Minimum Cross-Entropy Reasoning: A Statistical Justification

Manfred Jaeger
Max-Planck-Institut fur Informatik
Im Stadtwald
66125 Saarbrueken
Germany

Abstract

Degrees of belief are formed using observed evidence and statistical background information. In this paper we examine the process of how prior degrees of belief derived from the evidence are combined with statistical data to form more specific degrees of belief. A statistical model for this process then is shown to vindicate the cross-entropy minimization principle as a rule for probabilistic default-inference.

1 Introduction

A knowledge based system incorporating reasoning with uncertain information gives rise to quantitative statements of two different kinds: statements expressing statistical information and statements of degrees of belief. "10% of applicants seeking employment at company X who are invited to an interview will get, a job there" is a statistical statement. "The likelihood that I will be invited for an interview if I apply for a job at company X is about 0.6'1 expresses a degree of belief.

In this paper, both of these kinds of statements are regarded as probabilistic, i.e. the numbers appearing in these statements are assumed to obey the rules of probability theory. A degree of belief is viewed as a constraint on a set of possible (subjective) probability values.

A very expressive extension of the language of first-order logic for representing the two types of probabilistic information has been defined by Halpern [1990] and Bacchus [1990]. For the purpose of the present paper, it will be sufficient to restrict attention to a much simpler language based on propositional logic. Adapting the notation of Halpern and Bacchus, we will consider knowledge bases in an extension of the language L(S) of propositional logic over the propositional variables S = $\{V_1, \ldots, V_n\}$ incorporating expressions of the form

$$[\phi \mid \psi] \geq p \qquad \text{(i)}$$

and

$$\mathrm{prob}(\phi(e) \mid \psi(e)) \geq p, \qquad (2)$$

where ϕ, ψ are formulas in L(S), p is a real number in [0,1], and e is a new symbol not in S.

The intended meaning of such a knowledge base is that formulas in the underlying propositional language express certain properties that a given object or, in a wider sense, a given *event* may or may not possess. Formulas of the form (!) make a statement with what statistical probability property in the domain of discourse ϕ holds given that property ' holds, and formulas of the form (2) are used to express that for the specific event e it is believed that the probability of *e* having property ϕ, given that c has property ψ, is $\geq p$.

Using the propositional variables A ($\approxeq$ applicant), I ($\approx$ interview), and J ($\approx$ job), the two example sentences above may be symbolized by [J | A ∧ I] = 0.1 and prob(I(1) A(i)) ∈ [0 55, 0.65]vhich the definite knowledge ¬(I ∧ A) → ¬J ("no job without an application and an interview") might be added. (The symbols — and ∈ here used are definable abbreviations, not an extension of the syntax.)

Semantics for this representation language are given by probability measures on L(S): one probability measure interprets the statistical expressions (1), and one probability measure *v,* is needed for every event symbol *e* used in the knowledge base to interpret the degree of belief expressions (2).

A probability measure on L(S) is fully determined by the probability values of the *N* :— 2" *atoms* of the language, i.e. the expressions in the set

$$\Gamma := \{\wedge_{i=1}^{n} \tilde{V}_i \mid \tilde{V}_i \in \{V_i, \neg V_i\}\}.$$

We denote by $\Delta(\Gamma)$ the set of probability measures on r.

Usually, the constraints (1) on the statistical measure ;;,, and constraints (2) on the belief measure *v*£ given in the knowledge base will only lead to reasonably narrow bounds for probabilities on a few formulas, while others remain largelv undetermined. In this situation we will be looking for a rule of probabilistic inference that selects from the multitude of probability measures consistent with the given constraints the ones that seem to be most reasonable or plausible.

Several types of inference rules here can be distinguished. First, there are those that only apply to sets of constraints on probabilities of a single kind. Since in many applications only one kind of probabilistic information is represented in the knowledge base, this type of inference is the one most frequently encountered. Entropy maximization is the most common rule of this type. A second type of probabilistic inference is given by the random worlds formalism of Bacchus *et al* [1992]. Here constraints on statistical probabilities are used to derive degrees of belief. In [Bacchus *et ai,* 1994] it is also shown how this method can be extended to make the resulting subjective probabilities also depend on given prior degrees of belief. It is this third kind of probabilistic inference, where information about both statistical and subjective probabilities are used to complete the set of degrees of belief, that we, too, are concerned with in this paper.

That statistical information is relevant for the formation of degrees of belief appears to be obvious: given the statements of the introductory example - and no further information whether I'm either more or less likely than everybody else to get a job after an interview - I will conclude that my chances for finding employment at company X are approximately 0.06.

This form of reasoning has been called *default reasoning about probabilities* in [Jaeger, 1994a], because, just as in logical default reasoning, information about what, is generally true in the domain of discourse is used to derive conclusions about specific objects not strictly implied by the knowledge base. In this example the default inference only consisted of {a slightly generalized form of) *direct inference* ([Carnap, 1950]), which is not applicable in more general cases. In [Jaeger, 1994b] it was therefore proposed to use cross-entropy minimization as a generalization of direct inference that is applicable under much more general circumstances. This was motivated mainly by the logical properties that the resulting inference rule possesses - properties that are intuitively reasonable. In the present paper we are taking a more fundamental approach to the issue: first, it is undertaken to provide a precise epistemological analysis of the principles that underly default reasoning about probabilities. From this analysis a concrete statistical model for the interpretation of the probabilistic information will be derived. It then turns out that this model validates the cross-entropy minimization principle for default reasoning about probabilities.

2 The Formation of Beliefs: an Interpretation

In this section we will derive an analysis of the principles underlying default reasoning about probabilities. Two examples will serve as a guide towards this analysis.

Example 2.1 I'm playing a game of dice with a friend who just has made the roll of the die that will decide the game: if she rolls a four or better, she wins; if a three or less turns up, I win. The die has come to rest out of my sight, but the outcome has been observed by my friend. By the somewhat satisfied expression on her face I gather that I will less likely have won than lost this game. "Less likely" I'm here willing to quantify by a probability of 0.3, so that my degree of belief in the current toss t having the property $p := R_1 \vee R_2 \vee R_3$ with $R_1 :=$ " i has turned up" is given by the subjective probability 0.3.

Example 2.2 Scanning channels on TV we tune in to a mystery film /. We just catch the last part of a spectacular car chase, apparently taking place in a European city. These two observations induce us to believe that the film has been an expensive production with probability > 0.7, and is of American origin only with probability $< 0\text{-}5$.

The two uncertain events described in these examples are of a somewhat different nature: the first one is a product of what can only be understood as a random process. The uncertain event in the second example, however, is not random in the classical sense that a toss of a die, or the drawing of a card from a shuffled deck is random. The film, a fragment of which we have happened to see on TV, is not broadcast at that time as a result of being drawn from a gigantic urn containing all mystery films. Given that we are partially ignorant of the deterministic chain of events that led to the screening of that particular film at that particular time, however, for us endows this events with all the features of randomness. Some partial knowledge we may possess of the actual chain of events causing the given observation, and the ignorance about some other of it's parts, combines to the imperfect, perception of that chain of events as a *random mechanism.*

Interpreting an observed uncertain event r as a realization of some random mechanism, provides a means for defining one's degrees of belief: based on our model of a random mechanism, we can consider a long (hypothetical) sequence of events that are independent realizations of the same random mechanism. Moreover, we can imagine all the elements of the sequence to provide us with the same evidence as c. Such an imaginary sequence of events we call a *thought experiment.* Our degree of belief (point- or multi-valued) that e has property φ now can be defined as a bound on the relative frequency with which we imagine events in the thought experiment to have 0.

This interpretation of degrees of belief is the content of the following postulate.

Postulate 1: *The degree of belief that an uncertain event c has property φ is the predicted hound on the relative frequency of φ in a long (imaginary) sequence of events, each of which, is a realization of the random mechanism modeling the chain of events that produced e, and each of which provides the same evidence that has been observed in e.*

Speaking of a "long sequence of events" here is a somewhat sloppy terminology. In principle, a thought experiment must be considered as an infinite sequence of events; degrees of belief are defined by a prediction of the limiting relative frequencies in initial segments of increasing length from this sequence.

By this postulate it is not claimed that we are always able to precisely specify a random mechanism corresponding to the observed event in the sense of reducing it, for example, to the random draw from a set of well-defined alternatives according to well-defined chances.

Our image of the random mechanism may well contain unknown parameters.

For an illustration of this, consider a variation of example 2.1: suppose that I have a vague suspicion that my friend has a loaded die up her sleeve that enables her to roll a six at will, and that she occasionally will use this die instead of the fair one. Finding myself in the same situation as described previously, I will now have to incorporate into my model of the random mechanism that produced the crucial toss of the die the possibility that in that toss the loaded die was in fact supplanted for the fair one. The result might be a model of a random mechanism consisting of first a random draw of one of either a fair or a loaded die, and a subsequent toss of that die. However, feeling unable to evaluate the likelihood for my friend to have cheated at the observed toss, I am unable to specify the respective probabilities for the two dice to be drawn. This makes my thought experiment, depend on an unknown parameter. Depending on it's value, the predicted frequency of *p* will have any value between 0 (always a loaded die is being tossed), and 0.3 (only the fair die is being used). Consequently, my degree of belief in *p(t)* now will be the interval [0,0.3],

In our interpretation, then, the vagueness of a degree of belief in part is caused by an uncertainty about the parameters of the thought experiment.

Postulate 1 gives a semantic interpretation of the meaning of a degree of belief, but does not attempt to give a rule for their computation. Particularly, the question of how to construct a random mechanism for the thought experiment, and how to translate the evidence into a predicted bias for the outcome of realizations of the random mechanism, are outside the scope of the statement made in that postulate. They, too, are outside the scope of this paper, where our immediate concern is with interpreting knowledge bases including statements of degrees of belief, but not containing the primary evidence which initiated these degrees of belief.

The interpretation of degrees of belief here given, however, does provide guidance for finding a specific rule by which degrees of belief stated in the knowledge base should be combined with statistical information.

Example 2.1 (continued): What, in the situation described previously, will be my degree of belief in the proposition R,(t) (i = 1,2,3)? The observation I have made only provides evidence that bears on the probability of *p(t),* but does allow me to discriminate between the three alternatives Ri(t), R_2(t)X R;s(0- However, I do have the information that the statistical probability of each of the R, in tosses of a fair die is 1/6. Specifically, this means that each of the R, has an equal statistical probability. This statistical knowledge determines my prediction of the outcome of the thought experiment associated with the present event *t:* I will expect that here, too, each of the three alternatives R_1,R_2,R_3 will appear with equal frequency 0.3/3 = 0.1. Similarly, for *i* = 4,5,6, a degree of belief 0.7/3 will be assigned to R,(t).

Example 2.2 (continued): While[1] a commercial break has stopped the flow of useful information, we have time to make up our mind whether we want to continue watching that mystery film. Having a preference for films with a happy end, we first attempt to estimate the likelihood for this film to have one. None of the evidence provided in the short scene we have seen directly suggests either a happy or an unhappy ending. Fortunately, however, we do have recourse to statistical information with what relative frequency happy endings have occurred in the great number of mystery films (distinguished by their having combinations of the properties A (American) and E (expensive)) shown on television in the last few years. Using our syntax for the representation of statistical probabilities, let this information consist, of

$$[\mathrm{HE} \mid A \wedge E] = 0.9 \qquad [\mathrm{HE} \mid A \wedge \neg E] = 0.7$$
$$[\mathrm{HE} \mid \neg A \wedge E] = 0\ 7 \qquad [\mathrm{HE} \mid \neg A \wedge \neg E] = 0.5$$

Here it is far from obvious what prediction for the relative frequency of happy endings in the thought experiment we should derive from these statistics and our prior predictions about the frequencies of A and E. It is easy, though, to obtain some bounds for the plausible values of this frequency.

For an upper bound we may suppose that in the hypothetical sequence of mystery films the relative frequency of those of the four properties $A \wedge E, \ldots, \neg A \wedge \neg E$ is maximal {within the given bounds that the relative frequency of property A is at most 0.5, and that of E at least 0.7) for which the conditional statistical probability [HE | ■] has the greatest values. This is achieved by assuming an outcome of the thought experiment in which both the relative frequency of $A \wedge E$ and $\neg A \wedge E$ are 0.5, i.e. every film in fact turns out to be expensive, and the number of American films is maximal. For such a sequence then a relative frequency

$$[\mathrm{HE} \mid A \wedge E] \cdot 0\ 5 + [\mathrm{HE} \mid \neg A \wedge E] \cdot 0.5 = 0.45 + 0.35 = 0.8$$

of happy endings should be predicted.

Similarly, by considering an outcome of the thought experiment in which the number of expensive or American films is minimal, a lower bound of 0.64 is obtained for the expected frequency of HE.

What is the rationale for using statistical information, in the way described by these examples, for the prediction of the outcome of a thought experiment? Clearly, here a close connection between the random mechanism, realizations of which constitute our thought experiment, and the statistical probability distribution (partially) described by the statistical data must be assumed: our understanding of the random mechanism producing the toss of the die in example 2.1 is characterized by the assumption that we observed an unmanipulated toss of a fair die. In the film-example the screening of that film at that time is perceived to be a random draw from the set of all screenings of mystery films by arbitrary networks at arbitrary times.

Thus, in both examples the random mechanism used as an explanation of the chain of events producing the observed event is *equivalent* to the statistical distribution

- equivalent in the sense that when we consider an arbitrary series of realizations of the random mechanism, i.e. one in which it is not supposed that each realization supplies us with some specific evidence, then we would predict that the relative frequencies in this series agree with the statistical data.

Postulate 2: *Default reasoning about probabilities rests on the assumption that the observed event e is a realization of a random mechanism, equivalent to the statistical probability distribution.*

Postulate 2 only describes a precondition that must be fulfilled in order to combine degrees of belief with statistical information. It gives no hint whatsoever by what operational rule this combination will actually be performed.

A key observation that will be instrumental for a derivation of a specific analytical rule for this combination can be made by reconsidering the arguments used above in deriving bounds on R_1(f) and HE(/): in both cases, the predictions for the relative frequencies of these properties in the thought experiments as, respectively, 0.1 and [0.64,0.8] were obtained by only arguing from the prior beliefs derived from the evidence, and from the statistical data, but were completely independent of the evidence itself.

When from a prior subjective probability of 0.3 for *p(t),* and the statistical data available for tosses of fair dice, a degree of belief of 0.1 is derived for R_1 (t), this is done by simply considering a random sample of tosses of a die, in which the relative frequency of the property *p* happens to be 0.3. For this imaginary sample it is no longer necessary to assume that each of it's elements occurs in a setting analogous to the one; of the original toss. Similarly in the film example; assume that the scene we have seen does not provide any more relevant information with respect to the actual film / having any of the properties A, E or HE. Then, in order to predict the relative frequency of HE in the thought experiment associated with /, an arbitrary sample of mystery films with less than one half American and more than 70% expensive productions will be considered. If the original film happens to be black and white, and we have no statistical information referring to the property of being black and white, then we will not assume that every film in the random sample is black and white too, this property being recognized as irrelevant.

To obtain a more precise notion of what it means that the given evidence does not provide any more relevant information, we say that a set Ψ of degrees of belief *exhausts the evidence with respect to* L(S) if, based on the evidence alone, and without any statistical information, we are unable to assign degrees of belief to properties definable in L(S) any more specific than the ones in ty. The way in which statistical data is used to define degrees of belief now is described in a third postulate.

Postulate 3: *//Ψ is a set of degrees of belief exhausting the evidence obtained about an event c with respect to* L(S), *then the predicted frequency of a property φ €* L(S) *in the thought experiment associated with e is calculated as the expected relative frequency of the property φ in a large random sample of events, given that the relative frequencies of prope $\psi \in$ L(S) n that sample is within the bounds prescribed by Ψ.*

As before in postulate 1, it was here preferred to use the imprecise term "large sample[1]', when, in fact, we should more accurately speak about limiting frequencies as the sample size tends towards infinity.

3 The Statistical Model

To implement the rule for the derivation of degrees of belief formulated in postulate 3 in a mathematical model, the concepts of a random sample and relative frequency of a property in such a sample, which, as yet, have only been used intuitively, must be formalized.

The mathematical model for a random observation of a single event, is provided by a *random variable:* a function defined on some probability space equipped with a probability measure P, taking values in the set of possible events. Since we distinguish different events only with respect to properties definable in L(S), we may use the simpler model of a random variable taking values in T (this being the set. of equivalence classes of events with regard to these properties). Such a random variable A' now is a model of a randomly sampled event, if it's distribution is equal to the statistical probability measure u on T, i.e.

$$\mathrm{P}^X(\alpha) := \mathrm{P}(\{\omega \in \Omega \mid X(\omega) = \alpha\}) = \mu(\alpha)\ (\alpha \in \Gamma).$$

The model of a sequence of random events, initial segments of which constitute random samples of increasing size, is an infinite sequence $X_1, X_2, \ldots$ of independent Γ-valued random variables, all distributed according to μ.

For every n, the first n elements of this sequence define a new random variable P_n^X on Ω with values in $\Delta(\Gamma)$, their *empirical distribution:*

$$\mathrm{P}_n^X(\alpha) := \frac{1}{n}\sum_{i=1}^{n} 1_\alpha(X_i) \quad (\alpha \in \Gamma),$$

with 1_α the indicator variable of α, i.e. $1_\alpha(X_i(\omega)) = 1$ if $X_i(\omega) = \alpha$, and 0 else.

$\mathrm{P}_n^X(\alpha)$ thus describes the relative frequency of α in a sample of size n. What we have to investigate now is the limiting value we should expect for $\mathrm{P}_n^X(\alpha)$ as $n \to \infty$, given that P_n^X approaches a limit consistent with Ψ.

The set Ψ, primarily regarded as a collection of statements of degrees of belief, by a slight abuse of notation, may also be regarded as a subset of $\Delta(\Gamma)$: the subset of probability measures satisfying the constraints stated in Ψ. We can then define for $\delta \geq 0$:

$$\Psi(\delta) := \{\nu \in \Delta(\Gamma) \mid \exists \nu' \in \Psi \ \ |\nu - \nu'| \leq \delta\}$$

with $|\nu - \nu'|$ the Euclidean distance of ν and ν'.

We now have assembled the mathematical counterparts of most of the informal concepts of postulate 3. What still remains unexplained is the notion of an "expected frequency". Theorem 3.1 will show that, in our model, we can give a very strong formal meaning to this notion of expectation; it also provides an explicit description of the frequencies to be expected.

To prepare the theorem, we now give a brief reminder of the essential definitions regarding cross-entropy minimization.

For $\mu = (\mu_1, \ldots, \mu_N)$ with $\mu_k > 0$ $(k = 1, \ldots, N)$ and $\nu = (\nu_1, \ldots, \nu_N) \in \Delta(\Gamma)$, the cross-entropy of ν with respect to μ is defined by

$$CE(\nu, \mu) := \sum_{\substack{i \in \{1, \ldots, N\} \\ \nu_i > 0}} \nu_i \ln \frac{\nu_i}{\mu_i}.$$

To treat the general case, without the restriction to measures μ with only strictly positive components, the definition of CE must be extended in a way which makes it also attain the value ∞, and some additional considerations for the cases when this happens have to be added. $CE(\cdot, \mu)$ is a strictly convex function, so that for every closed and convex $J \subset \Delta(\Gamma)$ there exists a unique $\nu_0 \in J$ with $CE(\nu_0, \mu) < CE(\nu, \mu)$ for all $\nu \in J$, $\nu \neq \nu_0$. This ν_0 is denoted $\pi_J(\mu)$.

In the statement of theorem 3.1, for sequences $(\delta_n), (\delta'_n)$ of real numbers, we use the intuitive notation $(\delta_n) \geq (\delta'_n)$ to signify that $\delta_n \geq \delta'_n$ for all n, and $(\delta_n) \searrow 0$ to say that $\delta_n \geq 0$ for all n, and $lim_{n \to \infty} \delta_n = 0$.

Theorem 3.1 Let $X_1, X_2, \ldots$ be a sequence of independent random variables taking values in $\Gamma = \{a_1, \ \ , a_N\}$ with distribution $\mu \in \Delta(\Gamma)$ $(\mu_k > 0,\ k = 1, \ldots, N)$. Let $\Psi \subseteq \Delta(\Gamma)$ be closed and convex. Let $\nu_0 := \pi_\Psi(\mu)$. Then there exists a sequence $(\delta_n) \searrow 0$, such that for all $(\delta'_n) \geq (\delta_n)$ with $(\delta'_n) \searrow 0$, there exists $(\epsilon_n) \searrow 0$, such that

$$lim_{n \to \infty} \mathrm{P}(\ |\mathrm{P}^X_n - \nu_0| \leq \epsilon_n \mid \mathrm{P}^X_n \in \Psi(\delta'_n)) = 1. \qquad (3)$$

In a version of this theorem also allowing for measures μ with 0-components. the additional assumption must be made that $CE(\nu, \mu) < \infty$ for at least on $\nu \in \Psi$. The proof of this theorem is essentially an application of the Sanov-theorem for multinomially distributed random variables (see e.g. [Bahadur, 1971]), and, in broad outline, is similar to the proof of a related result in [van Campenhout and Cover, 1981]. The details will be given in [Jaeger, ?].

A few comments may be useful to better understand the role that in theorem 3.1 is played by the sequence (δ_n). When Ψ has interior points, then the theorem actually is true for $(\delta_n) = 0$, i.e. the whole process of approximating Ψ by the sequence $\Psi(\delta'_n)$ can be done without. On the other hand, consider $\Psi := \{\nu \in \Delta(\Gamma) \mid \nu_1 = r\}$ where r is some irrational number. Then P^X_n can never take a value in Ψ, each component of P^X_n being of the form q/n for some $q \in \mathbf{N}$. Hence, conditioning on $\{\mathrm{P}^X_n \in \Psi\}$ in (3) would mean to condition on the empty set, and the conditional probability in (3) would be undefined for every n. Moreover, for each n, P^X_n can only take on finitely many values, so that for sufficiently fast decreasing sequences $(\delta_n) > 0$, even $\{\mathrm{P}^X_n \in \Psi(\delta_n)\}$ will be empty for all n. Thus, the condition of (δ'_n) "slowly" tending to 0 in theorem 3.1 makes sure that sufficiently many possible values of P^X_n are in $\Psi(\delta'_n)$.

Theorem 3.1 provides a clear answer to what frequency of *φ E* L(S) we should expect in the random samples described by postulate 3, provided is closed and convex, as is the case when is defined by a set of sentences (2): when looking at sufficiently large samples, with a probability arbitrarily close to certainty, this relative frequency will be arbitrarily close to ψ(μ)(Φ).

For the die-example, the minimum cross-entropy solution for the given constraints and statistical distribution is (0.1,...,0.1,0.7/3,...,0.7/3). The upper bound of 0.8 derived for prob(HE(/)) in the film- example corresponds to the. minimum cross-entropy measure with respect to the statistical distribution with //.(E) = 1 and //(A | E) = 0.5. The lower bound derives from statistical measures *μ* with μ(A) = 0 and μ(E) < 0.7. The minimum cross-entropy measure μ for other statistical measures *μ* satisfying the statistical constraints of example 2.2 will yield values μ(HE) in between 0.64 and 0.8. (All these results are derivable from elementary properties of rross-entropy minimization, e.g. the axioms given in [Shore and Johnson, 1980].)

By the epistemic analysis of section 2, we obtain a good insight under what conditions (an ideal agent's) default reasoning about probabilities, when reconstructed from information given in a formal knowledge base, is adequately modeled by cross-entropy minimization: first, we must make the assumption of postulate 2, i.e. that the agent who's degrees of belief are encoded in the knowledge base considers the random mechanism he or she associates with the event e to be equivalent to the statistical probabilities stated in the knowledge base. Second, it must be assumed that the given degrees of belief exhaust the evidence, i.e. that the knowledge base reflects all the relevant information the agent has about e. Observe that this second condition is a typical idealization that always has to be made to justify application of a non-monotonic inference rule (probabilistic or logical) to a knowledge base.

4 Comparison and Conclusion

Traditionally, the meaning of degrees of belief often is defined in terms of preferences between acts (e.g. the acceptance of certain bets), the utility of which will depend on some uncertain proposition. By eliciting from a person suitable statements of preference, his or her degree of belief about the proposition can be defined by a unique (subjective) probability value. The most, influential presentation of this approach probably is [Savage, 1954]. This view of degrees of belief is stronger than the one we used here, in the sense that they are always defined to be point-valued. Nevertheless, the two definitions ere not incompatible: the thought experiment explanation focuses on how an agent arrives at a degree of belief without, trying to prescribe a method by which unique values will always result. The preference-paradigm concentrates on the measurement of degrees of belief, which can very well be imagined to have been formed by a thought, experiment.

Shafer and Tversky [1985] speak of "mental experiments1' that are performed to obtain probability judgments. Unlike the thought experiments described by postulate 1, Shafer and Tversky's mental experiments are not an abstract epistemic model for the meaning of degrees of belief, but designate a variety of ways in which, in concrete situations, specific evidence can be compared to well-defined chances. Thus, the (mental) drawing of a random sample of events according to some known statistics, as described in postulate 3, constitutes a mental experiment in the sense of Shafer and Tversky.

Paris and Vencovska [1992] have analyzed the problem of probabilistic inference from the same kind of knowledge bases as considered here. They base their approach on the semantic interpretation that a subjective probability represented by prob(0(e)) in fact describes a statistical probability *[*Φ| S<]: the statistical probability of Φ in the ideal reference class S_r of elements that are "similar" to e. As a natural consequence of this view, there is little room for the distinction of different types of inference rules made in section 1: since essentially we are left with only one type of probabilities, there is only room for inference rules to be applied simultaneously to degrees of belief and statistics.

Paris and Vencovska show that when entropy-maximization is applied to their knowledge bases (which must also include a clause stating that $[S_e]$ is small), then the effect of the general statistical information on the inferences made about the specific statistical terms $[\Phi \mid S_c]$ is defined by cross-entropy minimization. Together with a justification of the maximum entropy method ([Paris and Vencovska, 1990]), this provides a justification for minimum cross-entropy inferences. This derivation of the minimum cross-entropy principle, however, is of a completely different nature than the one presented here, because the justification of the maximum-entropy method is based on logical arguments alone (just as in the well known work by Shore and Johnson [1980]): it, is shown that if an inference process satisfies certain logical principles, i.e. behaves adequately when applied to knowledge bases of certain syntactic structures, then it will have to be entropy maximization.

An argument of this kind can only be used to show that cross-entropy minimization is the adequate formalism for default reasoning about probabilities when it is taken for granted that at least one such formal process exists an assumption that in itself is not corroborated by an axiomatic derivation. It, might, very well be that there are other axioms that are intuitively reasonable for default reasoning about, probabilities, but are not satisfied by the minimum cross-entropy principle. In that case we would have to conclude that no completely adequate formal process exists. For this reason it has here been attempted to elucidate the process of the formation of degrees of belief based on statistical information in human reasoning by looking at its epistemic basis rather than by giving a normative (partial) description of its behaviour. It was then shown that with the unfolding interpretation of a degree of belief as a prediction of the outcome of a thought experiment, the reasoning process itself can be captured in a statistical model validating minimum cross-entropy reasoning. Such a derivation of the minimum cross-entropy principle from a semantic model provides valuable evidence that it does, in fact, not have counterintuitive logical properties, since these would have to correspond to flaws in the semantic, model.

References

[Bacchus *et al,* 1992] F. Bacchus, A. Grove, J.Y. Halpern, and D. Roller. From statistics to beliefs. In *Proc. of National Conference on Artificial Intelligence (A A A I- 92),* 1992.

[Bacchus *et ai,* 1994] F. Bacchus, A. Grove, J.Y. Halpern, and D. Roller. Generating new beliefs from old. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence.* Morgan Raufmann, San Francisco, CA, 1994.

[Bacchus, 1990] F. Bacchus. *He-presenting and Reasoning With Probabilistic Knowledge.* MIT Press, 1990.

[Bahadur, 1971] R.R. Bahadur. *Some limit theorems in statistics.* CBMS-NSF Regional Conference Series in Applied Mathematics; 4. SIAM, Philadelphia PA, 1971.

[Carnap, 1950] R. Carnap. *Logical Foundations of Probability.* The University of Chicago Press, 1950.

[Halpern, 1990] J.Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence,* 46:311-350, 1990.

[Jaeger, ?] M. Jaeger. Phd thesis. In preparation.

[Jaeger, 1994a] M. Jaeger. A logic for default reasoning about probabilities. In *Proceedings of the Tenth, Conference on Uncertainty in Artificial Intelligence.* Morgan Kaufmann, San Francisco, CA, 1994.

[Jaeger, 1994b] M. Jaeger. Probabilistic reasoning in terminological logics. In *Principles of Knowledge Representation an Reasoning: Proceedings of the Fourth International Conference (KR.94).* Morgan Raufmann, San Francisco, CA, 1994.

[Paris and Vencovska, 1990] J.B. Paris and A. Vencovska. A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning,* 4:183-223, 1990.

[Paris and Vencovska, 1992] J.B. Paris and A. Vencovska. A method for updating that justifies minimum cross entropy. *International Journal of Approximate Reasoning,* 7:1-18, 1992.

[Savage, 1954] L. J. Savage. *The Foundations of Statistics.* Wiley, New York, 1954.

[Shafer and Tversky, 1985] G. Shafer and A. Tversky. Languages and designs for probability judgment. *Cognitive Science,* 9:309 339, 1985.

[Shore and Johnson, 1980] 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('l):26-37, 1980.

[van Campenhout and Cover, 1981] J. M. van Campenhout and T. M. Cover. Maximum entropy and conditional probability. *IEEE Transactions on Information Theory,* IT-27(4):483-489, 1981.