag{7}$$

Note that we write elements of a CBS as small letters, if convenient. There should be no problem with this notation.

Equation (7) allows the decomposition of any $B|A$ into its basic conditionals. We shall need this property further on. Given all semantical, syntactical and logical prerequisites, we are now ready to study information measures on conditionals.

3. From Information to Probability

3.1 Conditionals and Information

If we know that A is true, how much do we learn coming to know that B is true? In other words: How much information do we get from learning that an arbitrary object in our population with property A has also property B and how much does this information count? We intuitively agree that this information is some real number and that the system of such real numbers on $L|L$ should have some obvious properties:

- they should be nonnegative,
- if A is implicant of B, no additional information should result from learning that B is true, given A is true,
- if $B|A$ is c-independent of $D|C$ or vice versa, informa- tion(s) of learning that $B|A$ *and* $D|C$ are true should add,
- it should meet our day by day experience like if somebody "informs us that in a hand of cards there is a spades' ace".

These characteristics can be formulated as axioms. We call a real-valued function *inf* on $L|L$ an information measure if it fulfills the following axioms:

A1 $inf\left(B|A\right) \geq 0$ for *all* $B|A \in L|L$,

A2 $inf\left(B|A\right) = 0$ for $A \subset B$,

A3 $inf\left(\left(B|A\right) \wedge \left(D|C\right)\right) = inf\left(B|A\right) + inf\left(D|C\right)$ for $B|A$ c-independent of $D|C$ or $D|C$ c-independent of $B|A$,

A4 $k^{-inf\left(\bigvee_i B_i|A\right)} = \sum_i k^{-inf\left(B_i|A\right)}$ for all sets of pairwise disjoint B_i and for some $k>1$.

A1 is evident for measures, A2 was explained earlier and so was A3. The crucial point is A4. We shall make a special effort to show that A4 meets our intuition and is not just a tricky deviation to probability theory.

We first derive a straight forward consequence of A3. $B|A$ is c-independent of A and $(B|A)\wedge A=BA$. Hence from A3 we have immediately $inf(BA)=inf(B|A)+inf(A)$ and thus $inf(B|A)=inf(BA)-inf(A)$. Dividing A4 by $k^{inf(A)}$ yields the unconditioned form

A4' $k^{-inf\left(\bigvee_i B_i\right)} = \sum_i k^{-inf\left(B_i\right)}$.

As $k^{-inf}=k^{-a\cdot inf}$ for $a=\log_k k$ changing the basis in (A4') means only changing the scale of *inf*. So the choice $k=2$ is not restrictive and it suffices to justify A4 in the binary and unconditioned form

A4" $2^{-inf\left(\bigvee_i B_i\right)}=\sum_i 2^{-inf(B_i)}$.

Each B_i is a proposition and *inf* is intended to measure the information we get from learning that B_i is true. The more "laborious" the description of B_i the higher the information if we learn it to be true. Consider the special case $\Omega=\{\omega=(b_1,\ldots,b_L)$ and $b_l=0\vee 1\}$. Let the B_i be subsets of Ω built by fixing some of the variables, $b_l=0$ or $b_l=1$, and leaving the others free, $b_l=*$. The information we need to determine whether or not some ω lies in such a B_i equals the number $inf(B_i)$ of fixed elements in B_i, by intuition. Mind the fact that the number of elements in a B_i is $2^{*(B_i)}$ where $*(B_i)$ counts the free variables. Therefore, $inf(B_i)=L-*(B_i)$.

If now B and B_i are such sets and if $B=\bigvee B_i$ is a disjunction of pairwise disjoint B_i, then the equation

$2^{*(B)}=\sum_i 2^{*(B_i)}$ and

$$2^{L-\inf B}=\sum_i 2^{L-\inf B_i}\ , \qquad (8)$$

respectively, is evident. Dividing by 2^L yields (A4").

We esteem equation (8) an intrinsic property of any abstract information measure *inf* and it should even hold in the absence of a naïve integer information count as in the above example.

Note that A1-A4 determine the inf-measures' properties like nonnegativety and some sort of additivity. In general there is an infinite number of measures on $L|L$. The measure becomes fixed if fixed for all $v\subset\Omega$, for instance. A3 and A4 then permit the calculation of $inf(B|A)$ for any $(B|A)$. As the determination of all $inf(v)$ is a laborious thing, we relate on a better manner to built an inf-measure in section 4.

There are some immediate consequences of A1-A4. The proofs are obvious and so we leave them to the reader.

Lemma 1 (Properties of *inf*)

i) $inf(\Omega) = 0,\ inf(\emptyset) = \infty$.

ii) $inf(A) = inf(\overline{A})$ implies $inf(A) = inf(\overline{A}) = 1$

iii) From $B \supset A$ we get $inf(B) \leq inf(A)$.

The occurrence of the all-event supplies no information at all, the impossible event would yield an infinite amount of information if occurring, *inf* is monotonous with respect to inclusion and information of equally informative complementary events measures 1. This information unity we call [bit].

3.2 Information and Probability

The higher $inf(B|A)$ the more we learn when $B|A$ becomes true. We learn much from the unexpected, unlikely or unpro-bable, we learn little from the expected, likely or probable. In some idioms, as in the native tongue of these authors, pro-bable = "wahrscheinlich" has the downright meaning "seems to be true". There should be a reverse measure to *inf* which expresses this likelihood to be true.

A natural property of this reverse measure would be the additivity on disjoint events, we feel, and its mere functional dependency on *inf*; furthermore we shall opt for normali-sation. This gives rise to the following.

Definition 3 (Probability)

i) If the function h is such that for any B_1, B_2 with $B_1 B_2 = f$, $h(inf(B_1 \vee B_2)) = h(inf(B_1)) + h(inf(B_2))$ and $h(inf(\Omega))=1$, then $h(inf(B))$, all B, is called the *inf*-corresponding probability measure on L and is written as *pinf*.

ii) For h like in i) $h(inf(B|A))$ is called the conditional's probability or the (conditioned) probability of B given A.

Definition 3 determines for any *inf*-measure the functional relation between *inf* and *pinf*, which we show in the next theorem.

Theorem 1 (Relation between *inf* and *pinf*)

For any measure *pinf* obeying definition 3i) and ii) the following statements hold:

i) $pinf(B) = 2^{-inf(B)}$, all $B \in L$.

ii) $pinf(B|A) = pinf(BA) / pinf(A)$, all $B|A \in L|L$.

Proof: i) For disjoint B_1, B_2 we have from axiom A4 $inf(B_1 \vee B_2) = -ld\left(2^{-inf(B_1)} + 2^{-inf(B_2)}\right)$

and from definition 3 i)

$h(inf(B_1 \vee B_2)) = h\left(-ld\left(2^{-inf(B_1)} + 2^{-inf(B_2)}\right)\right) = h(inf(B_1)) + h(inf(B_2))$. Now taking $g(\cdot) = h(-ld(\cdot))$ we conclude

$g\left(2^{-inf(B_1)} + 2^{-inf(B_2)}\right) = g\left(2^{-inf(B_1)}\right) + g\left(2^{-inf(B_2)}\right)$.

This functional equation on $IR^+ x IR^+$ has the solution $g(\cdot)=a \cdot id(\cdot)$ and hence $h(\cdot)=a \cdot 2^{-(\cdot)}$.

Because of $h(inf(\Omega))=a \cdot 2^{-0}=1$ we have $a=1$. This concludes the proof of i).

ii) $h(inf(B|A)) = 2^{-inf(B|A)} = 2^{-inf(BA)+inf(A)} = 2^{-inf(BA)} / 2^{-inf(A)} =$

$pinf(BA) / pinf(A)$

This concludes the proof of ii).

Because of theorem 1 there is a one to one correspondence between *inf* and *pinf*, respectively. *pinf* is nonnegative, additive and normalised and thus it is a probability measure. Sometimes we write *infp* instead of *inf*, if the focus is on the information related to the probability measure p. Note that $infp(B|A) = - ld\, p(B|A)$ is an information measure for any p.

Perhaps this is the right place to emphasize that every fact we learned in elementary (finite) probability theory has its counterpart in information theory, of course. Conditional probability now is conditional information, the law of Bayes appears in a new shape and p-independence now is *inf*-independence. The last property reads e.g.: B is *inf*-independent of A iff $inf(B|A)=inf(B)$. And this makes sense!

Any information measure *inf* attributes a certain amount of [bit] to a $B|A \in L|L$. The higher this value the less we expected $B|A$ to become true. What is the average uncertainty in *inf*? This will be the focus of the next section.

3.3 Uncertainty

$inf(B|A)$ for any $B|A \in L|L$ measures the information we receive when we come to know that B is true given A (is true). The less we learn the more certain we were beforehand; the more we learn the less certainty we had. There is an overall intrinsic uncertainty in the measure inf on $L|L$.

Because of equation (7) at the end of section 2.2 $inf(B|A)$ is calculable for each $B|A$, if the measure is fixed on any CBS. Any conditional basic system suffices to determine inf on the whole $L|L$. Note that in (7) every BA and every A is represented as a disjoint disjunct of independent conjuncts of elements from CBS. So because of axioms A3 and A4 $inf(B|A)$ is available, if only inf is available for the CBS.

Therefore, to measure the overall uncertainty on $L|L$ we should restrict our attention to a(n arbitrary) CBS.

The considerations so far give rise to the following definition.

Definition 4 (Uncertainty of inf)

Let inf and $pinf$ be an information and its related probability measure on the CBS S, respectively. Then

$$\sum_{f|e \in S} pinf(fe) \cdot inf(f|e) \tag{9}$$

is the uncertainty in inf.

The following observations permit a better understanding of definition 4.

- From $f|e$ we receive $inf(f|e)$ if $f|e \wedge e$ becomes true or if fe becomes true, respectively.
- If $f|e$, $h|g$ are disjoint, then fe and hg are disjoint (cf. definition 1). Consequently (9) contains a weighted sum of their corresponding information, as it should be.
- If either $f|e$ is c-independent of $h|g$ or $h|g$ is c-independent of $f|e$, the weighted information should also appear in (9), even if they are not disjoint. More on that in the proof of the next theorem.
- Uncertainty measures in [bit]

Definition 4 makes sense only if uncertainty does not depend on the special CBS. This is garanteed by the following theorem.

Theorem 2 (Uniqueness of uncertainty)

The uncertainty in *inf* as in formula (9) is the same for any CBS *S* and equals $\sum_{v|t} pinf(v|t)\cdot inf(v|t)$.

Proof: Let *S* be an arbitrary CBS. Due to Observation 1 we have

$$\begin{aligned}
&\sum_{v|t} pinf(v|t)\cdot inf(v|t) = \sum_{v|t} pinf(v|t)\cdot inf\left(\bigwedge_{\substack{f|e\in S \\ v\subset fe}} f|e\right) \\
&= \sum_{v|t} pinf(v|t)\cdot \sum_{v\subset fe} inf(f|e) = \sum_{v|t}\sum_{v\subset fe} pinf(v|t)\cdot inf(f|e) \\
&= \sum_{f|e\in S} pinf(fe)\cdot inf(f|e)
\end{aligned}$$

Since the CBS was chosen arbitrarily, this concludes the proof.

Note that the $v=v|t$ are the elementary (unconditioned) events in Ω and so $\sum_{v|t} pinf(v|t)\cdot inf(v|t)$ can be written as

$$\sum_{v} pinf(v)\cdot inf(v), \tag{10}$$

which is the well-known entropy of *pinf* on Ω, as $inf(v)=-ld\ pinf(v)$ from theorem 1. The entropy measures in [bit].

(10) measures uncertainty in any CBS – and so in $L|L$ – only on elementary events, no real conditional is involved. This might cause uneasiness, but this uneasiness should be prevented by the proof of theorem 2 and by example 2.

Example 2 (Conditional structure in *inf* or *pinf*)

Consider again the CBS of example 1 ii): $\{v_1\}\cup\{v_2|v_1\}\cup\ldots\cup\{v_L|v_{L\text{-}1}\ldots v_1\}$. Because of theorem 2 (or by direct verification) we have

$$\begin{aligned}
&\sum_{v} pinf(v)\cdot inf(v) = \sum pinf(v_1)\cdot inf(v_1) + \\
&\sum pinf(v_1v_2)\cdot inf(v_2|v_1) + \ldots + \sum pinf(v_L\ldots v_1)\cdot inf(v_L|v_{L-1}\ldots v_1)
\end{aligned}$$

This can be developed further to

$$\sum pinf(v_1)\cdot inf(v_1)+\sum_{v_1} pinf(v_1)\sum_{v_2} pinf(v_2|v_1)\cdot inf(v_2|v_1)+\ldots+$$
$$\sum_{v_{L-1}..v_1} pinf(v_{L-1}..v_1)\sum_{v_L} pinf(v_L|v_{L-1}..v_1)\cdot inf(v_L|v_{L-1}..v_1)$$

The last expression in example 2 intuitively represents conditional structure in *inf* better than (10), as it involves conditionals. But it is the same!

Information was first, then came probability. (10) cha-racterises uncertainty in either measure. We are now prepared to process in section 4.2 information measures for the purpose of learning and answering questions from the acquired knowledge, subsection 4.1 provides preliminaries.

4. Information, Knowledge and Inference

4.1 Uncertainty Change

Theorem 2 of section 3.3 is essential for the inference process to be developed here. Remind the fact that for any CBS its inherent uncertainty is the same and equates entropy. We now assume that – for whatever reason – the information measure *infp* has changed to *infq*. With this change the average uncertainty is altered, too. As it suffices to consider p.e. the CBS $\{v\}$, this average uncertainty change is

$$\sum qinf(v)\cdot(infp(v)-infq(v)). \tag{11}$$

The information *divergence* for any v to become true, measures $infp(v)$-$infq(v)$. And in the light of the new information measure the *average* change is (11).

In the set of all *inf*-measures there is a distinguished element, denoted $infp^0$. $infp^0$ is the one with maximum uncertainty:

$$infp^0 = \arg\max_{qinf} \sum qinf(v)\cdot infq(v). \tag{12}$$

It is well-known, that for this $infp^0$ all $infp^0(v)$ are equal and measure $infp^0(v)=ld|\Omega|$, all v. In $infp^0$ we learn an equal amount of [bit] for any v to become true. Maximum uncertainty corresponds to minimum knowledge. The more uncertainty reduction is still possible in an *inf*-measure, the less we know! $infp^0$ corresponds to absolute ignorance.

The average uncertainty change against ignorance is of great importance.

$$\sum qinf\left(infp^{0}(v) - infq(v)\right) \tag{13}$$

measures average uncertainty in *infq* on an absolute scale. If it is 0, nothing we know and maximum uncertainty reduction is still possible. If it is high, much we have learned and little uncertainty reduction is expected. (13) measures adapted knowledge in *infq* in [bit].

With these reflections we should be able to develope the concept of knowledge acquisition and inference under information fidelity.

4.2 Knowledge Processing under Information Fidelity

Resuming, we dispose of the language $L|L$ on Ω and a set of information measures on $L|L$. To determine such an information measure, needs the specification of not less than $|\Omega|$ real numbers which must satisfy the axioms A1-A4. This is inconvenient even for a modest cardinality of Ω. Besides this technical argument, totally fixing an *infq* on $L|L$ does not meet our conceptual understanding of thinking and reasoning. The reader is asked to go back to be beginning of section 2 and repeat the there explanations about vague populations, for which only partial information is available.

In the information theoretical context "partial information" means that for some conditionals $B_i|A_i, i=1\ldots I$, we know their uncertainty in [bit], for others we don't. There is still a high degree of freedom to choose an appropriate uncertainty measure for the remaining conditionals. To do so we must *extend* *infq* from $infq(B_i|A_i), i=1\ldots I$, to the whole $L|L$. And then we can answer any question $inf(H|G)=?$

Even worse, very often the incomplete knowledge about the vague population must be reflected in the light of temporary situative assumptions, different from such knowledge. What can be answered to the question $inf(H|G)=?$ in such a case. This is the subject of the present section.

The knowledge processing concept we propose here consists of three steps: Knowledge Acquisition, Query and Response.

Knowledge Acquisition

The incomplete knowledge about the population is provided as a set of I conditionals $B_i|A_i, i=1\ldots I$, and their respective uncertainty rk_i. More precisely, the rk_i represent an estimation of how much uncertainty reduction in [bit] we expect learning that an arbitrary object of the population with property A_i is also B_i. To require such estimations

perhaps is unusual. Keep in mind, however, that the estimation of a subjective probability is equivalent. The latter we believe to do better as a mere consequence of habit.

To infer from these knowledge pieces solve the optimization problem

$$\begin{aligned} infp^{*} &= \arg\min \sum_{v} qinf(v) \cdot \left(infp^{0}(v) - \inf q(v) \right) \\ s.t. \quad & infq\left(B_i \middle| A_i\right) = rk_i, i = 1...I. \end{aligned} \tag{14}$$

*infp** is the information measure which adapts ignorance to the given knowledge pieces and does not create any additional knowledge. Respecting the provided facts, (14) seeks a minimal uncertainty change and hence adapts knowledge under information fidelity. *infp** is a correct extension on $L|L$, it is the knowledge base. This is what knowledge acquisition is concerned with.

Query

The incomplete knowledge about the evident temporary situation consists of J conditionals $F_j|E_j, j=1\ldots J$, and their respective uncertainty rs_j. The rs_j measure in [bit] the assumed temporary uncertainty reduction, different from the one acquired in *infp**. As the assumption is temporary this information is not to be acquired but abandoned after the query is over! We reflect basic knowledge in view of evidence to answer a question.

To temporarily adapt the knowledge base *infp** to evidence, solve the optimization problem

$$\begin{aligned} infp^{**} &= \arg\min \sum_{v} qinf(v) \cdot \left(infp^{*}(v) - infq(v) \right) \\ s.t. \quad & infq\left(F_j \middle| E_j\right) = rs_j, j = 1...J. \end{aligned} \tag{15}$$

*infp*** is the information measure with minimum information theoretical divergence from *infp**. (15) adapts *infp** to the assumptions under information fidelity.

Response

Response to a question $inf(H|G)=?$ is the answer

$$\inf p^{**}(H|G) = ra. \tag{16}$$

The answer expresses in [bit] the concluded uncertainty in $H|G$. How much do we learn if we come to know that an object with property G also is H. The conclusion was entailed from basic knowledge *and* from evidence.

In either case (14) and (15), the reason for minimizing the uncertainty change against $infp^0$ and $infp*$, respectively, was intuitiv and reflected our resistance against not intended revision of former convictions. This intuition experiences a deep justification when studying knowledge revision in more detail in section 4.4.

The essential of this information processing (14), (15), (16), besides of its fidelity, is its measurability. Knowledge, remaining uncertainty, inferential strength and even significance of the answer are measurable in [bit]. With the notation $R(\inf q, \inf p) = \sum_v q \inf(v)(\inf p(v) - \inf q(v))$ the following statements detail our proposition.

- $R(infp*,infp^0)$ is information gain against ignorance and hence measures knowledge, cf. the end of section 4.1. It measures in [bit].

- $R(infp**,infp^0)$ is information against ignorance in the evident situation. It is ad hoc knowledge. Ad hoc know ledge might be more or less than basic knowledge. If it is less, basic knowledge is inadequate for the actual query. It measures in [bit].

- $R(infp**,infp*)$ is the strength of inference. The more the evident situation departs from basic knowledge, the higher the ("intellectual") effort to adapt it. It measures in [bit].

- Maximum knowledge $R(infp^\infty,infp^0)$ is reached for $infp^\infty(v)=0$, some v (and infinite for all remaining worlds). Then $(0\cdot\infty=0)$ we get $R(infp^\infty,infp^0)=ld|\Omega|$, c.f. section 4.1. Maximum knowledge is limited and so is the remaining uncertainty $ld|\Omega|-R(infp*,infp^0)$ in a knowledge base. It measures in [bit].

- $R\left(infp**\left(H|G\right),infp**\left(\overline{H}|G\right)\right),\left(infp*\left(H|G\right),infp*\left(\overline{H}|G\right)\right)$ is the directed divergence of the question's uncertainty under $infp**$ and under $infp*$. It is the significance which evidence puts upon the question and measures in [bit].

Of course, (14), (15), (16) are equivalent to respective optimization problems in probability space instead of information space. This fact is due to the one-to-one correspondence between $pinf$ and $infp$, c.f. theorem 1 in section 3.2. These probabilistic optimization problems were studied in earlier works like [RME, 1996], [ROD, 2000], [ROD, 2001]. The there developed iterations to solve them, go back to [CSI, 1975]. Besides the equivalence we write down the probabilistic versions here, too. This facilitates future considerations in the next section. So (14), (15), (16) correspond to (14'), (15'), (16'):

$$\begin{aligned} p*inf = \operatorname{argmin} \sum_v qinf(v)\left(inf\, p^0(v) - inf\, q(v)\right) \\ s.t. \qquad qinf\left(B_i|A_i\right) = x_i, \quad i=1..I, \\ and \; -ld\, x_i = rk_i, \quad i=1..I, \end{aligned} \tag{14'}$$

$$p^{**}inf = \operatorname{argmin} \sum_{v} qinf(v)(infp^{*}(v) - infq(v))$$

$$\begin{aligned} s.t. \quad & qinf(F_j | E_j) = y_j, \ j = 1..J, \\ & and \ -ld\, y_j = rs_j, \ j = 1..J, \end{aligned} \tag{15'}$$

$$p^{**}inf(H|G) = z \qquad and \ -ld\, z = ra. \tag{16'}$$

SPIRIT is a shell which supports both, the probabilistic and the information theoretical version of knowledge processing. In the remainder of this section we emphasize the latter. To do so, we now study the famous inference problem Lea Sombé, called after a group of logicists which evaluated different concepts of non-monotonous formalisms and published the results in [SOM, 1992].

Example 3 (Lea Sombé)

There are only four variables and their respective values characterising the society in which Lea Sombé lives: To be a student or not (*S=y/n*), to be young or not (*Y=y/n*), to have marital status single, married or corporate life *M=s/m/c*, to be parent or not (*P=y/n*).

If we learn that a student is young, we come to know little: 0.15 [bit]. To hear that a young person is single provides 0.33 [bit], and to hear in turn a single to be young yields 0.5 [bit]. By far not all young people are students and so we learn 1.7 [bit], if this is true for a young person. Little we learn from hearing that students with children are not single, 0.15 [bit]. And finally the information that members of the society living in corporate lifes, are young, counts 0.33 [bit]. That is all we know about the fictitious society, not more and not less. In terms of the present paper this knowledge reads:

$$\begin{aligned} & infq(Y = y | S = y) = 0.15 \ , \ infq(M = s | Y = y) = 0.33, \\ & infq(Y = y | M = s) = 0.50 \ , \ infq(S = s | Y = y) = 1.70, \\ & infq(M = m \vee M = c | S = y \wedge P = y) = 0.15 \ , \ infq(Y = y | M = c) = 0.33. \end{aligned}$$

The information measure *infp** derived from this partial information about the society by solving (14), is shown for all configurations

SYMP	*infp**	*SYMP*	*infp**	*SSYMP*	*infp**	*SYMP*	*infp**
yysy	7.88	*ynsy*	12.24	*nysy*	2.51	*nnsy*	3.34
yysn	3.06	*ynsn*	7.43	*nysn*	2.51	*nnsn*	3.34
yymy	6.44	*ynmy*	7.08	*nymy*	6.42	*nnmy*	3.53
yymn	6.97	*ynmn*	7.62	*nymn*	6.42	*nnmn*	3.53
yycy	5.62	*yncy*	10.37	*nycy*	5.60	*nnsy*	6.82
yycn	6.15	*yncn*	10.91	*nycn*	5.60	*nnsn*	6.82

This was what knowledge acquisition is concerned with.

We now use the knowledge base for a query and ask the system: What if Lea Sombé is a student and has a child, is she young? The query is performed by evidencing $S=y \wedge P=y$ and asking $Y=y$? More precisely, the evident situation is characterised by $infq(S=y \wedge P=y)=0$. If $S=y \wedge P=y$ by assumption is true, no information we get from learning that this in fact realises. Calculating $infp^{**}$ according to (15) provides the following table for all configurations:

SYMP	infp**	SYMP	infp**	SYMP	infp**	SYMP	infp**
yysy	3.39	ynsy	7.75	nysy	∞	nnsy	∞
yysn	∞	ynsn	∞	nysn	∞	nnsn	∞
yymy	1.95	ynmy	2.59	nymy	∞	nnmy	∞
yymn	∞	ynmn	∞	nymn	∞	nnmn	∞
yycy	1.13	yncy	5.88	nycy	∞	nncy	∞
yycn	∞	yncn	∞	nycn	∞	nncn	∞

Calculating $infp^{**}(Y=y)$ yields 0.30 [bit]. Remind that inf obeys axiom 4 of section 3.1 and verify the result. Reflecting all disposable knowledge about the fictitious society and assuming a person undoubtedly to be student and parent, leaves a considerable uncertainty about its youth, namely 0.30 [bit].

Things change significantly for Lea Sombé, if she is evidently a single student. Is she young? Here the answer is $inf^{**}(Y=y)=0.07$. In other words: Almost nothing we learn if we come to know that Lea Sombé, besides of being a single student, is also young. We expected it!

Note that in example 3, different from many logical con-cepts, there is no instantiation of Lea Sombé, but just an evidentiation of her temporary properties. Lea Sombé as well might be represented by the variable $LS=yes/no$ in the knowledge base and in this case LS must be connected to her properties by conditionals, to make a query. Instantiation of $LS=yes$ then permits a correct conclusion. Instantiation and evidentiation are equivalent.

This was what knowledge acquisition, query and response in an information theoretical environment is concerned. In the next section we study the iterative solution procedure of (14), (15), (16). Each iteration will be shown to be a con-sequent adaptation process of the information measure to one of the corresponding restrictions in (14) or (15).

4.3 Knowledge Adaptation in the Light of Infor-mation Theory

The mathematical optimization problems (14), (15) of the last section are equivalent to (14'), (15'). To verify this again, take the restrictions in (14'), (15') to the dual logarithm and multiply by – 1.

The solution of the probabilistic versions (14'), (15') is performed by a generalized Iterative Proportional Fitting (IPF) procedure, whose convergence was first shown by Csiszàr [CSI, 1975] in a very abstract measure theoreti-

cal form. We briefly repeat the IPF as we need it here, and put the essentials as theorem 3. The information theoretic version is attached as a lemma. Each single adaptation step (iteration step) is a mere application of Kuhn Tucker theory; so we leave a proof to the reader or refer to [MEY, 1998], theorem 3.7 or [ROD, 2000].

Theorem 3 (Iterative Proportional Fitting)

The solution of the optimization problem (14')

$$p^*inf = \arg\min \sum_v qinf(v) ld \frac{qinf(v)}{p^0 inf(v)}$$
$$s.t. \qquad qinf(B_i|A_i) = x_i, \quad i = 1...I$$

is the limit distribution of $pinf^0, pinf^1, \ldots, pinf^{k-1}, \ldots, pinf^k, \ldots$ and

$$pinf^k(v) = pinf^{k-1}(v) \cdot \alpha_k \cdot \begin{cases} 1 & v \subset \overline{A}_i \\ \left[\frac{x_i}{1-x_i} \cdot \frac{pinf^{k-1}(\overline{B}_i A_i)}{pinf^{k-1}(B_i A_i)} \right]^{-x_i} & v \subset \overline{B}_i A_i \\ [\quad -"- \quad]^{1-x_i} & v \subset B_i A_i \end{cases}$$

for $i=k$ mod I. Provided that there exists a feasible solution!

The α_k are normalizing factors; IPF runs through all conditionals repetitively and actualizes the probability of B_iA_i and $\overline{B}_iA_i$, respectively. After each actualization of conditional i the distribution obeys $qinf(B_i|A_i)=x_i$, but does not yet fulfil the others, in general. In the long run the limit distribution contains *all* knowledge at once and *all* what can be concluded from the given information.

There are fast implementations even for "large" distributions. How such distributions can be broken down to marginals and iterated efficiently, is beyond the scope of this contribution, see e.g. [MEY, 1998].

Note that to iterate (15') instead of (14') it suffices to replace $pinf^0$ by $pinf^*$ and $q(B_i|A_i)=x_i, i=1\ldots I,$ by $q(Fj|Ej)=y_j, j=1\ldots J.$

The IPF procedure of theorem 3 is equivalent to an iterative transformation of information measures, which we now put as a lemma.

Lemma 2 (Iterative Additive Fitting)

The solution of the optimization problem (14')

$$infp^* = \arg\min \sum_v qinf(v)\left(infp^0(v) - infq(v)\right)$$
$$s.t. \qquad inf\left(B_i|A_i\right) = rk_i, \quad i = 1...I$$

is the limit information measure of $infp^0, infp^1, ..., infp^{k-1}, infp^k, ...$ and

$$infp^k(v) = infp^{k-1}(v) + \beta_k +$$
$$\begin{cases} 0 & v \subset \overline{A}_i \\ x_i\left[\left(infp^{k-1}\left(B_i|A_i\right) - rk_i\right) - \left(inf^{k-1}\left(\overline{B}_i \mid A_i\right) - \overline{rk}_i\right)\right] & v \subset \overline{B}_i A_i \\ (x_i - 1)[\quad -"- \quad] & v \subset B_i A_i \end{cases}$$

for $i=k$ mod I. Here $\overline{rk}_i = -ld\left(1 - x_i\right)$.

Proof: Taking the iteration of theorem 3 to the dual logarithm and multiplying by –1 we get

$$infp^k(v) = infp^{k-1}(v) + ld\alpha_k +$$
$$\begin{cases} 0 & v \subset \overline{A}_i \\ x_i ld\left[\dfrac{x_i}{1-x_i} \cdot \dfrac{pinf^{k-1}\left(\overline{B}_i A_i\right)}{pinf^{k-1}\left(B_i A_i\right)}\right] & v \subset \overline{B}_i A_i \\ (x_i - 1)ld[\quad " \quad] & v \subset B_i A_i \end{cases}$$

Calculating the logarithm further, using the equation

$\dfrac{pinf^{k-1}\left(\overline{B}iAi\right)}{pinf^{k-1}\left(BiAi\right)} = \dfrac{pinf^{k-1}\left(\overline{B}_i|A_i\right)}{pinf^{k-1}\left(B_i|A_i\right)}$ and reordering terms we conclude the proposition. β_k=-$ld\alpha_k$ is the summand which makes $inf(\Omega)$=0.

The iteration of lemma 2 is of purely information theoretical nature. It transforms $infp^{k-1}$ into $infp^k$ on the whole Ω and respects the conditional structure according to $\overline{A}_i, \overline{B}_i|A_i, B_i|A_i$. $inf^{k-1}\left(B_i|A_i\right) - rk_i$ is the divergence of $B_i|A_i$'s information measure and the desired value rk_i, $inf^{k-1}\left(\overline{B}_i|A_i\right) - \overline{rk}_i$ is the respective divergence for the complementary conditional event. Remind the fact that there should be a penalty for divergence from above and from

below, as we demand equality. $\left|\left(inf^{k-1}\left(B_i|A_i\right)-rk_i\right)-\left(inf^{k-1}\left(\overline{B}_i|A_i\right)-\overline{rk}_i\right)\right|$ balances them both. Now multiplying this expression by x_i and x_i-1, respectively, updates information for both, $\overline{B}_i|A$ and $B_i|A_i$. Note that for a positive [–"–], the value $inf^{k-1}(v), v \subset \overline{B}_i A_i$, needs to be augmented as the difference $inf^{k-1}\left(B_i|A_i\right)-rk_i$ exceeds $inf^{k-1}\left(\overline{B}_i|A_i\right)-\overline{rk}_i$. And likewise $infp^{k-1}(v), v \subset B_i A_i$, should be diminished. That is exactly what the factors' x_i and x_i-1 sign does, and the factors perform a proportional contraction.

In the present section, to each iteration of the algorithmic solution of (14) and (15) an information theoretical interpretation was given. In the next section we study the limit measure $infp^*$.

4.4 Adaptation of Future Knowledge

Once the optimization problem (14) has been solved and thus the information measure $infp^*$ has been derived, we are ready for queries. But what if at a later occasion more information about the domain is available, how to adapt $infp^*$ to that? Was $infp^*$ constructed with the necessary care to make it ready for further adaptation processes? We ask the reader to repeat section 2.1. Information acquisition about a domain was characterized as an ongoing process, never finished since the vague population can not be described in terms of an ultimate probability distribution.

The following adaptation occurs when after having learned

$$infq\left(B_i|A_i\right)=rk_i, i=1...I_1,$$

we come to know the next package of knowledge pieces

$$infq\left(B_i|A_i\right)=rk_i, i=I_1+1,...,I_2,$$

compatible with the first one. (Any third, fourth or fifth adaptation proceeds in the same way).

Instead of (14) we now have to solve the problem $\left(infp^{1*}=infp^*\right)$:

$$\begin{aligned} infp^{2*} &= \arg\min \sum_v qinf(v)\left(infp^{1*}(v)-infq(v)\right) \\ s.t. \quad & infq\left(B_i|A_i\right)=rk_i, i=1...I_2. \end{aligned} \tag{17}$$

Note that (17) does not process ad hoc knowledge as does (15) inasmuch as further knowledge adaptation is

different from answering a query in a temporary evident situation. In (17) we iterate *all* knowledge pieces, the old and the new ones. In (15) the new information does not meet the old one, in general. It can be shown that (17) is equivalent to

$$infp^{2*} = \arg\min \sum_{v} qinf(v)\left(infp^{0}(v) - infq(v)\right) \quad s.t. \quad infq\left(B_i \middle| A_i\right) = rk_i, i = 1 \ldots I_2. \tag{18}$$

Either starting from ignorance $infp^0$ or from $infp^*$ does not make any difference. The result of iterating *all* restrictions $i=1 \ldots I_2$ is always $infp^{2}*$. The proof of this fact is implicitly given in [CSI, 1975] and in a less abstract from in [MEY, 1998], lemma 2.4. We therefore do not repeat it here.

The equivalence of (17) and (18) is an important property of our knowledge acquisition concept. There is no "time-dependence" or order-dependence of acquisition. All knowledge so far, even after many acquisition steps is a unique condensation of all packages.

In section 4.2 we charactized the choice of $infp^*$ and $infp^{**}$ by solving (14) and (15) as the intuitive resistance against not intended revision of former convictions. Now, in the light of a deeper insight of consecutive adaptation steps we prove $infp^*$ and $infp^{**}$ to be the only rational choice. We restrict our attention to $infp^*$, as the reflections for $infp^{**}$ are identical.

Given the restrictions in (14), the facts $infq(B_i|A_i)=rk_i$, $i=1 \ldots I$, provided to build the knowledge base, are the imperative to choose a feasible $infq$ in an adequate manner. Keeping in mind that the provided facts are information against ignorance $infp^0$, knowledge should grow with a growing number of facts. Roughly speaking, there should be a certain kind of monotony.

More formally: Let $R_1=\{infq(B_i|A_i)=rk_i, i=1 \ldots I_1\}$

and $R_2=\{infq(B_i|A_i)=rk_i, i=I_1+1, \ldots, I_2\}$

be two subsequently provided sets of conditionals and let $R=R_1 \cup R_2$. A choice function CH is function which chooses a unique element out of the set of all R_1(or $R=R_1 \cup R_2$) - feasible *inf*-measures. To be a *rational* choice function, CH should have the above mentioned monotony property which we now put as a definition.

Definition 5 (Rational Choice Function)

A choice function CH is rational if for no R^1 there exists a R^2 in such a way that ($\mathsf{R}= \mathsf{R}_1 \cup \mathsf{R}_2$) for all R - feasible *infq* the inequality $\sum qinf(v)\left(inf\, p^0(v) - inf\, q(v)\right) < \sum \bar{p}inf(v)\left(inf p^0(v) - inf \bar{p}(v)\right)$ holds. Here $inf\,\bar{p} = \mathrm{CH}(\mathsf{R}^1)$.

In other words: A rational choice $inf\,\bar{p} = \mathrm{CH}\,(\mathsf{R}^1)$ should not be such that for all *infq*, compatible with a richer set of facts $\mathsf{R}=\mathsf{R}^1 \cup \mathsf{R}^2$, the uncertainty change against $infp^0$ is less than that of $inf\,\bar{p}$ against $infp^0$. This reflects our understanding of knowledge growth. Now the only rational choice is the one which selects the *inf*-measure with minimum relative entropy. We put this as a lemma.

Lemma 3 (MINREL is Rational)

The only rational choice $\mathrm{CH}(\mathsf{R}^1)$ for any R^1 is

$$infp^* = CH\left(\mathsf{R}^1\right) = \operatorname{argmin}\sum_{v} qinf(v)\left(inf\, p^0(v) - inf\, q(v)\right) \qquad (19)$$
$$s.t. \qquad infq \text{ is } \mathsf{R}^1 \text{ - feasible}$$

To prove the above lemma assume $infp' \neq infp^*$ to be a rational choice and construct a set R^2 of facts which, together with R^1, makes $infp^*$ unique. For this unique *inf*-measure in R - which is richer than R^1 - we have $\sum p^* \mathrm{inf}(v)\left(inf\, p^0(v) - inf\, p^*(v)\right) < \sum p' inf(v)\left(inf p^0(v) - infp'(v)\right)$.

This constradicts the rationality of *infp'*. Note that for the construction of R^2 you might choose the facts $infq(v) = rk(v) = infp^*(v)$, all v.

Adaptation of subsequent information packages implies in ongoing knowledge revision steps, as developed here. Such adaptation turns out to be much more sophisticated if newer is not compatible with older information. A full treatment of knowledge revision under information fidelity will be given in a separate paper.

5. The Shell XSPIRIT

The shell XSPIRIT 3.0, a Java-version, is a professional tool for information and knowledge processing as developed here. It processes probabilities rather than information measures, but an additional module allows to handle both. After the user defined variables as boolean, nominal and ordinal, he might construct conjunctions, disjunctions and negations of literals and form composed conditionals, in an arbitrary hierarchy. The variables' attributes might be associated with real number (utilities). The shell at any time calculates expected utilities and so supports modelling decision problems.

Once the composed conditionals and their corresponding inf-values are supplied by the user, XSPIRIT automatically generates a hyper-tree on the set of all variables in such a way that variables "contained" in the conditional lie in the same hyper-edge. The shell then calculates iteratively the knowledge base making use of the Markovian property of the global distribution with respect the hyper-edges.

Any conditional can be put active or inactive, thus allowing to use it for knowledge acquisition, query or response.

Bayes Nets are importable and XSPIRIT also provides frenquentistic learning rather than the acquisition of subjective probabilities or estimated information values.

The biggest example calculated with XSPIRIT until now was a business-to-business model for advertisement evaluation of a leading German steel-company. It counts appr. 1300 conditionals and more than 80 variables, not all of which are binary. The CPU-time for knowledge acquisition or query is below 5 seconds on a PC.

The reader interested in the shell SPIRIT might visit the homepage www.fernuni-hagen.de/BWLOR.

References

[CAL, 1991] P. G. Calabrese: *Deduction and inference using conditional logic and probability*, in: I.R. Goodman, M. M. Gupta, H. T. Nguyen, G. S. Rogers (Eds.): Conditional Logic in Expert Systems, Elsevier Science, Amsterdam, 71-100, 1991.

[CSI, 1975] I. Csiszàr: *I-Divergence Geometry of Probability Distributions and Minimisation Problems*, The Annals of Probability, Vol. 3, No. 1, 146-158, 1975.

[DUP, 1990] D. Dubois, H. Prade: *The logical view of conditioning and its application to possibility and evidence theories*, Internat. J. Approx. Reason. 4, 23-46, 1990.

[DUP, 1991] D. Dubois, H. Prade: *Conditioning, Non-Monotonic Logic and Non-Standard Uncertainty Models*, in: I. R. Goodman, M. M. Gupta, H. T. Nguyen, G. S. Rogers (Eds.): Conditional Logic in Expert Systems, Elsevier Science Amsterdam, 115-158, 1991.

[FIN, 1972] B. de Finetti: *Induction and Statistics*, Wiley, New York, 1972.

[HAR, 1928] R.V.L. Harley: *Transmission of Information,* The Bell Systems Technical Journal, 7, 535-563, 1928.

[JEN, 1994] F. V. Jensen; F. Jensen: *Optimal Junction Trees*, in: Proceedings of the 10th Conference on Uncertainty in Artifical Intelligence (UAI-94), 360-366, ed. Ramon Lopez de Mantaras, David Poole, Morgan Kaufman Publishers, San Francisco, California, 1994.

[KIS, 2001] G. Kern-Isberner: *Conditionals in nonmonotonic reasoning and belief revision*, Springer Lecture Notes in Artificial Intelligence 2087, 2001.

[MEY, 1998] C.-H. Meyer: *Korrektes Schließen bei unvollständiger Information,* Europäische Hochschulschriften, Peter Lang, Frankfurt a.M., 1998.

[RES, 1969] N. Rescher: *Many-Valued Logics*, McGraw-Hill, New York, 1969.

[RME, 1996] W. Rödder, C.-H. Meyer: *Coherent knowledge processing at maximum entropy by SPIRIT*, in: E. Horvitz, F. Jensen (Eds.): Proc. 12th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, San Francisco, CA, pp. 470-476, 1996.

[ROD, 2000] W. Rödder: *Conditional Logic and the Principle of Entropy*, Artificial Intelligence, 117, 83-106, 2000.

[ROD, 2001] W. Rödder: *Knowledge Processing under Information Fidelity*, Proc. 12th Int. Joint Conf. on AI (IJCAI – 2001), Seattle, 749-754, 2001.

[SHA, 1948] C. E. Shannon: *Mathematical Theory of Communication,* The Bell System Technical Journal, 27, 379-423 + 623+656, 1948.

[SOM, 1992] L. Sombé: *Schließen bei unsicherem Wissen in der Künstlichen Intelligenz*, Vieweg, Wiesbaden, 1992.