# Taking the Sting out of Subjective Probability

We analyze the subjective (Bayesian) view of probability by giving an analogue of 'truth conditions' for subjective probabilities: we define what has to be the case in a domain in order for the subjective assessment *the probability of event $E$ is $p$* to be accurate in the domain. Our analysis is radically different from existing approaches, in that we assume from the beginning that subjective probabilities can be *useful* (in that they lead to good predictions concerning some aspects of a domain) yet by no means *true* or *optimal* (in that they do not lead to good predictions concerning all aspects of a domain). We show how, by taking this fact into consideration, some of the problems associated with subjective probabilities disappear.

## 1 Introduction

In the Bayesian approach to reasoning under uncertainty, probability is viewed as a subjective notion. While the Bayesian approach has strong theoretical foundations (see, e.g. De Finetti (1974); Savage (1954)) and typically performs well in practice, several aspects of the interpretation of subjective probability remain controversial and problematic. In order to deal with such concerns, one may attempt to define the *meaning* of subjective probabilities in non-probabilistic terms. In this paper, we give such a 'meaning'. Our proposal differs markedly from existing ones in that it explicitly accounts for subjective probabilities that lead to good predictions of only *some*, not *all* aspects of a domain. In this way we manage to avoid some of the problematic aspects of (the traditional interpretations of) subjective probability. Specifically, we argue that traditional use of subjective probability sometimes leads to overly strong inferences. This seems to be an intuitively-felt con-

*The author is currently supported by a TALENT-grant awarded by the Netherlands Organization for Scientific Research (NWO). He wishes to thank Phil Dawid, Jorma Rissanen, Paul Vitányi and Johan van Benthem for encouragement and stimulating discussions pertaining to this research.

Submitted for
*Proceedings of LLC8*
D. Barker-Plummer, D. Beaver, J. van Benthem, P. Scotto di Luzio (eds.)

cern[1] of many researchers, especially those who are more accustomed to logic-based approaches to uncertainty such as non-monotonic logics (Brewka, 1991). The goal of this paper is not only to propose a new interpretation of subjective probability, but also to show, in a language accessible to such researchers, that (a) the problem of making overly strong inferences can be rigorously shown to exist (see Example 6), but that (b) this problem can in principle be solved, too.

**The Meaning of 'the Meaning of Subjective Probability'** Two classic approaches (due to Savage (1954) and De Finetti (1974)) define the meaning of subjective probability by giving a precise answer to the following question:

**Behavior Question** What does the statement 'Agent's subjective distribution is $P$' entail in terms of *behavior* of the Agent if the Agent's task is to make decisions and/or predictions about the domain under consideration?

A different, less well-known approach is to try to define a probabilistic analogue of what a logician might call 'truth conditions': in the field of mathematical logic, one needs a precise definition of 'what has to be the case in a domain in order for the sentence $\Gamma$ to be true in that domain'. In a theory of subjective probability, one would like a precise definition of 'what has to be the case in a domain in order for the subjective distribution $P$ to be *accurate* in that domain'. In other words,

**Fundamental Question** What does the statement 'Agent's subjective distribution $P$ is *accurate*' imply for the domain under consideration?

In this paper we define a 'meaning' of subjective probability by giving a precise answer to the Fundamental Question. We use this answer to arrive at an answer for the Behavior Question. Our answer to the Fundamental Question is a modification of an existing approach based on the notion of *calibration* (Dawid, 1982, 1985). It leads to an answer to the Behavior Question that is in some ways radically different from existing answers. The reason is that we take as our starting point the following idea:

**Wrong-Yet-Useful (WYU) Principle** In many situations, the subjective probabilities of an agent will be very *useful* (in that they

[1] communicated to this author by several AI Researchers at several conferences; see also the discussion in (Cheeseman, 1988).

lead to good predictions or decisions concerning some aspects of the domain) yet by no means (universally) *true* or *correct* or *optimal* (in that they do not lead to good predictions concerning some other aspects of the domain and/or another subjective probability distribution would generally lead to better predictions).

In contrast to most other authors we take this realistic assumption explicitly into account in our answer to the Fundamental Question. This leads to the notion of a *safe agent*. We will show that 'safe agents' refuse to make overly strong inferences.

**Contents of This Paper** Sections 2 and 3 are written at an informal level. Section 2 introduces subjective probability. Section 3 introduces 'calibration' and, using this concept, indicates in exactly what way subjective probabilities may lead to overly strong inferences. Section 4 and 5 introduce the mathematics needed to appreciate our meaning of probability, and Sections 6 and 7 discuss this meaning and how it avoids overly strong inferences. We end with a conclusion in which we examine whether we have really 'taken the sting out of' subjective probability.

## 2 Subjective Probability and its Uses

Here we give two examples within the context of a domain in which subjective probabilities are routinely used. The examples will hopefully show (a) the *usefulness* of subjective probability and (b) the fact that people, at least in some cases, more or less know how to interpret subjective probabilities (both (a) and (b) seem to be disputed by some statisticians, cf. Cheeseman (1988)).

**Example 1 [the professional weather forecaster]**
Consider a weather forecaster who on each day $i$ ($i \in \{1,2,3,\ldots\}$) has to predict whether or not it will rain on day $i+1$. Let $X_i \in \{0,1\}$ indicate whether it rains ($X_i = 1$) on day $i$ or not ($X_i = 0$). The forecaster's predictions are probabilistic, that is, on each day $i$ the forecaster will declare 'the probability that it will rain tomorrow is $p_{i+1}$', $p_{i+1}$ being some number between 0 and 1. To arrive at these predictions, the weather forecaster may use some complicated model of 'weather physics'; in general, the predictions $p_{i+1}$ may depend on knowledge the weather forecaster has about the past. Hence $p_{i+1}$ will be a function of (1) the outcomes of $X_1, \ldots, X_i$, and (2) the outcomes of some external variables we will call $Z_1, \ldots, Z_i$; the value of $Z_j$ summarizes information available to the weather forecaster about (for example) local air

pressure, temperature and various other observable quantities that are relevant to the weather at day $j+1$, but are observable before day $j+1$.

How should we interpret the weather forecaster's predictions? The predictions $p_1, \ldots, p_i$ are better not viewed as 'true' probabilities or even as 'approximations' to true probabilities (Dawid, 1982): first of all, it is not at all clear whether a 'true' distribution $P^*$ according to which the $X_i$ are distributed can even be assumed to exist. Second, even if we accept the existence of such a distribution, then the $p_i$ may lead to quite good predictions even if they are completely different from $p_i^*$, the 'true' probabilities that it will rain on day $i$ given the values of $X_1, \ldots, X_{i-1}$ and $Z_1, \ldots, Z_i$: the weather forecaster will most likely use some reasonable, but far from 'optimal' or 'correct' model of weather physics (this will be further illustrated in Example 2 below). It may thus be a better idea to regard $p_{i+1}$ as the weather forecaster's *subjective* probability, his 'degree of belief' that it will rain on day $i+1$ given his knowledge about the domain ('weather physics') and about the values of $X_1, \ldots, X_i$ and $Z_1, \ldots Z_i$.

**From Individual Predictions to Subjective Probability Distributions** It turns out that without loss of generality, we may describe the forecaster's prediction scheme as one single probability distribution defined over sequences of outcomes of arbitrary length. Let us give some rough idea of how this works. Note that each prediction $p_{i+1}$ is actually a *function* of $(x_1, z_1), \ldots, (x_i, z_i)$, where $(x_i, z_i)$ are the actually observed values of $X_i$ and $Z_i$ on day $i$. This suggests interpreting the $p_i$'s as *conditional* distributions, $p_{i+1}$ being 'the probability that it rains on day $i+1$ conditional on $X_1 = x_1, Z_1 = z_1, \ldots, X_n = x_i, Z_i = z_i$'. In our subjective interpretation, this can be read as 'the weather forecaster's degree of belief that it rains on day $i+1$ given (that he has observed) that $X_1 = x_1, Z_1 = z_1, \ldots, X_n = x_i, Z_i = z_i$. By applying the definition of conditional probability,, one verifies the following: for each choice of prediction functions $p_1, p_2, \ldots$, there exists a unique distribution $P$ (defined over all sequences of arbitrary length) such that the conditional probabilities coincide with the $p_i$, i.e.

$$P(X_{i+1} = 1 \mid X_1 = x_1; Z_1 = z_1; \ldots; X_i = x_i; Z_i = z_i) = p_{i+1}. \quad (1.1)$$

for all $i$, and all values $(x_1, z_1), \ldots, (x_i, z_i)$ in the range of $(X_1, Z_1), \ldots, (X_i, Z_i)$. Conversely, for each distribution $P$ defined over sequences of outcomes of arbitrary length, there exists a unique set of functions $p_1$ (a constant), $p_2$ (a function of the outcome of $X_1$ and $Z_1$, $p_3$ (a function of the outcome of $X_1, X_2, Z_1$ and $Z_2$) etc. such that (1.1) holds for all $i$ and all values $(x_1, z_1), \ldots, (x_i, z_i)$.

We can do this in general: with any Agent who, at each time step $i$, uses some (computable) function of the values that variables $X_i$ and $Z_i$ in the domain took at times $1, \ldots, i$ to arrive at a probability distribution $P_{i+1}$ over the variable $X_{i+1}$, we can associate a single probability distribution $P_{Agent}$ that 'encodes' all the probabilistic predictions made by Agent. It is not hard to prove this formally; for details, see (Dawid, 1985, Section 3) who calls the distribution $P_{Agent}$ a *probability forecasting system*. We will simply call $P_{Agent}$ the Agent's *subjective probability distribution* for the domain under consideration[2].

**The Need for Subjective Probability** The example exhibits a domain for which probabilities that are best seen as 'subjective' are routinely used. Moreover, most people (not just mathematicians!) have some intuitive idea about what the weather forecaster's predictions 'mean', i.e. what they tell us about the world (in many countries, the television's weather forecast is indeed formulated in terms of 'precipitation probability'). The following example makes clear that such subjective distributions may continue to be useful even if other subjective distributions would lead to much better predictions, thereby illustrating our Wrong-Yet-Useful Principle (Section 1).

**Example 2 [the naive weather forecaster]** Imagine a person who has lived in the same city for years and who has observed that, on average, it rained about 30% of the time. If asked: 'what is the probability of rain tomorrow' the person may answer that it is about 0.3. Assuming that the mechanisms that underly the region's weather do not change quickly over time, this prediction will be in a certain sense be correct or at least accurate, even though it provides much less information than the prediction made by the professional weather forecaster. The exact sense in which this prediction is accurate will be explained in the next section.

## 3 Calibration

We have just seen that subjective distributions will often be wrong-yet-useful. Using the concept of *calibration*, we will now indicate how

[2]The reader may object that this view of subjective probability is too narrow, since one can imagine domains in which the predictions at times $i, i+1, \ldots$ concern events in different sample spaces, some of which are of an inherently unrepetitive nature. Our results can be trivially extended to this case; it is only for ease of exposition that we stick to the 'repeated trials' case (Dawid, 1985).

this leads to a serious problem for subjective probability. Calibration is a means of determining the accuracy of a subjective distribution for a given sample. It has been extensively studied by A.P. Dawid (1982, 1985). For a modern treatment as well as additional references, we refer to Kalai et al. (1999). We give an informal introduction by means of an example.

**Example 3** Imagine that the weather forecaster of Example 1 always uses discretized probabilities. For concreteness let $p_i \in \Delta$ where $\Delta = \{0.1, 0.2, 0.3, \ldots, 0.9\}$. Consider the period from day 1 to day $n$, where $n$ is some large number. Intuitively, when we say that the forecaster is 'calibrated' within this period we mean that on those days $i$ for which the forecast was $p_i = 0.1$, it rained approximately 10% of the time; on those days for which it was $p_i = 0.2$, it rained approximately 20% of the time etc.

A forecaster can be calibrated without being in any sense 'optimal'. For example, in scenarios in which it rains approximately 30% of the time, our naive weather forecaster will automatically be calibrated even though his predictions are much less precise than those of a calibrated professional weather forecaster.

**Example 4** [**a problem for subjective probability: the power of probability calculus**] We are now in a position to give a first, informal account of a serious problem for subjective probability. Note that the naive weather forecaster predicts $p_{i+1} = 1/3$ on *each* day $i$; both if it rained on day $i$ ($X_i = 1$) and if it did not ($X_i = 0$). Of course, the frequency of rain on those days which follow a day on which it rained will in general be quite different from the frequency of rain on those days which follow a day on which it did not rain. So, the predictions of the naive forecaster may be quite accurate (calibrated) if we look at a large set of *consecutive* days; yet if we split this set into several subsets, the predictions may be quite bad (uncalibrated, i.e. probabilities diverging from frequencies) for *all* of these subsets. Of course, the naive weather forecaster may be aware of this, and he may not *want* his probabilistic predictions to be used for predicting conditional events like 'the frequency of rain on day $i+1$ given that it rained on day $i$'. However, in conventional theories of subjective probability, an agent who uses subjective probability has *no means of specifying what his probabilities can be used for and what not.*

This example shows that if Agent uses subjective probabilities that are useful, but only for limited prediction tasks, then, in order to avoid bad predictions, it should be exactly clear (to the Agent) what these limited prediction tasks are. In a sense, probability calculus is too *powerful* to handle this situation: by the construct described earlier, a sequence of individual predictions $p_i$ automatically induces a unique full subjective distribution $P_{Agent}$ over all sorts of outcomes. These can in principle be used to arrive at predictions of the value of *any* random variable that can theoretically be defined over the domain, conditioned on the value of *any* other random variable that can theoretically be defined over the domain. In the simple example above, it would be easy to prevent this problem by simply stating that 'the probabilities $p_{i+1}$ should only be used to unconditionally predict the weather on day $i+1$'; the problem does not seem very serious here. But now let us return to the professional weather forecaster: even if he is a good forecaster, then his predictions, too, will most likely only be accurate for some random variables, not for all. For example, if the weather forecaster is asked to predict the weather on day $i + 5$ based only on information available to him on day $i$, the predictions may not be so good any more. But in this 'professional' case, it is not at all a priori clear what predictions that can be made by $P_{Agent}$ are trustworthy and what not. Dawid (1982) has proven an interesting theorem (see Example 6) which shows that *this problem cannot be avoided in common theories of subjective probability.* The cornerstone of our own theory will be to develop a formal means of deducing, for each subjective distribution $P_{Agent}$ what it can be reliably used for and what not. We will now start preparing for this.

## 4 Mathematical Preliminaries and Notation

We will restrict ourselves to agents that do not take into account the value of the external variables $Z_i$ when making predictions. Strictly speaking, this rules out agents like the professional weather forecaster of Example 1. The definitions and theorems we are to develop can be trivially modified to deal with this case too – at the expense of some additional notation. For ease of exposition we avoid this notation and neglect $Z_i$. Hence all our domains concern repeated realizations ('trials') of a random variable $X$ taking values in some finite set $\Omega = \{1, \dots, k\}$. A sample $x^n = (x_1, \dots, x_n)$ is then a sequence of $n$ realizations of the domain; each $x_i \in \Omega$. We define $\Omega^n$ to be the $n$-fold Cartesian product of $\Omega$. $\Omega^0 := \{\lambda\}$ (a special symbol denoting the 'empty sequence') and $\Omega^* := \cup_{i \geq 0} \Omega^i$. $\Omega^\infty$ is the set of infinite sequences of elements of $\Omega$. An agent that makes probability forecasts about the domain uses a probability distribution $P_{Agent}$. In some of our results, we assume the number of forecasts to be made to be bounded in advance by some number $n$; in other results it is unbounded. In the bounded case, $P_{Agent}$ may be thought of as a distribution over $\Omega^n$ or alternatively as the marginal distribution over $\Omega^n$ of a $P_{Agent}$ that is defined for the unbounded case. In the unbounded case, $P_{Agent}$ is a distribution over[3] $\Omega^\infty$. Conditional probabilities, random variables and expectations are all defined as usual.

We identify each distribution $P$ over $\Omega = \{1, \ldots, k\}$ with its probability mass function, i.e. the vector $(P(1), \ldots, P(k))^T$: if we write $P = r$ for some $r \in \mathbf{R}^k$, this means $(P(1), \ldots, P(k))^T = r$. We will compare such probability vectors to frequency vectors, which are averages of so-called *indicator vectors.* An indicator vector for sample space $\Omega$ is a vector-valued random variable $I : \Omega \to \mathbf{R}^k$ such that for $j \in \Omega$, $I(j)$ is the $k$-dimensional vector with a 1 in the $j$-th component and a 0 everywhere else. Analogously, for a sequence $x^n$ with $x_i = j$ we define $I_i(x^n)$ to be the vector with a 1 in the $j$-th component and a 0 everywhere else.

## 5 Calibration Defined

Let $P$ be a distribution over $\Omega^n$ and let $\Omega = \{1, \ldots, k\}$. The *prediction set* $\Delta$ *of* $P$ *relative to sample* $x^n$ is given by

$$\Delta := \{r \in \mathbf{R}^k \mid P(\cdot|x_1, \ldots, x_i)) = r \text{ for some } 0 \le i < n \ \}.$$

The prediction set is just the set of all different $k$-component probability vectors $p_{i+1} = P(\cdot|(x_1, \ldots, x_i))$ that are actually issued if $P$ is used for prediction on sample $x^n = (x_1, \ldots, x_n)$. For example, in Example 2, $\Delta = \{(0.3, 0.7)^T\}$. The prediction set is used to partition the set of time-points $\{1, \ldots, n\}$ into the sets $\mathcal{J}_r$ (one for each $r \in \Delta$) of time-points for which the same prediction was issued. Formally, for all $r \in \Delta$,

$$\mathcal{J}_r := \{i \in \{1, \ldots, n\} \mid P(\cdot|x_0, \ldots, x_{i-1}) = r\}. \tag{1.2}$$

We use $\#\mathcal{J}$ to indicate the number of elements of the set $\mathcal{J}$ and we use $|\cdot|$ for $\ell_1$-distance: for $v \in \mathbf{R}^k$, $|(v_1, \ldots, v_k)^T| := \sum_{i=1}^k |v_i|$.

**Definition 1 (calibration, simple case)** *Let $x^n$ be a given sample in $\Omega^n$ and let $P_{Agent}$ be an Agent's subjective distribution over $\Omega^n$ with prediction set $\Delta$ (relative to $x^n$). Let $\epsilon > 0$ and $I_i$ and $\mathcal{J}_r$ as defined*

[3]By this we mean that $P_{Agent}$ is a probability measure over $\Omega^\infty$ endowed with the $\sigma$-algebra generated by random variables $X_1, X_2, \ldots$.

*above. We say that $P$ is $\epsilon$*-calibrated on sample $x^n$ *iff*

$$\sum_{\{r \in \Delta \mid \mathcal{J}_r \neq \emptyset\}} |r - \frac{\sum_{i \in \mathcal{J}_r} I_i(x^n)}{\#\mathcal{J}_r}| \leq \epsilon, \tag{1.3}$$

This expresses that for all predictions $r$ that are ever issued, the deviation between the probabilities $(r_1, \ldots, r_k)$ and the frequencies of actual outcomes observed directly after prediction $r$ was made is small: the total deviation, summed over all different $r$, must be smaller than $\epsilon$.

**Selection Rules and Test Sets** We now extend the definition to the more interesting case where only a subsample of $x^n$ is used to measure calibratedness (Example 4). To this end we introduce 'selection rules' and 'test sets', constructs that have their roots in Von Mises' notion of 'collectives' (Dawid, 1985). Formally, a *test set selection rule* is a function $\psi : \Omega^* \to \{0, 1\}$. The *test set* $\mathcal{T}_\psi$ selected by selection rule $\psi$ for sample $x^n$ is the set

$$\mathcal{T}_\psi := \{i \in \{1, \ldots, n\} \mid \psi(x_1, \ldots, x_{i-1}) = 1\} \tag{1.4}$$

The interpretation is as follows: the test set $\mathcal{T}_\psi$ is the subset of time points $i$ on which $P_{Agent}$'s predictions are to be tested. That is, we want to measure whether $P_{Agent}$ is calibrated on the set of outcomes $\{x_i | i \in \mathcal{T}_\psi\}$; we call this set the *subsample of $x^n$ selected by $\psi$.*

**Definition 2 (calibration on a subsample)** *Let $x^n$, $P_{Agent}$ and $\Delta$ and $I_i$ and $\mathcal{J}_r$ and $\epsilon$ be as before. We say that $P_{Agent}$ is $\epsilon$*-calibrated on the subsample of $x^n$ selected by $\psi$ *iff*

$$\sum_{\{r \in \Delta \mid \mathcal{J}_r \cap \mathcal{T}_\psi \neq \emptyset\}} |r - \frac{\sum_{i \in \mathcal{J}_r \cap \mathcal{T}_\psi} I_i(x^n)}{\#(\mathcal{J}_r \cap \mathcal{T}_\psi)}| \leq \epsilon \tag{1.5}$$

We often abbreviate '$P$ is $\epsilon$-calibrated on the subsample of $x^n$ selected by $\psi$' to '$P$ is $\epsilon$-calibrated on test set $\mathcal{T}_\psi$ or even '$P$ is $\epsilon$-calibrated for $\psi$'. We note that our definition is *not* equivalent to either the definition of calibration given by Dawid (1982) or more modern definitions as in Kalai et al. (1999). There is no space here to go into the (somewhat subtle) differences.

**Example 5 [the naive weather forecaster, cont.]** Let us formalize the problem of Example 4. Consider a sample $x^n$ such that the frequency of rainy days is $f = 1/3$ if the full sample $x^n$ is taken into account, whereas the conditional frequencies are given by:

$$\begin{aligned} f_0 &= \frac{\#\{i \in \{2, \ldots, n\} \mid X_i = 1; X_{i-1} = 0\}}{\#\{i \in \{2, \ldots, n\} \mid X_{i-1} = 0\}} = 1/6 + \delta, \\ f_1 &= \frac{\#\{i \in \{2, \ldots, n\} \mid X_i = 1; X_{i-1} = 1\}}{\#\{i \in \{2, \ldots, n\} \mid X_{i-1} = 1\}} = 2/3 + \delta. \quad (1.6) \end{aligned}$$

Assume $\delta$ is very small (it is easy to show that samples $x^n$ exist satisfying (1.6) for arbitrarily small $\delta > 0$). The naive weather forecaster who predicts $P(X_{i+1} = 1|x^i) = 1/3$ at each day $i$ will be $\epsilon$-calibrated for the full sample $x^n$ for all $\epsilon \geq 0$. In contrast he will not be calibrated on either of the two subsamples of $x^n$ selected by test set selection rules $\psi_0$ and $\psi_1$ defined, respectively, by

$$\begin{aligned} \psi_0(\lambda) = 0; \psi_0((x_1, \ldots, x_i)) = |1 - x_i| \\ \psi_1(\lambda) = 0; \psi_1((x_1, \ldots, x_i)) = x_i \end{aligned} \quad (1.7)$$

On the subsample selected by $\psi_0$ ($\psi_1$), the frequency of rain is given by $f_0$ ($f_1$) rather than $f$.

**Example 6** [**a problem for subjective probability**] Using the calibration concept, Dawid (1982) proved the following theorem which implies that the problem mentioned in Example 4 is of a very general nature. Here we only state a special case expressed using the notation and terminology developed in this paper.

**Theorem (Dawid 1982)** *Define, for test set rule $\psi$, the event $\mathbf{M}_\psi \subseteq \Omega^\infty$ as follows:*

$$\begin{aligned} \mathbf{M}_\psi = \{(x_1, x_2, x_3, \ldots) \mid & \textit{ if } \lim_{n\to\infty} \textstyle\sum_{i=1}^n \psi(x^i) = \infty \textit{ then} \\ & \textit{for all } \epsilon > 0, \textit{ for all large enough } n, \\ P_{Agent} \textit{ is } \epsilon\textit{-calibrated on the subsample of } & (x_1, \ldots, x_n) \textit{ selected by } \psi\}. \end{aligned}$$

*Let $P_{Agent}$ be an arbitrary distribution over $\Omega^\infty$ and let $\psi$ be an arbitrary test set selection rule. Then $P_{Agent}(\mathbf{M}_\psi) = 1$.*

According to the traditional answers to the Behavior Question (Section 1), an agent that assigns high probability to an event $\mathbf{M}$ has a high ‘degree of belief’ in $\mathbf{M}$ actually taking place. This would imply that, for *every* $\psi$, *every* conceivable agent strongly believes that it will eventually (large enough sample) be very well-calibrated for the subsample selected by $\psi$. But neither our naive weather forecaster (who will fail against the functions $\psi_1$ and $\psi_0$ of Equation 1.7) nor any reasonable professional weather forecaster will truly believe that they will be well-calibrated for arbitrary $\psi$! Thus these agents - and most other agents one can think of - are self-inconsistent in that they do not believe in an event to which they assign a very high probability. In the next section, we introduce agents who do not have to suffer from this inconsistency.

## 6 The Fundamental and the Behavior Question

In this section we define a 'meaning' of subjective probability by answering the two questions raised in Section 1. The answers are based on calibration and on the concept of a *safe* agent.

**Safe Agents in Predictive Situations** A *predictive situation* is determined by a point in time $i$ and the outcomes of all random variables whose values have been observed by an agent up until time $i$, i.e. by $x^i = (x_1, \ldots, x_i)$. Two situations $s_1$ and $s_2$ are, by definition, *exchangeable*[4] for an agent if, according to the agent, they are indistinguishable for prediction purposes (this will be clarified later on).

A *safe* agent is an agent that views some of the situations in which it must make probabilistic predictions as exchangeable. It does so *not* because these predictive situations are inherently equivalent. Rather, due to lack of a priori knowledge about the domain or lack of experience with the domain or lack of computational resources, Agent *chooses* to treat them as indistinguishable, even if Agent may not be sure whether this is a good assumption (in some cases Agent may even be sure that it is not a good assumption). For example, our naive weather forecaster who views $X_1, X_2, \ldots$ all as independent random variables considers *all* predictive situations to be exchangeable – not because he really thinks that all these situations really are equivalent, but because he does not have any idea how he could *use* the different knowledge available to him in different situations to get better predictions. A safe Agent *acts as if* some situations are equivalent even though it does not want to assume that these situations are truly equivalent. We will now make this precise.

**Definition 3 (safe agent, situation selection function)** *A* safe agent *is characterized by the triple* $(\mathcal{S}_{Agent}, \chi_{Agent}, P_{Agent})$. $S_{Agent}$ *is a countable set, called the* set of predictive situations. $\chi_{Agent} : \Omega^* \rightarrow$

[4]The related notion of 'exchangeable random variables' has been studied extensively by De Finetti (1974). The notion of a 'safe' agent is, as far as we know, original.

*$\mathcal{S}_{Agent}$ is called the* situation selection rule *and $P_{Agent}$ is a distribution over $\Omega^\infty$. The pair $(S_{Agent}, \chi_{Agent})$ is called the agent's* situational structure.

Intuitively, $\mathcal{S}_{Agent}$ is the set of all possible predictive situations between which Agent distinguishes. $\chi_{Agent}$ is a function, which, for each time point $i$ and initial sample $x^i$ determines the situation Agent is in directly after having observed $x^i$ and before having to predict $X_{i+1}$.

**Example 7 [the naive and the Markov weather forecaster]** For the naive weather forecaster, $\Omega = \{0, 1\}$, $\mathcal{S}_{Agent} = \{s_1\}$ and $\forall i, \forall x^i \in \Omega^i : \chi_{Agent}(x^i) = s_1$ (the naive w.f. views all predictive situations as equivalent), and $P_{Agent}(X_{i+1} = 1 | x^i) = 1/3$.

In contrast, one may consider a Markovian weather forecaster who has observed that, over a long period of days $\{1, \ldots, n\}$, the frequency of rain depended on whether or not it rained on the previous day in the way indicated in Equation 1.6. If such a weather forecaster sets his predictive probabilities equal to his observed frequencies, then $S_{Agent} = \{s_0, s_1\}$; $\chi_{Agent}((x_1, \ldots, x_i)) = s_{x_i}$ and, for $i > 1$, for all $x^i \in \Omega^i$, $P_{Agent}(X_{i+1} = 1 | (x_1, \ldots, x_i)) = f_{x_i}$ with $f_{x_i}$ as in Equation 1.6.

**An Interpretation of Subjective Probability for 'Safe' Agents**
Why should an Agent be willing to make predictions using some distribution $P_{Agent}$ in the first place? Because it expects that its predictions will, in some sense, be accurate. Here, we will assume that this means that the Agent expects to be well-calibrated on a large enough sample (the reader who is suspicious of this idea is referred to Section 7, where it is shown that for safe agents, a quite different approach based on likelihood leads to essentially the same results).

But note that since our Agent is 'safe', it does not expect to be well-calibrated on a large sample $x^i$ under *all* test set selection rules $\psi$; for example, the 'safe' naive weather forecaster does not expect to be calibrated under the rules given by Equation 1.7. The question arises of what test set selection rules $\psi$ are permissible and what not. If we can identify the set of permissible $\psi$, then we have an answer to our Fundamental Question: we will simply say that a distribution $P_{Agent}$ is $\epsilon$-accurate for a sample $x^n$ iff it is $\epsilon$-calibrated for *all* subsamples of $x^n$ that can be selected by the 'permissible' $\psi$. Careful analysis of the previous examples suggests the following definition of 'permissible' selection rules: an Agent should be $\epsilon$-calibrated 'within each situation it distinguishes'. For example, the naive w.f. should be calibrated on the *full* sample $x^n$, because at *each* time point $i$ it is in the same situation $s_1$. The Markov w.f. should be calibrated for the test set $\{i | x_{i-1} = 1\}$

(consisting of *all* days on which it is in situation $s_1$) and the test set $\{i | x_{i-1} = 0\}$ (consisting of *all* days on which it is in situation $s_0$). This is achieved by the following definition:

**Answer to the Fundamental Question** Consider a safe agent $(\mathcal{S}_{Agent}, \chi_{Agent}, P_{Agent})$. For all $s \in \mathcal{S}_{Agent}$, define $\psi_s : \Omega^* \rightarrow \mathbf{R}$ to be the selection rule satisfying for all $0 \leq i < n$,

$$\psi_s(x^i) = 1 \Leftrightarrow \chi_{Agent}(x^i) = s. \tag{1.8}$$

That is, time-point $i+1$ is part of the test set $\mathcal{T}_{\psi_s}$ iff at time $i$, Agent was in situation $s$. *We define Agent to be $\epsilon$-accurate for sample $x^n$ iff, for all $s \in \mathcal{S}_{Agent}$, Agent is $\epsilon$-calibrated with respect to the subsample selected by $\psi_s$.*

Our answer to the Fundamental Question leads to an answer to the Behavior Question as follows: we already argued that Agent, if it is willing to use $P_{Agent}$ at all, expects that it will be $\epsilon$-calibrated with respect to all the $\psi_s$ (as defined in Equation 1.8). But if $P_{Agent}$ is $\epsilon$-calibrated for all $\psi_s$, then it can be logically deduced that it is also $\epsilon$-calibrated with respect to several more complicated selection rules $\psi$ with, for all $s$, $\psi \neq \psi_s$. Hence, a logically consistent agent should also expect to be $\epsilon$-calibrated for these more complicated selection rules. This motivates our answer to the Behavior Question:

**Answer to the Behavior Question, informally stated** Let $\psi : \Omega^* \rightarrow \{0, 1\}$ be some function. Agent should be willing to make predictions about random variable $\tau(X_{i+1})$ given that $\psi(x^i) = 1$ if and *only* if *for all $n$, $x^n$, from the premise that Agent is $\epsilon$-accurate with respect to $x^n$ it logically follows that Agent is $\epsilon$-calibrated with respect to the subsample of $x^n$ selected by $\psi$.* Since Agent expects the premise to hold, it should also expect the conclusion to hold. To make this answer precise, we need to characterize the selection rules $\psi$ for which an $\epsilon$-accurate Agent is calibrated. This is done in the following theorem:

**Theorem 1** *Let Agent be 'safe' with structure $(\mathcal{S}_{Agent}, \chi_{Agent})$. For given sample $x^n$, let $\mathcal{V}_\epsilon$ be the class of $\epsilon$-accurate probability distributions for this sample. Then for all $\epsilon > 0$, $\mathcal{V}_\epsilon$ contains at least one element. Moreover, (1) for all $P_{Agent} \in \mathcal{V}_\epsilon$, $P_{Agent}$ is $\epsilon$-calibrated for* all *test set selection rules $\psi$ that are '$\chi_{Agent}$-measurable', i.e. for which there exists a function $g : \mathcal{S}_{Agent} \rightarrow \{0, 1\}$ satisfying for all $0 \leq i < n$,*

$$\psi(x^i) = g(\chi_{Agent}(x^i)). \tag{1.9}$$

*(2) Let $\psi$ be such that there exists no $g$ that satisfies (1.9) for every $i \geq 0$, every $x^i \in \Omega^i$. Then there exist $n$, $x^n$ such that for small enough $\epsilon$, there exists a $P_{Agent} \in \mathcal{V}_\epsilon$ such that $P_{Agent}$ is* not *$\epsilon$-calibrated with respect to the subsample selected by $\psi$.*

The theorem is proved in the Appendix. Note that (1) and (2) are not quite each other's converse: (1) requires $g$ to exist for the actual sample (rather than all samples); (2) requires there is a sample for which $g$ does not exist. The reason for the asymmetry is that Agent may sometimes be *lucky* on the actual sample $x^n$; for example, the naive weather forecaster with $\chi_{Agent}(x^n) = s_1$ for all $n, x^n$ *may* turn out to be $\epsilon$-calibrated on the test set $\{i|x_{i-1} = 1\}$ for the particular sample $x^n$ even though this will not be the case for *all* samples for which the naive weather forecaster is $\epsilon$-accurate. This potential unluckiness is built into the following definition:

**Precise Answer to the Behavior Question** Let $\psi : \Omega^* \rightarrow \{0, 1\}$ be some function. Suppose a safe Agent $(\mathcal{S}_{Agent}, \chi_{Agent}, P_{Agent})$ observed an initial sample $x^l$ such that $\psi(x^l) = 1$. Agent is asked to predict the value that a function $\tau : \Omega \rightarrow \mathbf{R}$ assumes at the $(l+1)$st point in time. Then (a) either there is a $g$ such that (1.9) holds for all $i$, all samples $x^i$, or (b) no such $g$ exists. In case (a), Agent should predict $\tau(X_{l+1})$ by the value $E_{P_{Agent}(\cdot|x^l)}[\tau(X_{l+1})]$. In case (b), Agent should refuse to make any prediction.

**Example 8 [The naive and Markov forecasters, cont.]** The answer implies that the naive w.f. should refuse to make any predictions which are to be tested on a proper subset (such as for example $\{x_i|x_{i-1} = 1\}$) of the full sample. Concerning the Markov w.f., we intuitively expect an $\epsilon$-accurate forecaster to be not only $\epsilon$-calibrated on the subsamples $\{x_i|x_{i-1} = 1\}$ and $\{x_i|x_{i-1} = 0\}$, but also *on the full sample $x^n$*. That this is indeed the case is not directly clear from Equation 1.8, but it can easily be deduced from Equation 1.9.

Our answer to the Behavior Question avoids the problem indicated by Dawid's theorem (Example 6) since, in general, Agent does not take its own $P_{Agent}$ seriously for some $\psi$. On the other hand, we note that 'safety' is not necessarily a restriction: an Agent who *insists* on being prepared to make all kinds of predictions in all kinds of circumstances, may choose $\mathcal{S}_{Agent} = \Omega^*$ and $\chi_{Agent}(x^n) = x^n$ for all $n$. For such an Agent, a function $g$ satisfying (1.9) will always exist, so Agent is allowed to predict the value of $\tau(X_i)$ no matter what rule $\psi$ is used.

## 7 Accuracy in Terms of Likelihood

We can also assess the accuracy or truthfulness of a subjective distribution $P_{Agent}$ for a given sample in a manner quite different from calibration, namely by looking at how well $P_{Agent}$ *fits* the sample, that is, how probable ('likely') the sample is according to $P_{Agent}$. This measure of accuracy lies at the basis of many statistical procedures including 'Bayesian statistics' (Berger, 1985; Savage, 1954; De Finetti, 1974; Dawid, 1982), the statistical theory preferred by adherents of subjective probability. We show in our main theorem (Theorem 2) that, *for safe agents*, the new measure of accuracy leads to essentially the same answer to the Fundamental Question as the calibration criterion we used in the previous section. More specifically, for large samples, Bayesian statistical procedures tend to prefer models which are accurate according to the likelihood criterion. Theorem 2 shows that, for safe agents, these coincide with the models that are accurate in the sense defined in the previous section. This suggests that standard Bayesian methods, without any modification, can successfully be applied for statistical inference for safe agents. We regard this as evidence that our notion of 'safe agent' is a reasonable one.

**Definition 4** *For a given family of distributions $\mathcal{C}$ over $\Omega^\infty$, we say that $P' \in \mathcal{C}$ is* $\epsilon$-log-likelihood-optimal *within $\mathcal{C}$ for sample $x^n$ if*

$$\sup_{P \in \mathcal{C}} \{\log P(x^n)\} - \log P'(x^n) \leq \epsilon n.$$

If we want to determine whether $P$ is $\epsilon$-accurate (in the likelihood sense) for sample $x^n$, it would be too much to require that $P$ maximizes the likelihood within the class of *all* distributions $P$ over $\Omega^\infty$: since for each sample $x^n$ there exist $P$ such that $P(x^n) = 1$, this would turn both the naive and the professional w.f. into very inaccurate agents. Luckily, it turns out that the notion of a 'safe' agent as defined in the previous section leads to a natural definition of the class $\mathcal{C}$ of distributions under consideration.

Let $\Omega = \{1, \ldots, k\}$. Let Agent be 'safe' with situational structure $(\mathcal{S}_{Agent}, \chi_{Agent})$. Assume Agent *always* predicts the next $X_i \in \Omega$ using the same distribution whenever Agent is in the same predictive situation. Let $\mathcal{C}$ be the set of distributions over $\Omega^n$ that are available to such an Agent. Formally, $\mathcal{C}$ consists exactly of those $P_{Agent}$ that are '$\chi_{Agent}$-measurable', i.e. for which there exists a function $f : \mathcal{S}_{Agent} \rightarrow \mathbf{R}^k$ such that for all $i, 0 \leq i < n$, all $x^i$,

$$P_{Agent}(\cdot | x^i) = f(\chi_{Agent}(x^i)). \tag{1.10}$$

**Example 9** Consider the naive and the Markov weather forecasters of Example 7. For the naive w.f., $\mathcal{C} = \{P \mid \exists r \in [0,1] \ : \ \forall i, 0 \leq i < n, x^i \in \Omega^i \ : \ P(X_{i+1} = 1|x^i) = r\}$. For the Markov w.f., $\mathcal{C} = \{P \mid \exists r_1, r_2 \in [0,1] \ : \ \forall i, 0 \leq i < n, x^i \in \Omega^i \ : \ P(X_{i+1} = 1|x^i) = r_j \Leftrightarrow \chi_{Agent}(x^i) = s_j\}$.

For $P_{Agent}$ and sample $x^n$, let $p_{\min}$ be the smallest probability ever assigned by $P_{Agent}$ to any outcome, and let $n_{\min}$ be the situation in $\mathcal{S}_{Agent}$ that was encountered least often in $x^n$:

$$p_{\min} = \min_{0 \leq i < n} \ \min_{1 \leq j \leq k} \{P_{Agent}(X_{i+1} = j \mid x^i)\}$$
$$n_{\min} = \min_{\{s \in \mathcal{S}_{Agent} \mid \exists i < n : \chi_{Agent}(x^i) = s\}} \#\{0 \leq i < n \mid \chi_{Agent}(x^i) = s\}.$$

Theorem 2 says (item 1) that if $p_{\min} > 0$ (i.e. all the probabilities issued by Agent are bounded away from 0), then $\epsilon$-accurateness of $P_{Agent}$ implies $\delta$-log-likelihood optimality such that $\epsilon \to 0 \Rightarrow \delta \to 0$. If $n_{\min}/n > 0$ (i.e. frequencies of all outcomes in all situations are bounded away from 0), then (item 2) the converse holds too.

**Theorem 2** *Let $(\mathcal{S}_{Agent}, \chi_{Agent}, P_{Agent})$ characterize a safe Agent and let $x^n \in \Omega^n$ be a sample. We have (1) if $P_{Agent}$ is $\epsilon$-accurate for $x^n$ then $P_{Agent}$ is $p_{\min}^{-1}\epsilon^2$-log-likelihood-optimal for $x^n$ within the class $\mathcal{C}$ defined above. (2) If $P_{Agent}$ is $\epsilon$-log-likelihood optimal for $x^n$ within the class $\mathcal{C}$ then $P_{Agent}$ is $(n_{\min}/n)^{-1/2}\epsilon^{1/2}$-accurate for $x^n$.*

## 8 Conclusion; Did we really take the sting out of Subjective Probability?

The developments in this paper can be used to solve several problematic aspects of subjective probability. On the other hand, if subjective probability is to be used as a general means of dealing with uncertainty in arbitrary domains, then several additional problems pop up. Below we summarize what our proposal implies for some of the most well-known problematic aspects of subjective probability.

**What we achieved** We have developed the notion of a 'safe agent'. We have given a 'meaning' of subjective probability based on these safe agents and shown how our 'meaning' avoids the problem of making overly strong inferences. While Dawid's theorem (Example 6) shows that this problem is practically unavoidable in common theories of subjective probability, we showed that our 'safe agents' do not suffer from it. Apart from (hopefully) persuading those who are skeptical about the use of subjective probabilities that some of its inherent problems can be overcome, this work has at least two other applications we did not discuss in the main text, to be discussed in a future paper. These are (a) a 'safer' means of using subjective probabilities based on limited experience (i.e. learned from small samples); (b) dealing with *computational constraints*: Sometimes, given enough time and/or memory, an Agent could come up with a quite accurate subjective distribution $P_1$. In practice, Agent may be forced to use an alternative distribution $P_2$ which is less precise but much easier to compute. Our work shows how to come up with 'simple' $P_2$'s that are nevertheless safe to use.

**What we did not achieve** There are (at least) two aspects of subjective probability which are considered (at least by some people) problematic and on which our proposal does not have any direct bearing. These are (1) *use of 'qualitative' Subjective Probabilities* (Brewka, 1991) for which no (not even approximate) numerical values can be given; (2) *use of Subjective Probabilities over non-observable quantities* (i.e. probability distributions defined over spaces of probability distributions), as is customary in Bayesian statistics.

**What may be achieved – future work**
*1. Use of Subjective Probabilities in the presence of ignorance.* Our 'safe' agents are well-behaved in some situations involving ignorance (e.g. lack of knowledge to distinguish between two different situations). But some questions remain to be solved, for example: if an Agent has no idea whether $x_1$ is more or less likely than $x_2$, does it make sense to assign the *same* probability to $x_1$ and $x_2$? Relatedly, our 'safe' agents sometimes refuse to make predictions. What if they cannot afford this and they *have* to make a prediction in a situation in which they would rather refuse to? Preliminary work (Grünwald, 2000) suggests that our proposal can be extended to deal with these questions, too and makes a connection to the *Maximum Entropy Principle.*

*2. Dawid's Challenge.* Dawid (1982) focuses on a related but different consequence of his theorem (Example 6) than the one discussed here. In a nutshell, he argues that it follows from his theorem that there are (realistic!) situations in which a Bayesian (subjectivist) statistician may entertain a model that is very badly calibrated with respect to the available data; yet the Bayesian is quite confident that his model will be accurate (calibrated) on future data from the same source. This problem is alleviated, though by no means completely solved if only 'safe' predictions are allowed; the details are left for a future paper.

*3. Learning the Situational Structure.* Can an Agent *learn* from a given set of 'training data' a good situational structure $(\mathcal{S}_{Agent}, \chi_{Agent})$? We would like to have a means for Agent to decide whether the currently chosen structure is or is not a good one. Even if it is a good one, Agent should be able to decide whether there does not exist a better one, distinguishing between more different situations, being calibrated in each of these, and therefore leading to more precise predictions in a larger variety of situations. In order to achieve this we can apply a standard tool of Bayesian statistics: assume that Agent has predetermined a (countable) set of 'a priori conceivable situational structures' $(\mathcal{S}_1, \chi_1), (\mathcal{S}_2, \chi_2), \ldots$. We assume that Agent assigns each of them a *prior probability* $\pi((\mathcal{S}_i, \chi_i))$. For example, $(\mathcal{S}_i, \chi_i)$ may correspond to the structure entertained by an order-$(i-1)$-Markovian weather forecaster (Example 7), so that $(\mathcal{S}_1, \chi_1)$ is the naive weather forecaster's structure. Agent could then use standard Bayesian techniques (i.e. Maximum A Posteriori Probability) to determine, given sample $(x_1, \ldots, x_n)$, what structure is (currently) the best hypothesis to use. Readers familiar with Bayesian statistics (Berger, 1985) will recognize that this will automatically implement a form of 'Occam's Razor', preferring situational structures that strike a trade-off between simplicity (small number of situations) and likelihood (which, by Theorem 2, can be related to calibratedness). There is a connection to the Minimum Description Length Principle here (Barron and Cover, 1991). Working out the details of this is a major goal for future work.

## 9 Appendix

**Proof (sketch) of Theorem 1** Let $\mathcal{S}_0 = \{s \in \mathcal{S}_{Agent} \mid \chi_{Agent}(x^i) = s$ for some $0 \leq i < n$. Let $\mathcal{K}_s = \{i \in \{0, \ldots, n-1\} | \chi_{Agent}(x^i) = s\}$. Let, for $s \in \mathcal{S}_0$, $\delta_s \in \mathbf{R}^k$ be a vector such that $\sum_{1 \leq j \leq k} \delta_{s,j} = 0$. Let, for $s \in \mathcal{S}_0$, $P_s(\cdot)$ be the distribution over $\Omega$ with, for $1 \leq j \leq k$,

$$P_s(j) := \frac{\#\{i \in \mathcal{K}_s | x_{i+1} = j\} + \delta_{s,j}}{\#\mathcal{K}_s}. \tag{1.11}$$

**Proposition 1** *Let for all $0 \leq i < n$, $P_{Agent}(\cdot|x^i) := P_s$ for the $s$ satisfying $s = \chi_{Agent}(x^i)$. $P_{Agent}$ is $\epsilon$-accurate for all $\epsilon \geq \#\mathcal{S}_0 \cdot (\sum_{s \in \mathcal{S}_0} |\delta_s|)$.*

The proposition (proof trivial & omitted) shows that for all $\epsilon > 0$, $\mathcal{V}_\epsilon$ contains at least one element. For item (1): assume $P_{Agent}$ is $\epsilon$-accurate. Then by definition it is $\epsilon$-calibrated for the selection rules $\psi_s$

(Equation 1.8). Let $\mathcal{T}_{\mathcal{S}_g} = \bigcup_{s \in \mathcal{S}_g} \mathcal{T}_{\psi_s}$ where $\mathcal{S}_g := \{s \in \mathcal{S}_{Agent} \mid g(s) = 1\}$. Applying Definition 2 gives that $P_{Agent}$ is $\epsilon$-calibrated for $\mathcal{T}_{\mathcal{S}_g}$. But $\mathcal{T}_{\mathcal{S}_g} = \bigcup_{s \in \mathcal{S}_g} \mathcal{T}_{\psi_s} = \mathcal{T}_\psi$ (where $\psi$ is given by Equation 1.9). This is what we had to prove.

Item (2): from the non-existence of $g$ it follows that there is at least one $s_0 \in \mathcal{S}_0$ such that for some $n$, $x^n$, $\mathcal{T}_1 = \{i \in \{0, \ldots, n-1\} \mid \psi(x^{i-1}) = 1\} \cap \{i \mid \chi_{Agent}(i) = s_0\}$ and $\mathcal{T}_0 = \{i \in \{0, \ldots, n-1\} \mid \psi(x^{i-1}) = 0\} \cap \{i \mid \chi_{Agent}(i) = s_0\}$ are both non-empty. It is now not hard to show (again using the non-existence of $g$; we omit details) that there exists at least one $s_0 \in \mathcal{S}_0$ and an $x^n \in \Omega^n$ such that

$$\frac{\sum_{i \in \mathcal{T}_0} I_i(x^n)}{\#\mathcal{T}_0} \neq \frac{\sum_{i \in \mathcal{T}_0 \cup \mathcal{T}_1} I_i(x^n)}{\#\mathcal{T}_0 + \#\mathcal{T}_1} \tag{1.12}$$

Define $P_{Agent}$ by $P_{Agent}(\cdot|x^i) := P_s$ for the $s$ satisfying $s = \chi_{Agent}(x^i)$ where $P_s$ given by (1.11). By Proposition 1 it is possible, for every $\epsilon > 0$, to choose the vectors $\delta_s$ such that $P_{Agent}$ is $\epsilon$-accurate while at the same time, for all $s_1, s_2 \in \mathcal{S}_0$ we have $s_1 \neq s_2 \Rightarrow P_{s_1} \neq P_{s_2}$. Applying Definition 2 we can see from (1.12) that if we choose $\epsilon$ small enough, $P_{Agent}$ is not $\epsilon$-calibrated for the subsample selected by rule $\psi$, while, for all $\epsilon > 0$, $P_{Agent}$ is $\epsilon$-accurate. This is what we had to prove.

**Proof (sketch) of Theorem 2** The *relative entropy* or *Kullback-Leibler (KL) Divergence* (Cover and Thomas, 1991) between distributions $P$ and $Q$ is defined by $D(P||Q) := \sum_{x \in \Omega} P(x) \ln(P(x)/Q(x))$. As shown in Barron and Cover (1991), we have

$$\sum_{x \in \Omega} (P(x) - Q(x))^2 \leq D(P||Q) \leq \sum_{x \in \Omega} \frac{(P(x) - Q(x))^2}{Q(x)} \tag{1.13}$$

Let $Q \in \mathcal{C}$. Let $\mathcal{S}_0$ and $\mathcal{K}_s$ be as in the proof of Theorem 1. By Equation 1.10, $Q(\cdot|x^i)$ only depends on the situation $s \in \mathcal{S}_{Agent}$ with $s = \chi_{Agent}(x^i)$. Hence we can define distributions $Q_s$ over $\Omega$ (one for each $s \in \mathcal{S}_0$) such that $Q(\cdot|x_i) = Q_s$ iff $\chi_{Agent}(x^i) = s$. Then

$$\log Q(x^n) = \sum_{s \in \mathcal{S}_0} \sum_{i \in \mathcal{K}_s} \log Q_s(x_{i+1}). \tag{1.14}$$

By setting the derivative of (1.14) to 0 we see that $\sup_{P \in \mathcal{C}} \log P(x^n)$ is attained for (and only for) those $P$ with for all $s \in \mathcal{S}_0$ and for $1 \leq j \leq k$, $P_s(j)$ given by (1.11) with for all $s, j : \delta_{s,j} = 0$. Let $P$ be a distribution attaining this supremum and let $Q$ be an arbitrary distribution in C. Some basic rewriting of (1.14) and (1.11) gives us that $Q$ is $\epsilon$-log-likelihood optimal iff

$$\sum_{s \in \mathcal{S}_0} \frac{\#\mathcal{K}_s}{n} D(P_s || Q_s) \leq \epsilon. \tag{1.15}$$

Both items (1) and (2) of the theorem now follow by (1.15) and (1.13) together with some basic rewriting of the definition of $\epsilon$-accuracy.

## References

Barron, A. and T. Cover. 1991. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054.

Berger, J. 1985. *Statistical Decision Theory and Bayesian Analysis.* Springer Series in Statistics. Springer-Verlag, New York, revised and expanded second edition.

Brewka, G. 1991. *Nonmonotonic Reasoning: Logical Foundations of Commonsense*, volume 12 of *Cambridge Tracts in Theoretical Computer Science.* Cambridge University Press.

Cheeseman, P. 1988. An inquiry into computer understanding. *Computational Intelligence*, 4(1):58–66,67–129. Discussion: pages 67–129.

Cover, T. and J. Thomas. 1991. *Elements of Information Theory.* Wiley Interscience, New York.

Dawid, A. 1982. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605–611. Discussion: pages 611–613.

Dawid, A. 1985. Calibration-based empirical probability. *The Annals of Statistics*, 13(4):1251–1273. Discussion: pages 1273–1285.

De Finetti, B. 1974. *Theory of Probability. A critical introductory treatment.* John Wiley & Sons, London.

Grünwald, P. D. 2000. Maximum entropy and the glasses you are looking through. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI 2000).* Morgan Kaufmann Publishers.

Kalai, E., E. Lehrer, and R. Smorodinsky. 1999. Calibrated forecasting and merging. *Games and Economic Behaviour*, 29:151–169.

Savage, L. 1954. *The Foundations of Statistics.* Dover Publications.