T U M

I N S T I T U T F Ü R I N F O R M A T I K

Non-Monotonic Reasoning on Probability Models:

Indifference, Independence & MaxEnt

Part I – Overview

Manfred Schramm / Michael Greiner

TUM–I9509

März 1995

T E C H N I S C H E U N I V E R S I T Ä T M Ü N C H E N

TUM-INFO-03-1995-I9509-350/1.-FI

Typescript: LaTeX/Postscript

Druck: Mathematisches Institut und
Institut für Informatik der
Technischen Universität München

Non-Monotonic Reasoning on Probability Models:
Indifference, Independence & MaxEnt

Part I – Overview

Manfred Schramm and Michael Greiner
Institut für Informatik der Technischen Universität München, Germany

Keywords

Indifference, Independence, Maximum Entropy, Non-Monotonic Reasoning, Statistical Reasoning, Default Reasoning, Undirected Graphs, Decisions under Incomplete Knowledge, Simpson's Paradox

Abstract

Through completing an underspecified probability model, Maximum Entropy (MaxEnt) supports non-monotonic inferences. Some major aspects of how this is done by MaxEnt can be understood from the background of two principles of rational decision: the concept of Indifference and the concept of Independence. In a formal specification MaxEnt can be viewed as (conservative) extension of these principles; so these principles shed light on the "magical" decisions of MaxEnt. But the other direction is true as well: Since MaxEnt is a "correct" representation of the set of models (Concentration Theorem), it elucidates these two principles (e.g. it can be shown, that the knowledge of independences can be of very different information-theoretic value). These principles and their calculi are not just arbitrary ideas: When extended to work with qualitative constraints which are modelled by probability intervals, each calculus can be successfully applied to V. Lifschitz's Benchmarks of Non-Monotonic Reasoning and is able to infer some instances of them ([Lifschitz, 1988]). Since MaxEnt is strictly stronger than the combination of the two principles, it yields a powerful tool for decisions in situations of incomplete knowledge. To give an example, a well-known problem of statistical inference (Simpson's Paradox) will serve as an illustration throughout the paper.

1 Introduction

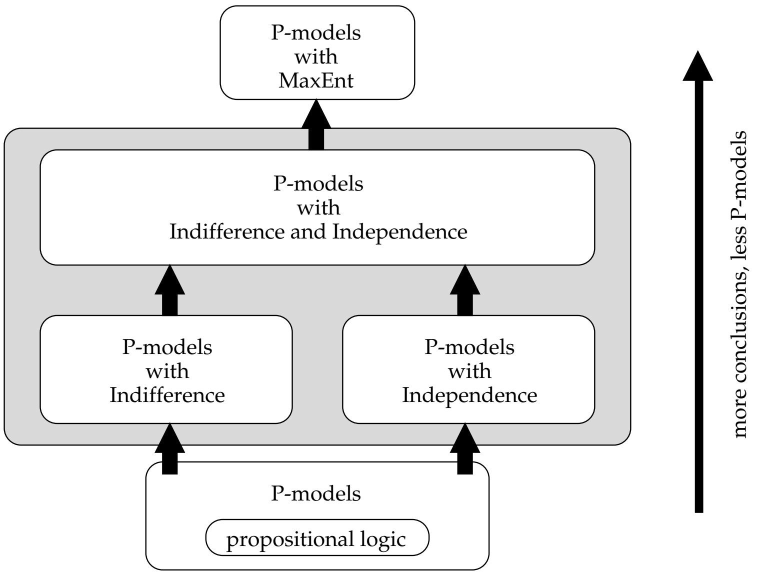
1.1 Background

If we want to model common sense reasoning, an important step will be the development of systems which can make decisions under incomplete knowledge. These decisions should be the best possible ones given the incomplete knowledge; they will show non-monotonic behaviour when the knowledge is increasing. Recently, probability theory has become more and more accepted as an appropriate tool for that purpose, especially in connection with the notion of entropy ([Paris & Vencovska, 1989], [Pearl, 1988], [Cheeseman, 1988]). Following [Cox, 1979], we consider probability theory as an adequate model for one-dimensional belief of propositional expressions[1]. Following [Adams, 1975], we consider the conditional probability to be much more adequate compared to the use of the Material Implication[2] of propositional logic when modelling the common sense connective "If, then" of the language. Following [Jaynes, 1982] we consider MaxEnt as an adequate method of choosing a probability model from an infinite set of possible models, when only linear constraints are present. Concerning MaxEnt it is still a problem to explain this method of inductive reasoning to any newcomers. Surely there are various ways. One possibility is to take some intuitively plausible axioms of rational reasoning and to show how MaxEnt is a necessary consequence of these axioms. This approach has been chosen quite a few times in the literature ([Shore & Johnson, 1980], [Skilling, 1988], [Paris & Vencovska, 1990]). Here we choose a slightly different approach; we take two strong properties, strong enough to define decision principles, and we show that MaxEnt concludes strictly stronger (see 6. and the figure below) than the two principles combined. Both principles seem to be different from MaxEnt at first glance, and although they seem to be well-known for a long time, they are far from clear when one looks at them in more detail:

- The **principle of Indifference**, viewed by [Jaynes, 1978] as a simple "demand of consistency", is sometimes mixed with the problem of modelling probabilities; this leads to arguments against this principle. Therefore we have to specify how we use this principle, especially in the presence of linear constraints.

- The **principle of Independence** is related to undirected graphs and to the Markov properties of its variables; it seems that it has not been used so far as a formal principle of reasoning (but see [Pearl, 1988]). If MaxEnt is derived from the usual axioms, only a special case of this principle is required for the proof.

So the paper proceeds from the bottom to the top of the following figure:



[1]the relation between belief, statistics and non-monotonic reasoning ist broadly discussed in [Bacchus et al., 1994]

[2]the Material Implication of two propositions (a, b), normally denoted by $(a \rightarrow b)$, is false iff the first proposition (antecedens) is true and the second one is false

First, the logic on **probability models** (P-models) is formally described and illustrated by use of Simpson's paradox. The principles of Indifference and Independence are then introduced as additional axioms on P-models. Some remarks about the relation between MaxEnt and these principles conclude this short presentation.

1.2 Mathematical formulation

In order to illustrate the following formal definitions we start with a small example of default-knowledge:

Default-Knowledge:	• normally animals do not fly • birds are animals • normally birds fly
Desired conclusion:	Animals, which are not birds, normally do not fly.

As we want to model the given common sense information in a probabilistic[3] way we have to construct an appropriate **measurable space** first:

In this example the set of all living beings is a suitable **reference space** Θ (see [Chow & Teicher, 1978]). Furthermore we get the following **events** (i.e. elements of the power set $\mathfrak{P}(\Theta)$ of Θ)

- $an \equiv$ "beings that are animals"
- $bi \equiv$ "beings that are birds"
- $fl \equiv$ "beings that can fly"

which we gather in the set $R := \{an, bi, fl\}$.[4] Of course, a more detailed splitting of the elements of R (for instance the information about birds can be split into information about nightbirds and birds that are active during the day) is possible but unnecessarily increases the complexity of the mathematical model.

For the formulation of the following principles it is sufficient to consider events of a discrete probability space that are built by the set operations $\cap$, $\cup$ and $\neg$ over R.[5] In general this leads to the set

$$\Omega := \left\{ \bigcap_{i=1}^{\ell} e_i \mid e_i \in \{a_i, \neg a_i\} \right\}^{6}$$

of **full conjunctions** over $R := \{a_1, \ldots, a_\ell\}$ where $a_i \in \mathfrak{P}(\Theta)$, $i = 1, \ldots, \ell$. It is a well-known fact from probability theory that the (maximal) 2^ℓ elements of Ω are mutually disjoint and span the set R (i.e. any a_i can be expressed by a disjunction of elements of Ω).

[3] and therefore set theoretic

[4] in general the reference space has to be a strict superset of all events that are mentioned either in the knowledge base or in the conclusion

[5] from an information theoretic point of view we consider R to be a minimal set of problem dependent variables (here an, bi and fl) whose combinations (via $\wedge$, $\vee$, $\neg$) are used to translate given information into formal sentences (see definition 3)

[6] $\neg a_i$ denotes the complement of a_i in Θ (which is not empty by the construction of Θ as strict superset)

Therefore the smallest (σ-) algebra $\mathfrak{A}(R)$ that contains R is identical to $\mathfrak{A}(\Omega)$ $[=\mathfrak{P}(\Omega)]$. For these reasons we restrict the set of elementary events (also called the **set of possible worlds**) to Ω instead of the underlying Θ and do not mention Θ any more.

Definition 1: Over a set $R := \{a_1, \ldots, a_\ell\}$ a **measurable space** $(\Omega, \mathfrak{A})$ ist defined by

- $\Omega = \left\{ \bigcap_{i=1}^{\ell} e_i \mid e_i \in \{a_i, \neg a_i\} \right\}$
- $\mathfrak{A} = \mathfrak{A}(\Omega) = \mathfrak{P}(\Omega)$.

Definition 2: Let $(\Omega, \mathfrak{A})$ be a measurable space over R with $\Omega = \{\omega_1, \ldots, \omega_n\}$. A (discrete) **probability measure** or **probability model** (P-model) P is an assignment of non-negative numerical values to the elements of Ω, which sum up to unity. In symbols:

$$p_i := P(\omega_i) \geq 0\,,\ \ i = 1, \ldots, n \ \ \text{and} \ \ p_1 + \ldots + p_n = 1\,.$$

The n-tupel $(p_1, \ldots, p_n)$ is called a **probability vector** (P-vector).

W_Ω (respectively V_Ω) denotes the set of all possible P-models (P-vectors) for $(\Omega, \mathfrak{A})$.

Definition 3: For given $(\Omega, \mathfrak{A})$, $P \in W_\Omega$, $a, b \in \mathfrak{A}$, $P(a) > 0$ and $I \subseteq [0, 1]$ the term

$$\langle P(b \mid a) = \delta;\ \delta \in I \rangle^{7}$$

is called a **sentence** in $(\Omega, \mathfrak{A})$. The sentence given above is called true in $P \in W_\Omega$ iff $P(b \mid a) \in I$. Otherwise it is called false.

Remarks:

- It is easy to see that $P(b \mid a) = \delta$ can be notated as a **linear equality** for the elementary probabilities p_i, $i = 1, \ldots, n$:

$$P(b \mid a) = \delta \ \Leftrightarrow\ P(a \cap b) = \delta \cdot P(a) \ \Leftrightarrow\ (1 - \delta) \cdot \sum_{i:\, \omega_i \in a \cap b} p_i - \delta \cdot \sum_{j:\, \omega_j \in a \cap \neg b} p_j = 0\,.$$

- The definition of a sentence can be extended to any term that can be transformed into a linear equation.

  Example:

$$P(b) - P(a) = \delta \ \Leftrightarrow\ \sum_{i:\, \omega_i \in b} p_i - \sum_{j:\, \omega_j \in a} p_j = \delta\,.$$

Definition 4: Let $\mathrm{DB} := \{s_1, \ldots, s_m\}$ be a set of m sentences in $(\Omega, \mathfrak{A})$. W_{DB} is defined as the set of all P-models $P \in W_\Omega$ in which $s_1, \ldots, s_m$ are true. In this context we call $s_1, \ldots, s_m$ **constraints** on W_Ω and W_{DB} the set of all elements of W_Ω that are **consistent** with the constraints in DB.

[7] $P(b \mid a) := P(a \cap b)/P(a)$ denotes the conditional probability of the event b given a.
Remark: $P(b) = P(b \mid \Omega)$

If W_{DB} consists of more than one element (here equivalent to infinitely many), the information in DB is incomplete for determining a single P-model. If W_{DB} is empty, the information in DB was inconsistent.
We want to model **incomplete information**, expressed by linear constraints (premises) over a set of P-models, so the case that there are "infinitely many elements in W_{DB}" will be our standard case.

Definition 5: A sentence s which is true in all P-models of W_{DB} is called a **conclusion** from DB, **in symbols**: DB $\|\!\sim s$. Therefore, adding a conclusion to DB will not change the set of models of W_{DB}.

A **belief in *a*** of a system now means to us that, if no other information is given and the system is forced to decide between a and $\neg a$, the system will decide for a (default decision). According to the relationship between probabilities and decisions, we model the belief in a as

$$\langle P(a) = \delta; \delta \in (0.5, 1] \rangle \in \mathrm{DB}\,.$$

Knowledge is expressed by probability 1 (a is known to be true iff $\langle P(a) = 1 \rangle \in \mathrm{DB}$). Therefore, if a sentence of the form $\langle P(a) = \delta; \delta \in (0.5, 1] \rangle$ for some propositional expression a is a conclusion from DB, the system will decide for a given the knowledge in DB. This interpretation of defaults is quantitative; especially this kind of belief means "in more than half of the cases". This is weaker than "in most cases" (similar to "normally"), but the quantitative meaning of *most* is context-dependent and therefore difficult to describe; the structure of the desired conclusions of *most* seems to be very similar to that of "more than half". So we opted for that interpretation. Conditional knowledge (belief, decisions) is of course expressed by conditional probabilities: $\langle P(b \mid a) = \delta; \delta \in (0.5, 1] \rangle$ means that if the system knows a (and nothing else), it believes (decides for) b.

Now we are able to handle the example from the beginning of this section:

$$\begin{aligned} \mathrm{DB}_1 \;:=\; & \{ \langle P(\neg fl \mid an) = \delta_1;\ \delta_1 \in (0.5, 1] \rangle\,,\ \langle P(an \mid bi) = 1.0 \rangle\,, \\ & \langle P(fl \mid bi) = \delta_2;\ \delta_2 \in (0.5, 1] \rangle \} \\ \equiv\; & \{ \langle v_2 + v_4 > v_1 + v_3 \rangle\,,\ \langle v_5 + v_6 = 0 \rangle\,,\ \langle v_1 > v_2 \rangle \} \end{aligned}$$

where $v_1 := P(an \cap bi \cap fl)$, $v_2 := P(an \cap bi \cap \neg fl)$, $v_3 := P(an \cap \neg bi \cap fl)$, $v_4 := P(an \cap \neg bi \cap \neg fl)$, $\ldots$, $v_7 := P(\neg an \cap \neg bi \cap fl)$, $v_8 := P(\neg an \cap \neg bi \cap \neg fl)$.

We have to decide whether

$$P(\neg fl \mid an \cap \neg bi) = \frac{P(an \cap \neg bi \cap \neg fl)}{P(an \cap \neg bi)} = \frac{v_4}{v_3 + v_4} \lesseqgtr 0.5\ ? \quad (\Leftrightarrow\ v_4 \lesseqgtr v_3\ ?)$$

From the last notation of DB_1 we get

$$v_2 + \underline{v_4} > v_1 + v_3 > v_2 + \underline{v_3}\,, \quad \text{i.e. } v_4 > v_3\,.$$

Therefore the desired conclusion is valid for any P-vector that is consistent with the constraints in DB_1, in symbols: $\mathrm{DB}_1 \|\!\sim \langle P(\neg fl \mid an \cap \neg bi) = \delta_3;\ \delta_3 \in (0.5, 1] \rangle$.

2 Conclusions on P-models

This kind of logic on P-models (**P-logic**), described so far, is of course strictly stronger than propositional logic, which can be embedded into P-logic as follows: Take the premises of propositional logic as knowledge with probability 1 into DB and look for expressions, being true in all remaining possible worlds. P-logic is surely useful, when modelling certain examples of reasoning (as already shown in the previous section this logic supports the desired conclusion from DB_1). Moreover the use of conditional probabilities instead of Material Implication avoids some of the well-known modelling problems with the Material Implication. Also P-logic allows for a richer language than propositional logic, but it still has the property of being monotonic (additional knowledge won't revise earlier decisions). However, we aim at something which is much stronger; because too many conclusions which seem to be intuitively true are not supported by this P-logic.

Example: DB_2

[Weak version of Simpson's Paradox ([Blyth, 1973], [Neufeld & Horton, 1990])]:

$$DB_2 \;=\; \{\, \langle P(c \mid a) = \delta_1; \delta_1 \in (0.5, 1] \rangle\,,\; \langle P(c \mid b) = \delta_2; \delta_2 \in (0.5, 1] \rangle \,\}^8$$

Desired conclusions: (c1) $DB_2 \mathrel{\|\!\!\sim} \langle P(c \mid a \cup b) = \delta_3; \delta_3 \in (0.5, 1] \rangle$

(c2) $DB_2 \mathrel{\|\!\!\sim} \langle P(c \mid a \cap b) = \delta_4; \delta_4 \in (0.5, 1] \rangle$

These conclusions seem intuitively obvious although they are not true in P-logic (or in statistics): We construct a counter-example by means of P-models, which fulfil the premises, but not the conclusions.

not (c1): Let $P(abc) = 6/18$[9], $P(ab\neg c) = 1/18$, $P(a\neg bc) = 1/18$, $P(a\neg b\neg c) = 5/18$, $P(\neg abc) = 1/18, P(\neg ab\neg c) = 4/18, P(\neg a\neg bc) = 0$ and $P(\neg a\neg b\neg c) = 0$. Then

$$\begin{aligned}
P(c \mid a) &= \frac{P(a \cap c)}{P(a)} = \frac{P(abc) + P(a\neg bc)}{P(abc) + P(ab\neg c) + P(a\neg bc) + P(a\neg b\neg c)} \\
&= \frac{6+1}{6+1+1+5} = \frac{7}{13} > 0.5\,, \\
P(c \mid b) &= \frac{P(b \cap c)}{P(b)} = \frac{P(abc) + P(\neg abc)}{P(abc) + P(ab\neg c) + P(\neg abc) + P(\neg ab\neg c)} \\
&= \frac{6+1}{6+1+1+4} = \frac{7}{12} > 0.5\,, \\
P(c \mid a \cup b) &= \frac{P(c \cap [a \cup b])}{P(a \cup b)} = \frac{P(abc) + P(a\neg bc) + P(\neg abc)}{1 - P(\neg a\neg bc) - P(\neg a\neg b\neg c)} \\
&= \frac{6+1+1}{18-0-0} = \frac{8}{18} < 0.5\,.
\end{aligned}$$

[8] if the system knows a, it believes (decides for) c; if the system knows b, it believes (decides for) c

[9] $abc \equiv a \cap b \cap c\,, \ldots$

not (c2): Let $P(abc) = 1/20$, $P(ab\neg c) = 5/20$, $P(a\neg bc) = 6/20$, $P(a\neg b\neg c) = 1/20$, $P(\neg abc) = 6/20, P(\neg ab\neg c) = 1/20, P(\neg a\neg bc) = 0$ and $P(\neg a\neg b\neg c) = 0$. Then

$$P(c \mid a) = P(c \mid b) = \frac{7}{13} > 0.5\,,\ \text{whereas}\ P(c \mid a \cap b) = \frac{1}{6} < 0.5\,.$$

This makes the Simpson problem a *common sense paradox*. Probability theory is too fine-grained to model common sense reasoning in general. The remaining degrees of freedom have to be filled up; to do this without adding information is still a problem, last but not least addressed by the MaxEnt-Program of Jaynes. Filling the degrees of freedom with correct methods will help to overcome the mistrust in statistics which can be found even among scientifically educated people. So our goal is to look for additional (context-sensitive) constraints (resp. principles), which are able to support rational decisions with incomplete knowledge (e.g. the desired conclusions of the last example DB_2). This will be done in the next sections.

3 Conclusions on P-models with Indifference

3.1 What does Indifference mean?

The history of this famous principle goes back to Laplace and Keynes. Let us quote [Jaynes, 1978] for a short and informal version of this principle:

> "If the available evidence gives us no reason to consider proposition a_1 either more or less likely than a_2, then the only honest way we can describe that state of knowledge is to assign them equal probabilities: $P(a_1) = P(a_2)$."

Three questions arise here:

a) How to make formally precise that a system has no reason to consider a_1 either more or less likely as a_2 in the presence of linear constraints?

b) Why should we use this principle?

c) Given a set of linear constraints of DB: is it possible to decide on the basis of this set which elementary events (and therefore which complex events) will be considered to be indifferent?

We will adress these questions on the following two pages.

3.2 Mathematical formulation

Let W_{DB} be the set of P-models of DB, V_{DB} the set of P-vectors of DB and $v \in V_{\text{DB}}$ a single vector. Now look for permutations Π with $\forall v \in V_{\text{DB}}\ \exists v^* \in V_{\text{DB}} : \ \Pi(v) = v^*$, in short form written as: $\Pi(V_{\text{DB}}) = V_{\text{DB}}$. It is well-known, that any permutation can be expressed by writing down its cycles, so we express Π by describing its cycles. The principle of Indifference now demands that all variables (we express the unknown probabilities of elementary events by variables) within the same cycle get the same value. We define the set I_{DB} as the collection of all the equations of any Π with the property $\Pi(V_{\mathrm{DB}}) = V_{\mathrm{DB}}$. s is a consequence of a set of linear constraints with the help of the principle of Indifference iff the following relation is valid: $DB \cup I_{\mathrm{DB}} \;||\!\!\sim s$.

3.3 The main argument for using Indifference: Consistency

If W_{DB} contains P-models with the property $P(a_1) < P(a_2)$ and $P(a_1) > P(a_2)$ and a_1 is indifferent to a_2 as defined above, an unknown future decision process based on this set of P-models might once choose a model with the property $P(a_1) < P(a_2)$ and might choose a P-model with $P(a_1) > P(a_2)$ at another time. Both models contain information which is not present in the database. On the basis of V_{DB} we notice that we won't be able to recognize if a permutation Π (of the kind $\Pi(V_{\mathrm{DB}}) = V_{\mathrm{DB}}$) has happened inside our machine which switches the values of some variables (this is equivalent to renaming the variables) and changes a model with the property $P(a_1) < P(a_2)$ into a model with the opposite property. Of course we don't want something we can't notice to have any influence on future (rational) decisions. That's what the principle of Indifference is able to prevent: it disposes of those degrees of freedom which our constraints do not address and which we therefore are not able to control in a rational manner.

3.4 Another argument for using Indifference: Model Quantification

Take $W_{\mathrm{I(DB)}}$ as the set of all P-models, which satisfy the constraints in DB and the equations in I_{DB}; take $V_{\mathrm{I(DB)}}$ as the corresponding set of all P-vectors. Given that the MaxEnt-solution of a problem with linear constraints is the correct representation of the set of P-models (what was proved by [Jaynes, 1982] via the Concentration Theorem, see section 6.2), it is possible to consider every Indifference model $w^i \in W_{\mathrm{I(DB)}}$ as MaxEnt-solution of a subproblem DB_i, where W_{DB_i} is an element of a certain partition of W_{DB} (the partition is formed by varying the values of additional constraints derived from models in $W_{\mathrm{I(DB)}}$). Then this P-model w^i is of course a correct representation of the set W_{DB_i}. If this is the case, only a minimum amount of information is necessary to replace the set W_{DB_i} by the model w^i (the amount tends to zero if the problem is modelled by a random experiment of size N and N grows large) and only a minimum of information is contained in I_{DB}. This means that statistically all models in $W_{\mathrm{I(DB)}}$ have a special representation status.

3.5 How to detect indifferent events by the matrix M of linear constraints

A sufficient condition for Π to have the property of $\Pi(V_{\mathrm{DB}}) = V_{\mathrm{DB}}$ is the existence of an permutation M_Π of the columns of M, which, followed by an permutation M_Λ of the rows of M, is equivalent to M (formally: $M_\Lambda \cdot M \cdot M_\Pi = M$).

Proof: Systems with the same matrix of equations have the same set of solutions.

Example: Let us take $\mathrm{DB}_3 := \mathrm{DB}_2 \cup \{\delta_1 = \delta_2 = \delta\}$. The matrix M of linear constraints has the entries

$v_1 :=$ $P(abc)$	$v_2 :=$ $P(ab\neg c)$	$v_3 :=$ $P(a\neg bc)$	$v_4 :=$ $P(a\neg b\neg c)$	$v_5 :=$ $P(\neg abc)$	$v_6 :=$ $P(\neg ab\neg c)$	$v_7 :=$ $P(\neg a\neg bc)$	$v_8 :=$ $P(\neg a\neg b\neg c)$	$=$
1	1	1	1	1	1	1	1	1
$1-\delta$	$-\delta$	$1-\delta$	$-\delta$	0	0	0	0	0
$1-\delta$	$-\delta$	0	0	$1-\delta$	$-\delta$	0	0	0

We obtain $\Pi(V_{\mathrm{DB}_3}) = V_{\mathrm{DB}_3}$ for the permutation $\Pi = \binom{v_1}{v_1}\binom{v_2}{v_2}\binom{v_3\ v_5}{v_5\ v_3}\binom{v_4\ v_6}{v_6\ v_4}\binom{v_7\ v_8}{v_8\ v_7}$.
Equations in I_{DB}: $\{ v_3 = v_5 ,\ v_4 = v_6 ,\ v_7 = v_8 \}$.

3.6 Examples (no rules) of the use of indifference

- $n = |\Omega|$ implies: $\emptyset \cup I_\emptyset \mathrel{||\!\!\sim} \langle P(\omega_i) = 1/n\rangle \quad \forall \omega_i \in \Omega$.
- Take DB_4 as equal to $\{ \langle P(b \mid a) = \delta_1; \delta_1 \in (0.5, 1]\rangle \}$.
  Conclusion: $\mathrm{DB}_4 \cup I_{\mathrm{DB}_4} \mathrel{||\!\!\sim} \langle P(b \mid a \cap c) = \delta_2; \delta_2 \in (0.5, 1]\rangle$.[10]
- Take DB_5 as equal to $\{ \langle P(b \mid a \cap c) = \delta_1; \delta_1 \in (0.5, 1]\rangle \}$.
  Conclusion: $\mathrm{DB}_5 \cup I_{\mathrm{DB}_5} \mathrel{||\!\!\sim} \langle P(b \mid a) = \delta_2; \delta_2 \in (0.5, 1]\rangle$.[11]

3.7 Summary (Indifference)

Two important arguments (consistency, quantification of possible worlds) justify the use of the principle of Indifference when decisions are necessary. Of course it does not solve the problem of modelling, which is the problem of defining Ω and encoding our knowledge. Some paradoxes of the use of Indifference are related to the selection of different Ω's and therefore different results of the principle of Indifference (see e.g. [Neapolitan, 1990], [Howson & Urbach, 1993]). The consistency (i.e. $V_{\mathrm{DB}} \neq \emptyset \Rightarrow V_{\mathrm{I(DB)}} \neq \emptyset$) of this principle can be proven by the convexity of V_{DB} in any component of the vectors v ($\in V_{\mathrm{DB}}$). Moreover the MaxEnt-Model fulfils all the equations of I(DB) (which means that the MaxEnt-Model w^* is an element of $W_{\mathrm{I(DB)}}$). The decisions based on P-models and the principle of Indifference are of course strictly stronger than that on pure P-models. The decisions have already the property of being non-monotonic, when additional information becomes available (indifferences might disappear, when new knowledge comes in).

[10] Indifference demands the equations $P(abc) = P(ab\neg c)$, $P(a\neg bc) = P(a\neg b\neg c)$, $P(\neg abc) = P(\neg ab\neg c) = P(\neg a\neg bc) = P(\neg a\neg b\neg c)$

[11] Indifference demands $P(ab\neg c) = P(a\neg b\neg c) = P(\neg abc) = P(\neg ab\neg c) = P(\neg a\neg bc) = P(\neg a\neg b\neg c)$

4 Conclusions on P-models with Independence

4.1 Basics

From the point of information theory, Independence of two events a and b in a P-model w is given, if any knowledge about the event a (like a has (or has not) happened) does not change the probability of b (and vice versa) in w (formally $P(b \mid a) = P(b)$). With the knowledge of Independence of the two events, the probability of the combined event becomes a function of the probability of the single events. If this is the case not only for single events, but for all values of a random variable, Independence allows to reduce the complexity (of calculating) and the space (for storing probability models ([Lewis, 1959]). In Bayesian Reasoning, Independence is well-known and commonly used when completing incomplete knowledge or when simplifying calculations (see e.g. [Pearl, 1988]). In our context the following questions arise:

a) How to make formally precise which kind of (conditional) Independence a system should demand?

b) Why should we use this principle?

c) Given a set of linear constraints of DB: is it possible to decide on the basis of this set which events will become independent?

4.2 Mathematical formulation

The principle of Independence is based on the construction of an undirected graph from the constraints in DB by the following rules: Let us take every variable from R as a knot and let us connect two variables by an edge, iff the two variables are both mentioned in the same constraint. Consider the resulting undirected graph as **Independence map** (I-map; see [Pearl, 1988]). We take all the statements of (conditional) Independence of the map and translate it into (non-linear) equations between events of Ω. We define U_{DB} as the set of all these equations.[12] s is a consequence of a set of linear constraints with the help of the principle of Independence iff the following relation is valid: $\mathrm{DB} \cup U_{\mathrm{DB}} \;\Vert\!\sim s$.

Example: The Independence map of DB_2 $(R = \{a, b, c\})$ is

$$\{a\} \;—\!—\; \{c\} \;—\!—\; \{b\}$$

This Independence map now demands that any event of $\mathfrak{A}_{\{a\}}$ is (conditionally) independent from any event of $\mathfrak{A}_{\{b\}}$, conditioned on an elementary event of $\Omega_{\{c\}}$.

4.3 First argument: Intuitive graphical representation

Some years ago, conditional Independence relations in P-models have been identified as a model for a set of axioms, which describe (and conclude) connections on undirected

[12] the set U_{DB} expresses many possible independences between subalgebras of $\mathfrak{A}(\Omega)$

graphs (an introduction to this topic can be obtained from [Pearl, 1988]). This means that (conditional) Independence relations could be detected by only qualitative information about a P-model: The quantitative information, encoded in the numerical values of its events, is not necessary (see e.g. [Pearl, 1988]). We find this approach very important for MaxEnt, because it clarifies the relation between MaxEnt and (conditional) Independence.[13]

4.4 Second argument: Quantification of possible worlds

Take $W_{\mathrm{U(DB)}}$ as the set of all P-models which fulfil the constraints in DB and the equations in U_{DB}; take $V_{\mathrm{U(DB)}}$ as the corresponding set of all P-vectors. Given that the MaxEnt-solution of a problem with linear constraints is the correct representation of the set of P-models, it is possible to consider every Independence model w^u ($\in W_{\mathrm{U(DB)}}$) as MaxEnt-solution of a subproblem DB_u, where W_{DB_u} is an element of a partition of W_{DB} (the partition is formed by varying the values of additional constraints derived from models in $W_{\mathrm{U(DB)}}$). Then this P-model w^u is of course a correct representation of the set W_{DB_u}. If this is the case, only a minimum amount of information is necessary to change from the set W_{DB_u} to the model w^u and only a minimum of information is contained in U_{DB}. This means that statistically all models in $W_{\mathrm{U(DB)}}$ have a special representation status.

4.5 Example (Model Quantification)

Consider an urn with N balls, R of which are red. Let us take out n balls without replacement. What is the most probable frequency of red balls in the sample to expect? We model this question with a Hypergeometric distribution and we count the maximum of models in the case of Independence (as to expect with the Independence map).

4.6 Summary (Independence)

Beside the important argument of reducing complexity two more arguments (intuitive graphical representation, quantification of possible worlds) justify the use of the principle of Independence when decisions are necessary. All demands of Independence, contained in U_{DB}, describe constraints of only little information-theoretic value to the problem; if the decisions are based on the method of MaxEnt, these constraints in U_{DB} have no influence on the decisions. So assumptions of Independence can be informative or not, depending on their relation to the I-map of the constraints. The consistency (i.e. $V_{\mathrm{DB}} \neq \emptyset \Rightarrow V_{\mathrm{U(DB)}} \neq \emptyset$) of this principle can be proven by the MaxEnt-Model, which fulfils all the non-linear equations of U_{DB} (what means that the MaxEnt-Model is an element of $V_{\mathrm{U(DB)}}$). The set U_{DB} (resp. the I-maps) will clarify the relation between MaxEnt and Independence. The decisions based on P-models and the principle of Independence are of

[13] an exact knowledge of this is useful, when the solution of a problem should be found by computers itself. This knowledge allows to separate "active" (independence) constraints from "inactive" constraints. The active constraints are necessary for the system, because they will change the result of the reasoning process, the inactive ones are fulfilled anyway by the reasoning process

course strictly stronger than those based on pure P-models. The decisions have already the property of being non-monotonic, when additional information gets available.

5 Conclusions on P-models with Indifference and Independence

It can be shown that a system using both the principle of Indifference and the principle of Independence concludes strictly stronger than the systems with the isolated principles. An example for this is again Simpson's Paradox: both conclusions of DB_2 become true in the joined system, but they are not supported in the single systems.

6 Conclusions on P-models with MaxEnt

6.1 Mathematical formulation

Having its origin in thermodynamics the concept of entropy plays a very important role in the description of irreversible events. As we can put the main emphasis for instance on an energetic or an information theoretic point of view there is a whole family of different concepts of entropy.

For the purpose of nonmonotic reasoning we chose the information theoretic aspect: Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ and $v = (v_1, \ldots, v_n) \in V_\Omega$. According to Shannon (1949) the entropy of v is given by the average number of binary decisions that is necessary to determine a certain element ω_i of Ω, if ω_i was selected a priori with probability v_i. An axiomatic approach is given by

Definition and Theorem 6: Let $v = (v_1, \ldots, v_n)$ be a P-vector[14]. The **entropy $\boldsymbol{H_n(v)}$ of $\boldsymbol{v}$** is characterized by the following properties:

(P1) $H_n : [0,1]^n \to \mathbb{R}\,, v \mapsto H_n(v)$ is a real valued function that is continuous in any argument v_i.

(P2) $\left\{H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)\right\}_{n \in \mathbb{N}}$ is an isotonic sequence in n.

(P3) For any given $\tau < n$ and $1 \le k_i \le \tau$ with $\sum\limits_{i=1}^{\tau} k_i = n$ the so-called **decomposition law** holds, that is:

$$
\begin{aligned}
&H_n(v_1, v_2, \ldots, v_n) \\
&\quad = \; H_\tau(y_1, y_2, \ldots, y_\tau) + \sum_{i=1}^{\tau} y_i \cdot H_{k_i - k_{i-1}}\left(\frac{v_{k_{i-1}+1}}{y_i}, \frac{v_{k_{i-1}+2}}{y_i}, \ldots, \frac{v_{k_i}}{y_i}\right),
\end{aligned}
$$

where $k_0 := 0$, $y_1 := v_1 + v_2 + \ldots + v_{k_1}$, $y_2 := v_{k_1+1} + v_{k_1+2} + \ldots + v_{k_2}$, $\ldots$, $y_\tau := v_{k_{\tau-1}+1} + v_{k_{\tau-1}+2} + \ldots + v_n$.

---

[14] i.e. $\sum\limits_{i=1}^{n} v_i = 1$ and $0 \le v_i \le 1$ for $1 \le i \le n$

It can be shown that the form of H_n is uniquely determined by the properties given above:

$$H_n(v) = -\sum_{i=1}^{n} v_i \cdot \log(v_i)$$

with an arbitrary logarithmic function log.

■

Further properties of $H_n(v)$:

(P4) H_n is a continuous and strictly convex function.

(P5) $H_n(v) \geq 0$ for all P-vectors v.

(P6) H_n has a unique maximum with value $\log(n)$ at the point $\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$.

(P7) $H_n(1,0,\ldots,0) = H_n(0,1,0,\ldots,0) = \ldots = H_n(0,\ldots,0,1) = 0$.

(P8) $H_{n+1}(v_1,\ldots,v_n,0) = H_n(v_1,\ldots,v_n)$.

(P9) $H_n(v) = H_n(\Pi(v))$ for an arbitrary permutation Π of the components of v.

As already stated in the previous sections we have to add additional contraints (i.e. information) to most of the inference problems with incomplete knowledge in order to get a unique answer.

By (P6) the discrete uniform distribution has the highest entropy value. On the other hand this distribution has the lowest variation and therefore the lowest amount of additional information between single elementary probabilities. In consequence we will try to find the P-vector with the highest entropy value[15] that is still consistent with the given a priori information. Two questions immediately arise:

a) Is the solution uniquely determined? Otherwise we would probably need a stronger principle that superceeds the principle of maximum entropy.

b) Does a solution even exist?

As shown in section 1.2 a priori information (i.e. the constraints in DB) can be notated in the form of linear constraints, in symbols $M \cdot v = b$, where $M \in \mathbb{R}^{m,n}$ and $b \in \mathbb{R}^m$. This leads to the following optimization problem which we call **MaxEnt-problem**:

$$\max_{v \in \mathbb{R}^n} \left\{ H_n(v) \mid M \cdot v = b\,,\ \sum_{i=1}^{n} v_i = 1\,,\ v_i \geq 0\,,\ i = 1,\ldots,n \right\}$$

Now we are able to give an answer to the previous questions:

Theorem 7: There is a unique solution of the MaxEnt-problem if the constraints in DB are consistent.

Proof: Immediately by the theorem of Kuhn-Tucker ([Luenberger, 1973]) and the strict convexity of H_n.

■

[15] i.e. the P-vector which is closest to the discrete uniform distribution

The exact handling of arbitrary sentences ($\delta \in I$, see definition 3) by combining problems of linear equalities (fixed δ) will be given in [Greiner & Schramm, 1995].

6.2 Model Quantification

MaxEnt has the best possible justification for decisions by the

Concentration Theorem: Let E be a random experiment with n possible results $\{E_1, \ldots, E_n\}$ and Ω be the set of possible results when E is repeated N times, i.e. $\Omega = \{\omega_1, \ldots, \omega_{n^N}\}$ where $\omega_i \in \{E_1, \ldots, E_n\}^N$ for $1 \leq i \leq n^N$. Let W_Ω be the set of all probability distributions $v_i = (v_{1i}, \ldots, v_{ni})$ with

$$v_{ji} := \frac{\text{total number of } E_j \text{ in } \omega_i}{N}$$

and $W_{\mathrm{DB}} (\subseteq W_\Omega)$ the set of distributions that fulfil a given set of m linear and independent constraints. Further let v^* be the distribution with maximum entropy in W_{DB}, $I := [H_n(v^*) - \Delta H, H_n(v^*)]$ and $1 \leq z \leq |W_{\mathrm{DB}}|$. Then asymptotically the following equation holds:

$$2 \cdot N \cdot \Delta H \;=\; \chi^2_{n-1-m; z/|W_{\mathrm{DB}}|},$$

where $\chi^2_{f;\alpha}$ denotes the α-quantile of the χ^2-distribution with f degrees of freedom.

Proof: [Jaynes, 1982].

■

Therefore by increasing the number N of random experiments more and more feasible P-models (i.e. elements of W_{DB}) will be in a infinitely small vicinity around the MaxEnt-distribution v^* (as H_n is a continuous function). In other words: With high probability a randomly chosen feasible P-model will be very close to v^*.

6.3 Relations between MaxEnt and Indifference/Independence

Theorem 8: MaxEnt complies with the demands of the principle of Indifference.

Proof: Without loss of generality the principle of Indifference may demand $v_i = v_j$ for two indices $i \neq j$. Let Π be the corresponding permutation matrix and v^* be the solution of the MaxEnt-problem. Suppose that $v_i^* \neq v_j^*$. As the principle of Indifference holds for v_i and v_j in V_{DB}, there is $\Pi(v^*) \in V_{\mathrm{DB}}$ with $v^* \neq \Pi(v^*)$ and $H_n(v^*) = H_n(\Pi(v^*))$ by (P9). However, this is a contradiction to theorem 7. Therefore $v_i^* = v_j^*$ holds.

■

Theorem 9: MaxEnt complies with the demands of the principle of Independence.

Idea of the proof: All the equations in U_{DB} have the form

$$v_i \cdot v_j = v_k \cdot v_\ell \, .$$

Sufficient for this equation to hold is the validity of

$$e_{\rho i} + e_{\rho j} = e_{\rho k} + e_{\rho \ell}$$

for all elements $e_{\rho\nu}$ of the matrix M, which can easily be shown by using the independence propositions of undirected graphs (see [Greiner & Schramm, 1994]).

∎

Theorem 10: MaxEnt decides strictly stronger than the joined principles of Indifference and Independence.

Proof: The proofs of theorem 8 and 9 show that both theorems are independent of each other. Therefore MaxEnt complies with the joined principles. The proof is completed by the fact that $V_{\mathrm{IU(DB)}}$[16] contains in most cases more than one P-vector (i.e. infinitely many). For example consider $\mathrm{DB}_6 := \{ \langle P(b \mid a) = 0.8 \rangle \}$ and the desired conclusion $\mathrm{DB}_6 \mid\!\mid\!\sim \langle P(\neg a) = \delta;\ \delta \in (0.5, 1] \rangle$: The principle of Indifference demands $P(\neg ab) = P(\neg a \neg b)$, the principle of Independence does not demand any new equation. The MaxEnt-solution

$$(P(ab) = 0.362\,,\ P(a\neg b) = 0.090\,,\ P(\neg ab) = 0.274\,,\ P(\neg a \neg b) = 0.274) \in V_{\mathrm{IU(DB)}}$$

supports the conclusion, whereas

$$(P(ab) = 0.640\,,\ P(a\neg b) = 0.160\,,\ P(\neg ab) = 0.100\,,\ P(\neg a \neg b) = 0.100) \in V_{\mathrm{IU(DB)}}$$

does not.

∎

7 Conclusions

The 5 logics (P-models, P-models with Indifference, P-models with Independence, P-models with both principles, P-models with MaxEnt) do not only clarify some theoretical relations between MaxEnt and these principles; they make sense by their own and are not an ad hoc concept: When applied to a special set of benchmarks for non-monotonic logics, collected by V. Lifschitz, each logic can infer some of the problems (MaxEnt, being strictly stronger, solves of course nearly all problems). This gives additional information about a problem; it makes explicit which assumptions are necessary to reach the desired conclusions. Concerning our background aim of modelling common sense reasoning we don't argue that in every day reasoning humans calculate the MaxEnt-distribution. Rather we argue that this is the formal solution of a general problem, parts of which might be solved informally (with less accuracy) very fast; a first idea for this is given by the qualitative reasoning in undirected graphs.

[16] i.e. the set of all P-vectors, which fulfil the constraints in DB and the equations in I_{DB} and U_{DB}

References

[Adams, 1975] E.W. Adams, "The Logic of Conditionals", D.Reidel Dordrecht Netherlands, 1975.

[Bacchus et al., 1994] F. Bacchus, A.J. Grove, J.Y. Halpern, D. Koller, "From Statistical Knowledge Bases to Degrees of Belief", *Technical Report* (available via ftp at `logos.uwaterloo.ca:/pub/bacchus`), 1994.

[Blyth, 1973] C. Blyth, "Simpson's Paradox und mutually favourable Events", *Journal of the American Statistical Association*, Vol. 68, p. 746, 1973.

[Cheeseman, 1988] P. Cheeseman, "An Inquiry into Computer Understanding", *Computational Intelligence*, Vol. 4, pp. 58-66, 1988.

[Chow & Teicher, 1978] Y.S. Chow, H. Teicher, "Probability Theory – Independence, Interchangeability, Martingales", Springer, New York, Heidelberg, Berlin.

[Cox, 1979] R.T. Cox, "Of Inference and Inquiry – An Essay in Inductive Logic", in: *The Maximum Entropy Formalism*, MIT Press, ed.: Levine & Tribus, pp. 119-167, 1979.

[Greiner & Schramm, 1994] M. Greiner, M. Schramm,"Nicht-Monotones Schließen auf Wahrscheinlichkeitmodellen, begründet durch die Prinzipien Indifferenz, Unabhängigkeit und maximale Entropie", Institut für Informatik der Technischen Universität München, Bericht I9440.

[Greiner & Schramm, 1995] M. Greiner, M. Schramm,"Non-Monotonic Reasoning on Probability Models: Indifference, Independence & MaxEnt. Part II - Algorithmic Approach by Numerical Optimization". To appear in the 2nd quarter of 1995.

[Howson & Urbach, 1993] C. Howson, P. Urbach,"Scientific Reasoning: The Bayesian Approach", 2nd Edition, Open Court, 1993.

[Jaynes, 1978] E.T. Jaynes, "Where do we stand on Maximum Entropy?", 1978, in: E.T. Jaynes: *Papers on Probability, Statistics and Statistical Physics*, pp. 210-314, Kluwer Academic Publishers, ed.: R.D. Rosenkrantz, 1989.

[Jaynes, 1982] E.T. Jaynes, "On the Rationale of Maximum-Entropy Methods", *Proceedings of the IEEE*, Vol. 70, No. 9, pp. 939-952, 1982.

[Lewis, 1959] P.M. Lewis, "Approximating Probability Distributions to Reduce Storage Requirements", *Information and Control 2*, pp. 214-225, 1959.

[Lifschitz, 1988] V. Lifschitz, "Benchmark Problems for Formal nonmonotonic Reasoning", *Lecture Notes in Artificial Intelligence Non-Monotonic Reasoning*, Vol. 346, pp. 202-219, ed.: Reinfrank et al., 1988.

[Luenberger, 1973] D.G. Luenberger, "Introduction to linear and nonlinear programming", Addison-Wesley, Reading, MA.

[Neapolitan, 1990] R.E. Neapolitan,"Probabilistic Reasoning in Expert Systems: Theory and Algorithms", John Wiley & Sons, 1990.

[Neufeld & Horton, 1990] E. Neufeld, J.D. Horton,"Conditioning on disjunctive knowledge: Simpson's paradox in default logic", *Uncertainty in Artificial Intelligence*

5, pp. 117-125, Elsevier Science, ed.: M. Henrion, R.D. Shachter, L.N. Kanal, J.F. Lemmer, 1990.

[Paris & Vencovska, 1989] J.B. Paris, A. Vencovska, "On the Applicability of Maximum Entropy to Inexact Reasoning", *Int. Journal of approximate reasoning*, Vol. 3, pp. 1-34, 1989.

[Paris & Vencovska, 1990] J.B. Paris, A. Vencovska, "A note on the Inevitability of Maximum Entropy", *Int. Journal of approximate reasoning*, Vol. 4, pp. 183-223, 1990.

[Pearl, 1988] J. Pearl, "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference", Kaufmann, San Mateo, CA, 1988.

[Shore & Johnson, 1980] J.E. Shore, R.W. Johnson, "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross Entropy", *IEEE Transactions on Information Theory*, Vol. IT-26, No. 1, pp. 26-37, 1980.

[Skilling, 1988] J. Skilling, "The Axioms of Maximum Entropy, Maximum-Entropy and Bayesian Methods in Science and Engineering, Vol. 1 – Foundations", Kluwer Academic, ed.: G.J. Erickson, C.R. Smith, Seattle Univ. Washington, 1988.

About the authors

Manfred Schramm (e-mail: `schramma@informatik.tu-muenchen.de`) has been working since 1990 in an automated reasoning group. His research interests include reasoning with incomplete knowledge, common sense reasoning, logics for belief and knowledge, non-monotonic reasoning and modelling with probability.

Michael Greiner (e-mail: `greiner@informatik.tu-muenchen.de`) has been working since 1992 on the comparison and development of numerical optimization methods for the performance analysis of computer systems. His research interests include probability theory and genetic algorithms.

This short article is based on the technical report [Greiner & Schramm, 1994].