[**To appear in:** *Proceedings of the 14th Symposium on Theoretical Aspects of Computer Science (STACS 97)*(6.12.96)]

A conditional-logical approach to minimum cross-entropy

Gabriele Kern-Isberner

FernUniversitaet Hagen,Fachbereich Informatik
P.O. Box 940, D-58084 Hagen, Germany
e-mail: gabriele.kern-isberner@fernuni-hagen.de

Abstract. The principle of minimum cross-entropy (ME-principle) is often used in the AI-areas of knowledge representation and uncertain reasoning as an elegant and powerful tool to build up complete probability distributions when only partial knowledge is available. The inputs it may be applied to are a prior distribution P and some new information $\mathcal{R}$, and it yields as a result the one distribution P^* that satisfies $\mathcal{R}$ and is closest to P in an information-theoretic sense. More generally, it provides a "best" solution to the problem "How to adjust P to $\mathcal{R}$?"
In this paper, we show in a rather direct and constructive manner that adjusting P to $\mathcal{R}$ by means of this principle follows a simple and intelligible conditional-logical pattern. The scheme that underlies ME-adjustment is made obvious, and in a generalized form, it provides a straightforward conditional-logical approach to the adaptation problem. We introduce the idea of a *functional concept* and show how the demands for *logical consistency* and *representation invariance* influence the functions involved. Finally, the ME-distribution arises as the only solution which follows the simple adaptation scheme given and satisfies these three assumptions. So a characterization of the ME-principle within a conditional-logical framework will have been achieved, and its logical mechanisms will be revealed clearly.

1 Introduction

Uncertain reasoning on the base of probability theory has a long tradition. From its early beginnings, probability theory has been regarded as an appropriate means to handle quantified degrees of (un)certainty in a logically consistent and sound way. In the fourties of this century, R.T.Cox [4] proved a fundamental connection between probability and conditionals: if any "measure of credibility" is to be attached to the conclusion B of a conditional statement "if A then B" presuming the antecedent A to be true, and if that measure is to follow two basic assumptions on commonsense inference, then it necessarily obeys the rules of probability.

Within the last decades, probabilistic inference has received increasing attention in the area of artificial intelligence. Its logical soundness provides a solid foundation for nonmonotonic reasoning methods (for a survey, cf. [6]; cf. [14], [2]), and probabilistic expert systems make use of the easy and consistent computability of (quantified) uncertainty (cf.[11], [15]). Usually, probabilistic knowledge is represented by a probability distribution P or by a system of compatible distributions over a set of (discrete or continous) variables. If evidence E is present, the knowledge base is instantiated by calculating the conditional probability $P(\cdot|\mathrm{E})$, applying the principle of direct inference (cf. [2, ch.5.1]). Thus the notion of conditionals is central to probabilistic inference, and vice versa, as the results of Cox cited above show.

Given a (consistent) set $\mathcal{R}$ of conditionals, each equipped with a probability, however, there are generally arbitrary many distributions which all fulfill the probabilistic conditionals in $\mathcal{R}$. Which of them should be chosen to be the "most adequate" one, which of them yields "best" inferences?

From a *statistical and information-theoretical* point of view, the answer should be clear. Many authors (cf. e.g. [5], [7]) argued in eloquent words as well as by proving important statistical properties, that representation of probabilistic knowledge should be performed via the *principle of maximum entropy.* The *entropy* $H(P) = -\sum_\omega p(\omega)\log p(\omega)$ (where the sum is taken over all elementary events ω) of a distribution P measures the uncertainty inherent to P. The idea of that principle is as follows: If the distribution P^* has maximum entropy under the constraints given, then P^* represents $\mathcal{R}$ most adequately, without any external information being added. This approach can be generalized to cases where prior knowledge P has to be taken into account. Here instead of (absolute) entropy the notion of *cross-entropy* (also called *relative entropy*) $R(Q,P) = \sum_\omega q(\omega)\log\frac{q(\omega)}{p(\omega)}$ as a measure of dissimilarity between two distributions Q and P is used. The *principle of minimum cross-entropy* states that if P^* solves

$$\min R(Q,P) = \sum_\omega q(\omega)\log\frac{q(\omega)}{p(\omega)} \tag{1}$$

s.t. Q is a probability distribution which satisfies $\mathcal{R}$

then P^* constitutes the optimal adaptation of the prior P to the new conditional knowledge in $\mathcal{R}$. We refer to both principles as the *ME-principle*, where the abbreviation *ME* stands both for *M*inimum cross-*E*ntropy and for *M*aximum *E*ntropy.

Two articles [17] and [13] provide essential insights into the *logical* aspects of ME-inference. Both show that minimizing cross-entropy resp. maximizing entropy may serve as a logically consistent inference pattern. Shore and Johnson [17] [8] succeeded in characterizing (cross-)entropy as the only functional whose optimization satisfies four (resp. five) fundamental axioms of probabilistic inference. A similar result is proved for entropy in [13] by Paris and Vencovská without assuming that inference is performed by optimizing a functional but relying heavily on solving linear equational systems.

The aim of our paper is to develop a characterization of ME-inference within a completely logical framework. We are going to start from a rather *simple and constructive conditional-logical scheme* for adjusting a prior distribution P to new conditional information $\mathcal{R}$ to yield some posterior distribution P^* which satisfies $\mathcal{R}$. This approach makes things explicit and tangible, and there will be no need to make use neither of optimization theory nor of linear equational systems. The ME-solution is a special instance of that scheme, but generally there are a lot of other solutions of that type. Then we will assume that these solutions arise from an *implicit functional concept* which is to represent the underlying logical inference pattern, and we will refine our scheme slightly to make this functional concept explicit. Now only two further postulates - that of *representation invariance* and a fundamental *logical consistency* - will be necessary to uniquely determine the special type of the solution. And we will prove that the ME-solution is indeed the only solution of that type.

Thus the aim of this paper will be achieved. The essential and striking benefit of the present paper compared to the appreciated works of Shore and Johnson and of Paris and Vencovská is that it thoroughly lays bare the logical mechanisms at work in ME-inference. Using a constructive and intuitive approach, it develops step by step the road to the entropy-optimal solution, making all assumptions explicit and intelligible. Thus the conditional-logical point of view provides a most clear and concise characterization of ME-inferences.

2 Probabilistic conditionals

We consider probability distributions P over a finite set $\mathcal{V} = \{V_1, V_2, V_3, \ldots\}$ of propositional variables V_i which are assumed to be binary. The dotted *literal* $\dot{v}_i \in \{v_i, \bar{v}_i\}$ stands for one of the two possible outcomes of the corresponding variable: v_i is to symbolize "V_i is true", and negation is indicated by barring, i.e. $\bar{v}_i = \neg v_i$. P is uniquely determined by the values of its probability function p applied to the elementary events $\omega = \dot{v}_1\dot{v}_2\dot{v}_3\ldots, p(\omega) = p(\dot{v}_1\dot{v}_2\dot{v}_3\ldots) = P(V_1 = \dot{v}_1, V_2 = \dot{v}_2, V_3 = \dot{v}_3, \ldots)$.

A propositional language $\mathcal{L} = \mathcal{L}(\mathcal{V})$ is defined in the usual way, using the letters $\mathcal{V}$ and the classical connectives $\wedge$ (resp. juxtaposition) and $\neg$. Its formulas are denoted by capital roman letters A,B,. .. To each propositional formula $\mathrm{A} \in \mathcal{L}$ a probability may be assigned via $p(\mathrm{A}) = \sum_{\omega:\mathrm{A}(\omega)=1} p(\omega)$, where the sum is taken over all elementary events ω, and $\mathrm{A}(\omega) = 1$ means that the complete conjunction corresponding to ω is a disjunct in the disjunctive normal form of A. Thus the correspondence between complete conjunctions and elementary events induces a probabilistic interpretation of $\mathcal{L}$ (on the base of the distribution P).

A *probabilistic conditional* (or a *probabilistic rule*) is an expression $\mathrm{A} \rightsquigarrow \mathrm{B}[x]$ with *antecedent* $\mathrm{A} \in \mathcal{L}$, *conclusion* $\mathrm{B} \in \mathcal{L}$ and probability $x \in [0,1]$. A *probabilistic fact* has the form $\mathrm{B}[x]$, $\mathrm{B} \in \mathcal{L}$, $x \in [0,1]$ and is considered to be equivalent to the conditional $\top \rightsquigarrow \mathrm{B}[x]$, where $\top$ represents a tautology. Let $\mathcal{L}^*$ denote the set of all probabilistic conditionals over $\mathcal{L}$. A semantical interpretation of probabilistic conditionals is given by conditional probabilities: If P is a distribution, we write $P \models \mathrm{A} \rightsquigarrow \mathrm{B}[x]$ iff $p(\mathrm{B}|\mathrm{A}) = \frac{p(\mathrm{AB})}{p(\mathrm{A})} = x$. In general, we have $x = p(\mathrm{B}|\mathrm{A}) = \frac{1}{1+\frac{p(\mathrm{A}\overline{\mathrm{B}})}{p(\mathrm{AB})}}$ iff $\frac{p(\mathrm{AB})}{p(\mathrm{A}\overline{\mathrm{B}})} = \frac{x}{1-x}$, so the quotient $\frac{p(\mathrm{AB})}{p(\mathrm{A}\overline{\mathrm{B}})}$ determines the probability of the conditional $\mathrm{A} \rightsquigarrow \mathrm{B}$. It represents the proportion of individuals or objects with property A which also have property B to those that B is not true of. Thus it is crucial for the acceptability of the conditional, not only within a probabilistic framework (cf. [12]).

The problem this paper is going to deal with can now be described in a formal manner:

> *Adjustment problem* (*)
> Given a prior distribution P and some set of probabilistic conditionals $\mathcal{R} = \{\mathrm{A}_1 \rightsquigarrow \mathrm{B}_1[x_1], \ldots, \mathrm{A}_n \rightsquigarrow \mathrm{B}_n[x_n]\} \subset \mathcal{L}^*$, how should P be modified to yield a posterior distribution P^* with $P^* \models \mathcal{R}$?

For (*) to be solvable at all, we now assume the prior P to be positive, i.e. $p(\omega) > 0$ for all complete conjunctions ω, and the set $\mathcal{R}$ to be consistent, that means there shall be some distribution Q that satisfies all rules in $\mathcal{R}$.

3 The principle of minimum cross-entropy

Cross-entropy is a well-known information-theoretic measure of dissimilarity between two distributions and has been studied extensively (for a brief, but informative introduction and further references cf. [16]; cf. [18]). It is also called *directed divergence* for it lacks symmetry, i.e. $R(Q, P)$ and $R(P, Q)$ differ in general, so it is not a metric. But cross-entropy is *positive*, that means we have $R(Q, P) \geq 0$, and $R(Q, P) = 0$ iff $Q = P$ (cf. [5], [16]).

If the adjustment problem (*) is solvable at all, then it has a solution P_e with minimal relative entropy to the prior P (cf. [5]), i.e. P_e solves (1). The condition $Q \models \mathcal{R}$ imposed on a distribution Q can be transformed equivalently into a system of linear equality constraints for the probabilities $q(\omega)$. Using the Lagrangian techniques, we may represent P_e in the form

$$p_e(\omega) = \alpha_0 p(\omega) \prod_{i:\mathrm{A}_i\mathrm{B}_i(\omega)=1} \alpha_i^{1-x_i} \prod_{i:\mathrm{A}_i\overline{\mathrm{B}_i}(\omega)=1} \alpha_i^{-x_i} \tag{2}$$

with the α_i's being exponentials of the Lagrange multipliers, one for each conditional in $\mathcal{R}$, and $\alpha_0 = \exp(\lambda_0 - 1)$, where λ_0 is the Lagrange multiplier of the constraint $\sum_\omega q(\omega) = 1$.

By construction, P_e satisfies all conditionals in $\mathcal{R}$: $p_e(\mathrm{B}_i|\mathrm{A}_i) = x_i, 1 \leq i \leq n$. So $\alpha_1, \ldots, \alpha_n$ are solutions of the nonlinear equations

$$\alpha_i = \frac{x_i}{1-x_i} \frac{\sum\limits_{\omega:\mathrm{A}_i\overline{\mathrm{B}_i}(\omega)=1} p(\omega) \prod\limits_{\substack{j\neq i \\ \mathrm{A}_j\mathrm{B}_j(\omega)=1}} \alpha_j^{1-x_j} \prod\limits_{\substack{j\neq i \\ \mathrm{A}_j\overline{\mathrm{B}_j}(\omega)=1}} \alpha_j^{-x_j}}{\sum\limits_{\omega:\mathrm{A}_i\mathrm{B}_i(\omega)=1} p(\omega) \prod\limits_{\substack{j\neq i \\ \mathrm{A}_j\mathrm{B}_j(\omega)=1}} \alpha_j^{1-x_j} \prod\limits_{\substack{j\neq i \\ \mathrm{A}_j\overline{\mathrm{B}_j}(\omega)=1}} \alpha_j^{-x_j}}, \tag{3}$$

$1 \leq i \leq n$, and α_0 arises as a normalization factor. If $x_i = 1$ for some i, (2) and (3) may still be applied by setting $\infty^0 = 1$ and $\infty^{-1} = 0$.

In spite of the deterrent complexity of the formulas above, (2) shows rather clearly how the ME-adjustment to a rule is carried out: Apart from the normalization factor α_0, only the probabilities of those complete conjunctions ω are changed which satisfy the antecedent of this rule. Furthermore, the new probability depends also on whether ω satisfies the conclusion or not. In particular, the probabilities of all conditionals valid in P whose antecedents are not touched by any of the rules in $\mathcal{R}$ remain unchanged. This means that the ME-adaptation respects one of the most fundamental principles of conditional logics: Asserting a conditional should only affect the knowledge about states which the conditional may be applied to.

This intuitive and reasonable *principle of conditional preservation* will give rise to the adaptation scheme which will be presented in the next section.

But first, we are going to illustrate the use of the ME-approach and the benefits of the representation formulas (2) and (3) by two simple but informative examples. The first example shows knowledge processing in the case of conflicting information, whereas the second example deals with transitive inference. All numerical results were obtained by using the probabilistic expert system SPIRIT which realizes entropy-optimal knowledge processing (cf. [15]).

Example 1. A knowledge base is to be built up representing "Typically, students are adults", "Usually, adults are employed" and "Mostly, students are not employed" with degrees of uncertainty 0.99, 0.8 and 0.9, respectively. Let A, S, E denote the propositional variables A=Being an Adult, S=Being a Student, and E=Being Employed. The quantified conditional information may be written as $\mathcal{R} = \{s \rightsquigarrow a[x_1], a \rightsquigarrow e[x_2], s \rightsquigarrow e[x_3]\}$, $x_1 = 0.9, x_2 = 0.8, x_3 = 0.1$. No prior information is at hand, so we start from the uniform distribution. We are interested in the probability of the conditional $as \rightsquigarrow e$ the antecedent of which combines the evidences a and s conflicting with respect to E.

SPIRIT now tells us $p_e\,(e|as) = 0.1009$, so the more specific information s dominates a, as we should expect. If we set $x_1 = 1$, defining the set of students to be definitely a subset of the set of adults, this preference of the more specific knowledge conveyed by the third rule may be proved in general. If P'_e denotes the ME-solution to the same problem with $x_1 = 1$ instead of $x_1 = 0.99$, we obtain $p'_e\,(e|as) = x_3$ by using (2) and (3).

Example 2. Let denote Y, S, C denote the three propositional variables Y=Being Young, S= Being Single and C=Having Children. We know (or assume) that young people are usually singles (with probability 0.9) and that mostly, singles do not have children (with probability 0.85). Here we have
$\mathcal{R} = \{y \rightsquigarrow s[x_1], s \rightsquigarrow \bar{c}[x_2]\}$ with $x_1 = 0.9, x_2 = 0.85$. Again we take the uniform distribution as prior information. A calculation with SPIRIT shows $y \rightsquigarrow \bar{c}[0.815]$, connecting both rules transitively. By use of the formulas (2) and (3), a more general transitive inference rule can be proved: For arbitrary x_1, x_2, we obtain for the ME-distribution P_e: $p_e\,(\bar{c}|y) = \frac{1}{2}\,(1 + 2x_1x_2 - x_1)$. Of course, the correctness of this formula is independent of the particular meanings of the propositional variables involved. So it states a general transitive inference rule for problems with an analogous knowledge structure.

4 A simple adaptation scheme preserving conditional structures

As we saw earlier in Section 2, ratios of probabilities determine the probabilities of conditionals. If the conditional knowledge inherent to P is to be preserved by the adjustment process (*) as far as possible, we have to preserve ratios of probabilities, the most fundamental of which being probabilities of elementary events.

Following Calabrese [3], a conditional $\mathrm{A} \rightsquigarrow \mathrm{B}$ can be represented as an *indicator function* $(\mathrm{B}|\mathrm{A})$ on elementary events, setting $(\mathrm{B}|\mathrm{A})(\omega) = \begin{cases} 1 & : \ \omega \in \mathrm{AB} \\ 0 & : \ \omega \in \mathrm{A\overline{B}} \\ u & : \ \omega \notin \mathrm{A} \end{cases}$ where u stands for *undefined*. This definition captures excellently the non-classical character of conditionals within a probabilistic framework. According to it, a conditional is a function that polarizes AB and $\mathrm{A\overline{B}}$, leaving $\overline{\mathrm{A}}$ untouched. Now we are able to phrase the principle of conditional preservation in a more general form: The adjustment process (*) shall alter a probability ratio only if at least one conditional in $\mathcal{R}$ explicitly gives rise to do so. For two complete conjunctions ω_1, ω_2, this means that we shall have $\frac{p^*(\omega_1)}{p^*(\omega_2)} = \frac{p(\omega_1)}{p(\omega_2)}$ if $(\mathrm{B}_i|\mathrm{A}_i)(\omega_1) = (\mathrm{B}_i|\mathrm{A}_i)(\omega_2)$ for all i. This leads to the following simple approach: Let $\mathcal{R} = \{\mathrm{A}_1 \rightsquigarrow \mathrm{B}_1\,[x_1], \ldots, \mathrm{A}_n \rightsquigarrow \mathrm{B}_n\,[x_n]\} \subset \mathcal{L}^*$ be a set of probabilistic conditionals. An adaptation $P * \mathcal{R}$ of the distribution P to the new conditional information $\mathcal{R}$ is constructionally defined by the scheme

$$p^*(\omega) = p * \mathcal{R}(\omega) = \alpha_0 p(\omega) \prod_{i:\mathrm{A}_i\mathrm{B}_i(\omega)=1} \alpha_i^+ \prod_{i:\mathrm{A}_i\overline{\mathrm{B}_i}(\omega)=1} \alpha_i^- \tag{4}$$

where α_i^+ and α_i^- are non-negative factors chosen in a suitable way so as to ensure that the rules of $\mathcal{R}$ are all valid in the posterior distribution P^* (as usual, the use of the small letter $p * \mathcal{R}$ indicates the probability function belonging to $P * \mathcal{R}$). Applying the constraints $p^*(\mathrm{B}_i|\mathrm{A}_i) = x_i$, $1 \leq i \leq n$, yields

$$\frac{\alpha_i^+}{\alpha_i^-} = \frac{x_i}{1 - x_i} \frac{\sum\limits_{\omega:\mathrm{A}_i\overline{\mathrm{B}_i}(\omega)=1} p(\omega) \prod\limits_{\substack{j \neq i \\ \mathrm{A}_j\mathrm{B}_j(\omega)=1}} \alpha_j^+ \prod\limits_{\substack{j \neq i \\ \mathrm{A}_j\overline{\mathrm{B}_j}(\omega)=1}} \alpha_j^-}{\sum\limits_{\omega:\mathrm{A}_i\mathrm{B}_i(\omega)=1} p(\omega) \prod\limits_{\substack{j \neq i \\ \mathrm{A}_j\mathrm{B}_j(\omega)=1}} \alpha_j^+ \prod\limits_{\substack{j \neq i \\ \mathrm{A}_j\overline{\mathrm{B}_j}(\omega)=1}} \alpha_j^-}, \tag{5}$$

with $\alpha_i^- = 0$ if $x_i = 1$, $1 \leq i \leq n$. Thus to each conditional $\mathrm{A}_i \rightsquigarrow \mathrm{B}_i\,[x_i]$ in $\mathcal{R}$, $1 \leq i \leq n$, a pair of "distortion factors" (α_i^+, α_i^-) is associated which are solutions of the equations (5). α_0 is a normalization factor that makes P^* a probability distribution. $P * \mathcal{R}$ shall denote any distribution of the form (4) which solves the adaptation problem $P^* \models \mathcal{R}$.

In spite of its simplicity and arbitrariness, this adaptation scheme actually has some important properties in common with the ME-approach. In [10] it is shown that it satisfies *system independence* as well as *subset independence* which Shore and Johnson [17] proved to be characteristic of ME-inferences.

5 The functional concept

Generally a solution to our adaptation problem provided by the scheme (4) is far from being unique. If e.g. P is to be adjusted to a nontrivial probabilistic conditional $\mathrm{A} \rightsquigarrow \mathrm{B}\,[x]$ the different solutions (α^+, α^-) of the corresponding equation (5): $\frac{\alpha^+}{\alpha^-} = \frac{x}{1-x}\frac{p(\mathrm{A}\overline{\mathrm{B}})}{p(\mathrm{AB})}$ result in different distributions. Thus the scheme (4) describes how the posterior distribution has to be built up once a solution to the equations (5) is fixed, but up to now any solution is possible.

Of course, this situation is not very satisfactory because we intuitively feel that there are "good" solutions and "bad" solutions. For instance, if the prior distribution P already satisfies the information contained in $\mathcal{R}$ it would be reasonable to expect that the scheme does not alter any probability, yielding the prior as the posterior distribution.

Demanding uniqueness means to assume a *functional concept* that underlies the finding of a "best solution" so that a unique distribution of type (4) arises in dependence of the prior knowledge P and of the new conditional information $\mathcal{R}$.

Returning to the small example given at the beginning of this section, we see that all different solutions (α^+, α^-) have the same quotient $\frac{\alpha^+}{\alpha^-} = \frac{x}{1-x}\frac{p(\mathrm{A}\bar{\mathrm{B}})}{p(\mathrm{AB})}$ which is determined by the constraint (5). In general, however, this need not be the case, different solutions may have different quotients, as can easily be checked when $\mathcal{R}$ contains more than one rule. But in any case, equation (5) reflects for each pair (α_i^+, α_i^-) which is part of a solution the dependencies of its quotient on the other conditionals in $\mathcal{R}$, on the prior knowledge P and – in a distinguished way – on the probability x_i of the conditional it corresponds to. Thus the quotient $\frac{\alpha_i^+}{\alpha_i^-}$ plays a central role, representing the core of the impact that the corresponding conditional is to have on the posterior distribution. If we look for something that should guide us in finding "good" solutions, these quotients seem to be most obvious.

Therefore we now assume the factors α_i^+ and α_i^- to be functionally dependent on their common quotient $\alpha_i := \frac{\alpha_i^+}{\alpha_i^-}$: $\alpha_i^+ = F_i^+(\alpha_i)\,, \alpha_i^- = F_i^-(\alpha_i)$. Of course the functions F_i^+ and F_i^- should not be arbitrary. They should be sufficiently regular and are to follow a pattern independent of the specific form of a rule $\mathrm{A}_i \rightsquigarrow \mathrm{B}_i\,[x_i]$, thus realizing a fundamental inference pattern. This last assumption is being expressed as

$$\alpha_i^+ = F_i^+(\alpha_i) = F^+(x_i, \alpha_i) \quad \text{resp.} \quad \alpha_i^- = F_i^-(\alpha_i) = F^-(x_i, \alpha_i) \tag{6}$$

so as to separate the numerical impact of the inferential strength x_i from the logical interdependencies of the rule represented by α_i. Under these assumptions, the equations (5) may now be transformed into

$$\alpha_i = \frac{x_i}{1-x_i}\frac{\sum\limits_{\omega:\mathrm{A}_i\overline{\mathrm{B}_i}(\omega)=1} p(\omega) \prod\limits_{\substack{j\neq i\\ \mathrm{A}_j\mathrm{B}_j(\omega)=1}} F^+(x_j,\alpha_j) \prod\limits_{\substack{j\neq i\\ \mathrm{A}_j\overline{\mathrm{B}_j}(\omega)=1}} F^-(x_j,\alpha_j)}{\sum\limits_{\omega:\mathrm{A}_i\mathrm{B}_i(\omega)=1} p(\omega) \prod\limits_{\substack{j\neq i\\ \mathrm{A}_j\mathrm{B}_j(\omega)=1}} F^+(x_j,\alpha_j) \prod\limits_{\substack{j\neq i\\ \mathrm{A}_j\overline{\mathrm{B}_j}(\omega)=1}} F^-(x_j,\alpha_j)}, \tag{7}$$

yielding a distribution P^* of the form

$$p^*(\omega) = \alpha_0 p(\omega) \prod_{i:\mathrm{A}_i\mathrm{B}_i(\omega)=1} F^+(x_i, \alpha_i) \prod_{i:\mathrm{A}_i\overline{\mathrm{B}_i}(\omega)=1} F^-(x_i, \alpha_i) \tag{8}$$

which solves the adaptation problem. The functions $F^+(x, \alpha)$ and $F^-(x, \alpha)$ are defined for $x \in [0, 1]$ and for non-negative real α , and they are related by

$$F^+(x, \alpha) = \alpha F^-(x, \alpha). \tag{9}$$

To symbolize the presence of the functional concept, we will use $*_F$ instead of $*$.

6 Logical consistency

Surely, the adaptation scheme (8) will be considered sound only if the resulting posterior distribution can be used as a prior distribution for further adaptations. This is a very fundamental meaning of *logical consistency.*

In particular, if we first adjust P only to a subset $\mathcal{R}_1 \subseteq \mathcal{R}$ and then use this posterior distribution to perform another adaptation to the full conditional information $\mathcal{R}$, we should have the same distribution as if we adjusted P to $\mathcal{R}$ in only one step. More formally, the operator $*_F$ should satisfy the following equation for any positive distribution P and any sets $\mathcal{R}_1, \mathcal{R}_2 \subset \mathcal{L}^*$ (provided that all adaptation problems are solvable):

$$P *_F (\mathcal{R}_1 \cup \mathcal{R}_2) = (P *_F \mathcal{R}_1) *_F (\mathcal{R}_1 \cup \mathcal{R}_2) \tag{10}$$

(10) is the first postulate we demand the adjustment operator $*_F$ to satisfy. According to the reflections above, it will be called the *postulate for logical consistency.* Applying it will give us a first important result in determining the functions F^+ and F^-:

Theorem 1. *If the adjustment operator $*_F$ satisfies the postulate for logical consistency (10) then F^- necessarily fulfills the functional equation*

$$F^-(x, \alpha\beta) = F^-(x, \alpha) F^-(x, \beta) \tag{11}$$

for all $x \in [0, 1]$, $\alpha, \beta \geq 0$.

Because of (9), F^+ satisfies (11) iff F^- does.

Proof. If (10) holds in principle for any adaptation carried out by $*_F$, it is surely valid for some special type of $P, \mathcal{R}_1$ and $\mathcal{R}_2$. So let P be any positive distribution over 3 variables A, B and C, let $\mathcal{R}_1 = \{a \rightsquigarrow c[x]\}$ and $\mathcal{R}_2 = \{b \rightsquigarrow c[y]\}$. Let $p_1, \ldots, p_8$ denote the prior probabilities of P in reverse order, i.e. $p_1 = p(abc), \ldots, p_8 = p(\bar{a}\bar{b}\bar{c})$.

The new distributions $P * (\mathcal{R}_1 \cup \mathcal{R}_2), P * \mathcal{R}_1$ and $(P * \mathcal{R}_1) * (\mathcal{R}_1 \cup \mathcal{R}_2)$ may be set up according to (8), with factors α, β (belonging to $P * (\mathcal{R}_1 \cup \mathcal{R}_2)$), $\tilde{\alpha}$ (belonging to $P * \mathcal{R}_1$) and α_1, β_1 (belonging to $(P * \mathcal{R}_1) * (\mathcal{R}_1 \cup \mathcal{R}_2)$). By equations (7), these factors may be calculated as follows:

$\alpha = \frac{x}{1-x}\frac{p_2F^-(y,\beta)+p_4}{p_1F^+(y,\beta)+p_3}; \quad \beta = \frac{y}{1-y}\frac{p_2F^-(x,\alpha)+p_6}{p_1F^+(x,\alpha)+p_5}; \quad \tilde{\alpha} = \frac{x}{1-x}\frac{p_2+p_4}{p_1+p_3};$
$\alpha_1 = \frac{x}{1-x}\left(\tilde{\alpha}\right)^{-1}\frac{p_2F^-(y,\beta_1)+p_4}{p_1F^+(y,\beta_1)+p_3}; \quad \beta_1 = \frac{y}{1-y}\frac{p_2F^-(x,\tilde{\alpha})F^-(x,\alpha_1)+p_6}{p_1F^+(x,\tilde{\alpha})F^+(x,\alpha_1)+p_5}.$

For $P*(\mathcal{R}_1\cup\mathcal{R}_2) = (P*\mathcal{R}_1)*(\mathcal{R}_1\cup\mathcal{R}_2)$ to hold, as required, we must have $\beta = \beta_1$ and $\alpha = \tilde{\alpha}\alpha_1$. Comparing β to β_1 above, observing (9) and using $\alpha = \tilde{\alpha}\alpha_1$, we obtain $F^-(x,\alpha) = F^-(x,\tilde{\alpha}\alpha_1) = F^-(x,\tilde{\alpha})\,F^-(x,\alpha_1)$. Because P, x and y can be chosen arbitrarily, this equation must hold for all non-negative real $\tilde{\alpha},\alpha_1$ and all $x\in[0,1]$.

The example in the proof of Theorem 1 above was not chosen accidentally. The way in which two conditionals with common conclusion should interact is one of the main issues in conditional logic. Sometimes it is referred to as the *antecedent conjunction problem* (cf. [12]).

Assume x to be held fixed and let for a moment $F_x^-(\alpha) := F^-(x,\alpha)$ be regarded only as a function of α. For real α',β', (11) yields $F_x^-\circ\exp(\alpha'+\beta') = (F_x^-\circ\exp(\alpha'))\,(F_x^-\circ\exp(\beta'))$. Therefore according to [1, p. 140], $F_x^-\circ\exp$ must have the form $F_x^-\circ\exp(\alpha') = e^{c\alpha'}$ for some constant c, and again taking into consideration the dependency on x, we obtain

$$F^-(x,\alpha) = \alpha^{c(x)} \quad \text{and} \quad F^+(x,\alpha) = \alpha^{c(x)+1}, \tag{12}$$

for any positive real α, with a (sufficiently well-behaved) function $c(x)$.

7 Representation invariance

Up to now, we neglected completely how (conditional) knowledge is represented in $\mathcal{R}$. Of course, the way in which we build up the adjusted distribution P^* already allows for the *logical equivalence* of propositional formulas, by definitions (4) and (5). But what about *probabilistic equivalences*, i.e. equivalences that are due to elementary probability calculus? For instance, the sets of rules $\{\mathrm{A} \leadsto \mathrm{B}\,[x],\ \mathrm{A}\,[y]\}$ and $\{\mathrm{AB}\,[xy],\ \mathrm{A}\,[y]\}$ are equivalent in this respect because each rule in one set is derivable from rules in the other. We surely expect the result of our adjustment process to be independent of the syntactic representation of probabilistic knowledge in $\mathcal{R}$:

$$P *_F \mathcal{R} = P *_F \mathcal{R}' \quad \text{if} \quad \mathcal{R} \quad \text{and} \quad \mathcal{R}' \quad \text{are probabilistically equivalent.} \tag{13}$$

(13) is called the *postulate for representation invariance*. It is the second stipulation that we expect the adjustment operator $*_F$ to comply with.

Proposition 2. *If the adaptation operator $*_F$ with F^-, F^+ of the form (12) satisfies the postulate for representation invariance (13) then for all real $x, x_1, x_2 \in [0,1]$ the following equations hold:*

$$c(x) + c(1-x) = -1 \tag{14}$$
$$c(xx_1 + (1-x)x_2) = -c(x)c(x_1) - c(1-x)c(x_2) \tag{15}$$

where $c(x)$ is defined by (12).

Proof. As in the proof of Theorem 1, the treatment of special cases revealing fundamental relationships will imply the desired equations. This proof involves intricate calculations, so we only sketch it by giving the main ideas. A detailed proof may be found in [9].
The most obious probabilistic equivalence is that of each two rules $\mathrm{A} \rightsquigarrow \mathrm{B}\,[x]$ and $\mathrm{A} \rightsquigarrow \overline{\mathrm{B}}\,[1-x]$. Respecting this implies equation (14).
Because of $p\,(b|a) = p\,(b|ac)\,p\,(c|a) + p\,(b|a\bar{c})\,p\,(\bar{c}|a)$ for arbitrarily chosen variables A, B and C, the two sets of rules $\mathcal{R} = \{a \rightsquigarrow c\,[x],\ ac \rightsquigarrow b\,[x_1],\ a\bar{c} \rightsquigarrow b\,[x_2]\}$ and $\mathcal{R}' = \{a \rightsquigarrow b\,[y],\ ac \rightsquigarrow b\,[x_1],\ a\bar{c} \rightsquigarrow b\,[x_2]\}$ with $y = xx_1 + (1-x)x_2$ are probabilistically equivalent for all real $x, x_1, x_2 \in [0,1]$. Setting up the distributions $P *_F \mathcal{R}$ and $P *_F \mathcal{R}'$ according to (8) and postulating their equality shows the second equation (15) by comparing the corresponding factors.

The next theorem is now an easy consequence.

Theorem 3. *If the operator $*_F$ is to meet the fundamental demands for logical consistency (10) and for representation invariance (13), then F^+ and F^- necessarily must be*

$$F^+\,(x,\alpha) = \alpha^{1-x} \quad \textit{and} \quad F^-\,(x,\alpha) = \alpha^{-x}. \tag{16}$$

Proof. Logical consistency implies $F^-\,(x,\alpha) = \alpha^{c(x)}$ (cf. (12)) where $c(x)$ satisfies equations (14) and (15) because of representation invariance. Therefore $c(x)$ fulfills a Cauchy functional equation, and according to [1, p. 44f.], $c(x)$ must be of the form $c(x) = kx$ with $k = c(1) = -1$. This shows (16).

8 Uniqueness

So far we have proved that the demands for logical consistency and for representation invariance uniquely determine the functions which we assumed to underly the adjustment process $P * \mathcal{R}$ in the way we described in Section 5. Applying (16) to (8) and (7) we recognize that the posterior distribution necessarily is of the same type (2) as the ME-distribution if it is to yield sound and consistent inferences.

So we nearly have achieved the aim of this paper, the characterization of ME-inference within a conditional-logical framework. But one step is still missing: Is the distribution which arises as a solution to (2) and (3) uniquely determined? Are there possibly several different solutions of type (2), only one of which being the ME-distribution? If we assume the functions F^- and F^+ to fulfill (16), is this*sufficient* to guarantee that the resulting operator $*_F$ satisfies logical consistency and representation invariance?

The question for uniqueness of the posterior distribution is at the centre of all these problems. If it can be answered positively, we will have reached our goal: The unique posterior distribution of type (2) must be the ME-distribution, $*_F$ then corresponds to ME-inference, and ME-inference is known to fulfill (10) and (13) as well as many other reasonable properties, cf. [13], [17], [18]. Indeed, this uniqueness will be affirmed by the next theorem. It is proved by making use of cross-entropy as an excellently fitting measure of distance for distributions of type (2).

Theorem 4. *There is at most one solution of the adaptation problem $P * \mathcal{R}$ of type (2).*

Proof. Let P_1^*, P_2^* be two distributions of type (2) with non-negative real factors $(\alpha_i)_{0 \leq i \leq n}$, $(\beta_i)_{0 \leq i \leq n}$ fulfilling equations (3). Calculating the cross-entropies between P_1^* and P_2^*, we see $R(P_1^*, P_2^*) = \log \alpha_0 - \log \beta_0$ and $R(P_2^*, P_1^*) = \log \beta_0 - \log \alpha_0$. Both equations together imply $R(P_1^*, P_2^*) = 0$ because cross-entropy is non-negative, and by using its positivity (cf. [16]), both distributions must be equal. This proves the theorem.

The following theorem summarizes our results in characterizing ME-adjustment within a conditional-logical framework:

Theorem 5. *Let $*_e$ denote the ME-adjustment operator, i.e. $*_e$ assigns to a prior distribution P and some set $\mathcal{R}$ of probabilistic conditionals the one distribution $P_e = P *_e \mathcal{R}$ which has minimal cross-entropy with respect to P among all distributions that satisfy $\mathcal{R}$ (provided the adjustment problem is solvable at all).*

Then $*_e$ is the only operator of type (4) that realizes a functional concept (6) and satisfies the fundamental postulates for logical consistency (10) and for representation invariance (13). $*_e$ is completely described by (2) and (3).

9 Conclusion

Starting from a conditional-logical point of view we found a completely logical characterization of the ME-solution to the adjustment problem (*) stated in Section 2. This characterization rests on three axioms: The existence of a functional concept, being described by (6), and the fundamental properties of logical consistency (10) and of representation invariance (13). All of these three principles are usually considered to be fundamental to any reasonable inference procedure. So the ME-solution arises in a very natural way, without imposing any external minimality demand. In fact, minimality is inherent to the idea of a functional concept and to logical consistency, and the simplicity of the scheme (4) may be regarded as minimizing the complextity of the conditional-logical approach.

Actually, there is a fourth axiom present, expressed by the approach (4). In this paper, we gave only an informative explanation for it realizing *conditional preservation.* In [9], a clear and exact formalization of this principle is given, and its equivalence to (4) is proved.

References

1. J. Aczél. *Vorlesungen ueber Funktionalgleichungen und ihre Anwendungen.* Birkhaeuser Verlag, Basel, 1961.

2. F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge: a Logical Approach to Probabilities.* MIT Press, Cambridge, Mass., 1990.
3. P.G. Calabrese. Deduction and inference using conditional logic and probability. In I.R. Goodman, M.M. Gupta, H.T. Nguyen, and G.S. Rogers, editors, *Conditional Logic in Expert Systems*, pages 71–100. Elsevier, North Holland, 1991.
4. R.T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–13, 1946.
5. I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Prob.*, 3:146–158, 1975.
6. M. Goldszmidt. Research issues in qualitative and abstract probability. *AI Magazine*, pages 63–66, Winter 1994.
7. E.T. Jaynes. *Papers on Probability, Statistics and Statistical Physics.* D. Reidel Publishing Company, Dordrecht, Holland, 1983.
8. R.W. Johnson and J.E. Shore. Comments on and correction to "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy". *IEEE Transactions on Information Theory*, IT-29(6):942–943, 1983.
9. G. Kern-Isberner. Characterizing the principle of minimum cross-entropy within a conditional logical framework. Informatik Fachbericht 206, FernUniversitaet Hagen, 1996.
10. G. Kern-Isberner. Conditional logics and entropy. Informatik Fachbericht 203, FernUniversitaet Hagen, 1996.
11. S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities in graphical structures and their applications to expert systems. *Journal of the Royal Statistical Society B*, 50(2):415–448, 1988.
12. D. Nute. *Topics in Conditional Logic.* D. Reidel Publishing Company, Dordrecht, Holland, 1980.
13. J.B. Paris and A. Vencovská. A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 14:183–223, 1990.
14. J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, Ca., 1988.
15. W. Roedder and C.-H. Meyer. Coherent knowledge processing at maximum entropy by spirit. In *Proceedings 12th Conference on Uncertainty in Artificial Intelligence*, pages 470–476, 1996.
16. J.E. Shore. Relative entropy, probabilistic inference and AI. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 211–215. North-Holland, Amsterdam, 1986.
17. J.E. Shore and R.W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26:26–37, 1980.
18. J.E. Shore and R.W. Johnson. Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, IT-27:472–482, 1981.

This article was processed using the LaTeX macro package with LLNCS style