# Probabilistic Abduction Without Priors

**Didier Dubois**
IRIT-CNRS
Université Paul Sabatier
Toulouse France

**Angelo Gilio**
Università di Roma "La Sapienza"
Roma, Italy

**Gabriele Kern-Isberner**
Dept. of Computer Science
University of Dortmund
Dortmund, Germany

## Abstract

This paper considers the simple problem of abduction in the framework of Bayes theorem, when the prior probability of the hypothesis is not available, either because there are no statistical data on which to rely on, or simply because a human expert is reluctant to provide a subjective assessment of this prior probability. The problem remains an open issue since a simple sensitivity analysis on the value of the unknown prior yields empty results. This paper tries to survey and comment on various solutions to this problem: the use of likelihood functions (as in classical statistics), the use of information principles like maximal entropy, Shapley value, maximum likelihood. We also study the problem in the setting of de Finetti coherence approach, which does not exclude conditioning on contingent events with zero probability. We show that the ad hoc likelihood function method, that can be reinterpreted in terms of possibility theory, is consistent with most other formal approaches. However, the maximal entropy solution is significantly different. Yet, it depends on likelihoods, and its form resembles the Shapley value.

## 1 INTRODUCTION

Consider the basic problem of Bayesian abduction: Let $H$ be a Boolean proposition interpreted as a hypothesis, a disease, a fault, a cause, etc. pertaining to the state of a system. Let $E$ be another proposition representing a hypothetically observed (that is, observable) fact, a symptom, an alarm, an effect, etc. Numerical assessments of positive conditional probability values $P(E|H) = a$ and $P(E|H^c) = b \leq a$ are supplied by an agent, who either uses available statistical data or proposes purely subjective assessments. The problem is to evaluate the relative plausibility of the hypothesis and its negation after observing event $E$. If a prior probability $P(H)$ is assigned and $b > 0$, the question is solved by Bayes theorem. But suppose no prior probability $P(H)$ is assigned and observation $E$ is made, or that probabilities $a$ or $b$ are set to zero. What can be said about the support given to hypotheses $H$ vs. $H^c$ upon observing $E$?

The aim of this note is to review past proposals for dealing with this problem and propose either new solutions or rigorous formalization of previously proposed solutions, in connection with various approaches to probability theory, and to imprecise probabilities, such as maximum entropy, Shapley value, conditional events, de Finetti's coherence setting, possibility theory and the like.

Here, by definition, we do not take for granted the Bayesian credo according to which whatever their state of knowledge, rational agents should produce a prior probability. Indeed the idea that point probability functions should be in one to one correspondence with belief states means that a probability degree IS equated to a degree of belief. Then, in case of total ignorance about $H$ agents should assign equal probabilities to $H$ and its complement, due to symmetry arguments. This claim can be challenged, and was challenged by many scholars (e.g., (Shafer 1976) (Dubois and Prade 1990), (Smets and Kennes 1994) (Walley 1991): Indeed agents must assign equal probabilities to $H$ and its complement, when they know that the occurrence of $H$ is driven by a genuine random process, and when they know nothing. The two epistemic states are different but result in the same probability assessment. Here, we take ignorance about $H$ for granted, assuming $P(H)$ is unspecified (in other words the agent refuses to bet on a value of $P(H)$), and see what was done in the past, what can legitimately be done to cope with ignorance, and how to formally justify various solutions to this problem.

Dagstuhl Seminar Proceedings 05321
Belief Change in Rational Agents: Perspectives from Artificial Intelligence, Philosophy, and Economics
http://drops.dagstuhl.de/opus/volltexte/2005/328

This paper is organized as follows: We put the problem into formal terms in the next section, and recall two classical approaches for its solution in section 3. In section 4, various information principles are applied to solve the problem, and compared with each other. Section 5 presents a novel maximal likelihood approach by taking the conditional probabilities as kind of mid-values. Section 6 concludes the paper with a summary and an outlook on further work.

## 2 FORMALIZING THE PROBLEM

In the whole paper, the following notations are adopted: $\Omega$ is the sure event, $AB$ is short for $A \wedge B$ (conjunction), and the complement of an event $A$ is denoted $A^c$. Moreover, we use the same symbol to denote an event and its indicator. The basic variables in the problem are denoted

$$x = P(EH)\,; y = P(E^cH)\,; z = P(EH^c)\,; t = P(E^cH^c)$$

Let $\mathcal{P} = \{P, P(E|H) = a, P(E|H^c) = b\}$ be the set of probability functions described by the constraints expressing the available knowledge. The variables $x, y, z, t$ are thus linked by the following constraints:

$$\begin{array}{l} x + y + z + t = 1 \text{ (normalization)},\\ x = a(x + y) \text{ corresponding to } \ P(E|H) = a,\\ z = b(z + t) \text{ corresponding to } \ P(E|H^c) = b. \end{array}$$

The set $\mathcal{P}$ is clearly a segment on a straight line in a 4-dimensional space $(x, y, z, t)$, namely, the intersection of three hyperplanes.

In the most general case, assuming $0 < b < 1$, the constraints can be written

$$x = \frac{a}{1-b}\,(1-b-t)\,, y = \frac{1-a}{1-b}\,(1-b-t)\,, z = \frac{b}{1-b}\,t\,,$$

with $0 \leq t \leq 1 - b$. Then, the set $\mathcal{P}$ is the segment bounded by the probabilities $(a, 1-a, 0, 0)$ and $(0, 0, b, 1-b)$. It can be checked that this result still holds when $a = b = 1$.

Note that $P(E|H), P(E|H^c)$ can be viewed as generic knowledge (sometimes interpreted causally) expressing the probabilities of observing events of the form $E$ in general when $H$ occurs or its contrary occurs. Then these probabilities refer to a population of situations where the occurrence of events of the form $E$ was checked when $H$ was present or absent. This population may be explicitly known (as in statistics) or not (for instance we know that birds fly but the concerned population of birds is ill-defined). On the contrary, the observation $E$ is contingent, it pertains to the current situation, and nothing is then assumed on the probability of occurrence of events of the form $E$ in the population. So it is not legitimate to interpret the observation $E$ as a (new) constraint $P(E) = 1$, which would mean that events of the form $E$ are always the case, while we just want to represent the fact that event $E$ has been observed now.

Suppose the prior probability $P(H)$ is provided by an agent. Clearly it must be interpreted in a generic way (in general events of the form $H$ have this propensity to be present) otherwise, if $P(H)$ were only the contingent belief of the agent now, one may not be able to use it on the same grounds as the conditional probabilities so as to uniquely define a probability function in $\mathcal{P}$ (since we do not interpret the contingent but sure observation $E$ as having probability 1). As a consequence, when the prior probability $P(H)$ is specified, our generic knowledge also includes the posterior probability $P(H|E)$, which we extract for the reference class $E$ (as we know the current situation is in the class of situations where $E$ is true). In a second (inductive) step, the value $P(H|E)$ can be used by the agents for measuring their belief in the hypothesis $H$ to be present now, given that $E$ is observed.

An objection to the above remark can be as follows: suppose that the agent interprets $P(E|H)$, $P(E|H^c)$ as contingent conditional belief degrees of observing $E$ if $H$ is present or not present in the current. In that case, since these values are interpreted as contingent uncertain beliefs, one may be tempted to interpret the observation of evidence in a strong way, as $P(E) = 1$, especially in the case where the prior probability of $H$ is unknown. Unfortunately, the equality $P(E) = 1$ is inconsistent with $P(E|H) = a$ and $P(E|H^c) = b$ since they imply $a \geq P(E) \geq b$. So the formal framework cannot support the interpretation of $P(E|H), P(E|H^c)$ as contingent conditional belief degrees.

## 3 TWO STANDARD APPROACHES

In the literature, two approaches exist that try to cope with ignorance of the prior probability. The first approach is based on varying the prior probability on the expression of $P(H|E)$ derived from Bayes theorem. Another classical approach in non-Bayesian statistics relies on the relative values of $P(E|H)$ and $P(E|H^c)$ interpreted as the likelihood of $H$ and its complement. In this approach, the idea of computing a posterior probability is given up.

### 3.1 IMPRECISE BAYES

The most obvious thing to do in the absence of prior is to perform sensitivity analysis on Bayes theorem. Let $P(H) = p$ be an unknown parameter. Then

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} = \frac{a \cdot p}{a \cdot p + b \cdot (1-p)} \cdot$$

But the value $p$ is anywhere between 0 and 1. Clearly the corresponding range of $P(H|E)$ is $[0,1]$. So this approach brings no information on the plausibility of the hypothesis, making the observation of evidence and the presence of the generic knowledge useless. This is rather counter-intuitive, as one feels prone to consider that evidence $E$ should confirm $H$ if for instance $a$ is high and $b$ is low. The above analysis presupposes $a \cdot p + b \cdot (1-p) \neq 0$.

Two cases result in $a \cdot p + b \cdot (1-p) = 0$. First the case when $P(E|H^c) = b = 0$ and $P(H) = p = 0$ (the case when $P(E|H) = a = 0$ implies $b = 0$ by construction); finally the case when $a = b = 0$, while $p > 0$. It can be checked that in these cases, the range of $P(H|E)$ still remains $[0,1]$

Notice that the direct reasoning above in general may be risky, because we are not sure of considering all (explicit or implicit) constraints. Therefore, the case when some conditioning events may have zero probability must be treated in the coherence framework of de Finetti.

The consistency of conditional probability assessments can be checked by a geometrical approach (see, e.g., (Gilio 1990) and (Gilio 1995)), or considering suitable sequences of probability functions (see, e.g., (Coletti and Scozzafava 2002)). The approach goes as follows on the (first) case: $P(E|H) = a > 0$; $P(E|H^c) = 0$; $P(H) = 0$. Assign probability vector $\mathbf{p} = (a, 0, 0, \gamma)$ to the family of conditional events $\mathcal{F} = \{E|H\,,\ E|H^c\,,\ H|\Omega\,,\ H|E\}$. To check coherence of $\mathbf{p}$, we have to consider the "constituents" (interpretations) generated by $\mathcal{F}$ and contained in the disjunction of the conditioning events $H \vee H^c \vee \Omega \vee E = \Omega$ (here, the sure event). They are $C_1 = EH\,,\ C_2 = E^cH\,,\ C_3 = EH^c\,,\ C_4 = E^cH^c$ Let $Q_h = (q_{h1}, \ldots, q_{hn})\,,$ denote the following 3-valued assignment induced by constituent $C_h$, to the $n$ conditional events in the family $\mathcal{F}$ :

$$q_{hj} = \begin{cases} 1, & \text{if } C_h \subseteq E_jH_j, \\ 0, & \text{if } C_h \subseteq E_j^cH_j, \\ p_j, & \text{if } C_h \subseteq H_j^c. \end{cases} \qquad (1)$$

Then, in geometrical terms, we introduce the points $Q_1 = (1,0,1,1)\,, Q_2 = (0,0,1,\gamma)\,, Q_3 = (a,1,0,0)\,, Q_4 = (a,0,0,\gamma)\,,$ and, denoting by $\mathcal{I}$ the convex hull of $Q_1, \ldots, Q_4$, we must check the (necessary) coherence condition $\mathbf{p} \in \mathcal{I}$. It amounts to checking the solvability of the linear system :

$$\begin{cases} x = a(x+y),\ \ z = 0,\ \ x + y = 0,\ \ x = \gamma(x+z), \\ x + y + z + t = 1,\ x \geq 0,\ y \geq 0,\ z \geq 0,\ t \geq 0. \end{cases}$$

The solution to this system, $(x, y, z, t) = (0, 0, 0, 1)$ is a probability function on the set of constituents $\{C_1, C_2, C_3, C_4\}$. The probabilities of conditioning events are $P(H) = 0, P(H^c) = 1, P(\Omega) = 1, P(E) = 0$.

Then, we must continue to check coherence on the sub-family of conditionals whose conditioning events have zero probability; that is, we now have to check the coherence of the assessment $\mathbf{p}_0 = (a, \gamma)$ on $\mathcal{F}_0 = \{E|H, H|E\}$. Constituents in $H \vee E$ are $C_1 = EH\,, C_2 = E^cH\,, C_3 = EH^c\,,$ with associated points: $Q_1 = (1,1)\,, Q_2 = (0,\gamma)\,, Q_3 = (a,0)\,.$ We can verify the condition $\mathbf{p}_0 \in \mathcal{I}_0$ ($\mathbf{p}_0$ belongs to the triangle $Q_1Q_2Q_3$) amounts to solving the linear system

$$\begin{cases} x = a(x+y),\ \ x = \gamma(x+z), \\ x + y + z = 1,\ x \geq 0,\ y \geq 0,\ z \geq 0. \end{cases}$$

It is satisfied for every $\gamma \in [0,1]$ by the vector $(x, y, z, t) = \Big(\frac{a\gamma}{a+\gamma(1-a)},\ \frac{(1-a)\gamma}{a+\gamma(1-a)},\ \frac{a(1-\gamma)}{a+\gamma(1-a)},\ 0\Big)$. The probabilities of conditioning events $H$ and $E$ are $P(H) = \frac{\gamma}{a+\gamma(1-a)} \geq 0\,,\ P(E) = \frac{a}{a+\gamma(1-a)} > 0\,.$

The set of conditioning events with zero probability is empty or equal to $\{H\}$. Hence, the assessment $\mathbf{p}_0 = (a, \gamma)$ is coherent for every $\gamma \in [0,1]$; therefore, the initial assessment $\mathbf{p} = (a, 0, 0, \gamma)$ is coherent for every $\gamma \in [0,1]$. In other words, the range of $P(H|E)$ remains $[0,1]$.

### 3.2 LIKELIHOOD APPROACH

Usual statisticians consider $P(E|H)$ to be the likelihood of $H$, $L(H)$. When $P(E|H) = 1$, $H$ is only fully plausible. When it is 0 (the probability $P(E|H^c)$ being positive) it rules out hypothesis $H$ upon observing $E$. But there is no formal justification given to the notion of likelihood, usually. However $L(H)$ can then be viewed as a degree of possibility as pointed out by (Dubois, Moral and Prade 1997): generally the quantity $P(A|B)$ is upper bounded by $\max_{x \in B} P(A|x)$ and as pointed out by (Coletti and Scozzafava 2002), if set-function $L$ is assumed to be inclusion-monotonic (as expected if we take it for granted that $L$ means likelihood), then $P(A|B) = \max_{x \in B} P(A|x)$ is the only possible choice if only $P(A|x)$ is known for all $x$.

In this sense the likelihood approach, common in non-Bayesian statistics (e.g. (Edwards 1972), (Barnett 1973)) comes down to interpreting conditional probabilities in terms of possibility theory. The quantity $P(E|H)$ can be used to eliminate assumption $H$ if it is small enough in front of $P(E|H^c)$, but knowing that $P(E|H) = 1$ is not sufficient to ascertain it.

However, as a possibility degree is also an upper probability bound, one may try to figure out if this approach

can be formally justified in this setting. Indeed, we are in a dilemma as the sensitivity approach is probabilistically founded but provides no information while the likelihood approach is informative but looks ad hoc in a probabilistic setting.

Note that the likelihood approach is also in agreement with a default Bayesian approach : in the absence of a prior probability, assume it is uniformly distributed. Then the posterior probability is $P(H|E) = \frac{a}{a+b}$, so that it is equivalent to renormalize the likelihood functions in the probabilistic style. This fact has been recurrently used to claim that the likelihood approach is like the Bayesian approach with a uniform prior.

However this is because in classical probability theory the notion of incomplete probabilistic model makes no sense. Even if the likelihood approach looks consistent with the uniform prior (Bayes) method, its meaning is radically different. The former has no pretence to compute precise posterior probabilities: results it provides are informative only if one of $a$ or $b$ is small (and not the other). The uniform prior results from applying the Laplace principle of insufficient reason. But, saying that the likelihood approach is a special case of the Bayesian approach is like saying that a uniform possibility distribution (equivalently: an unknown probability distribution) and a uniform probability distribution mean the same thing.

# 4 APPROACHES BASED ON INFORMATION PRINCIPLES

One way out of the dilemma of abduction without priors is to introduce additional information by means of default assumptions that are part of the implicit background knowledge. The idea is that in the absence of prior probability, one finds a (default) probability measure in $\mathcal{P}$ in some way, relying on principles of information faithfulness, maximal independence assumptions, or symmetry assumptions, respectively (Paris, 1994). Then the posterior beliefs of agents is dictated by the default probability thus selected. Unfortunately, as seen below the results obtained by means of the various principles are not fully consistent with each other.

## 4.1 MAXIMUM LIKELIHOOD

The maximum likelihood principle says that if an event occurred then this is because it was at the moment the most likely event. So the best probabilistic model in a given situation is the one which maximizes the probability of occurrence of the observed evidence. This principle is often used to pick a probability distribution in agreement with some data. For instance, assume we observe $k$ heads and $n-k$ tails from tossing a coin $n$ times. The probability function underlying the process is completely determined by the probability of heads, say $x$. To find the best value of $x$, one maximizes the likelihood $L(x) = P(E|x) = x^k \cdot (1-x)^{(n-k)}$, where $E =$ "$k$ heads and $n-k$ tails" and we find $x = \frac{k}{n}$. Interestingly, since $x$ completely defines the probability measure $P$ on {tail, head}, $P(E|x) = L(P)$, i.e. the likelihood of model $P$.

In our case, $E$ occurred, so it is legitimate to establish the agent's posterior (contingent) belief about $H$ assuming $P(E)$ is as large as possible under the constraints $P(E|H) = a < 1$ and $P(E|H^c) = b \leq a$. Again, in that case we interpret $P(E)$ as the likelihood of the probability function $P$ to be selected among those such that $P(E|H) = a, P(E|H^c) = b$, while the non-Bayesian statistics approach directly chooses between $H$ and $H^c$ on the basis of their likelihoods. Here we first try to select a plausible probabilistic model, with a view to solve the abduction problem in a second step.

Note that $P(E) = a \cdot p + b \cdot (1-p)$ whose maximum is $P(E) = a$, which unfortunately enforces $p = 1$. It comes down to assuming $P(H) = 1$, so that $P(H|E) = 1$, too. This is clearly too strong to be credible, even under a weak interpretation of the posterior probability ($H$ is present in the situation where $E$ was observed). However note that in this approach the constraint $P(E) = a$ is not added to mean that the probability of $E$ is indeed $a$ in the population. It just assumes that the population of realizations relevant for the current situation is the one where $E$ is as likely as possible, so that in the current situation, $\mathcal{P}$ can be restricted to $\{P \in \mathcal{P},\ P(E) \text{ is maximal}\}$.

In any case, this approach results in a dead end in the prior-free abduction problem. A way out of this difficulty is proposed in section 5.

## 4.2 MAXIMUM ENTROPY

A fairly popular informational principle is the maximization of entropy (e.g. Paris, 1994). Entropy quantifies the indeterminateness inherent to a probability distribution $P$ by $H(P) = -\sum_\omega P(\omega) \log P(\omega)$. Given a set $\mathcal{R} = \{(B_1|A_1)[x_1], \ldots, (B_n|A_n)[x_n]\}$ of probabilistic conditionals, the *principle of maximum entropy*

$$\max H(Q) = -\sum_\omega Q(\omega) \log Q(\omega)$$
$$\text{s.t. } Q \text{ is a probability distribution with } Q \models \mathcal{R}$$

solves (uniquely) the problem of representing $\mathcal{R}$ by a probability distribution without adding information unnecessarily. The resulting distribution is denoted by $ME(\mathcal{R})$. The maximal entropy solution is often interpreted as a least committed probability, in fact the one

involving maximal indeterminateness in a subsequent decision process. In fact, maximal entropy processes conditional dependencies especially faithfully, and independence between events is implemented only if no information to the contrary can be derived.

Using well-known Lagrange techniques, we may represent $ME(\mathcal{R})$ in the form

$$ME(\mathcal{R})(\omega) = \alpha_0 \prod_{\substack{1\le i\le n \\ \omega\models A_i B_i}} \alpha_i^{1-x_i} \prod_{\substack{1\le i\le n \\ \omega\models A_i B_i^c}} \alpha_i^{-x_i}, \qquad (2)$$

with the $\alpha_i$'s being exponentials of the Lagrange multipliers, one for each conditional in $\mathcal{R}$, and $\alpha_0$ simply arises as a normalizing factor. For further details, see, e.g., (Kern-Isberner 2001).

The maximum entropy solution to our problem can be computed as follows. Let $P_{me}$ be the maxent distribution in $\mathcal{P}$ and we use the notation $\alpha = \frac{a}{1-a}, \quad \beta = \frac{b}{1-b}$. Here, the probabilistic information given is represented by $\mathcal{R} = \{(E|H)[a], (E|H^c)[b]\}$, so $P_{me} = ME(\mathcal{R})$. Using equation (2) we get the following probabilities: $P_{me}(EH) = \lambda_0\lambda_a^+$, $P_{me}(E^cH) = \lambda_0\lambda_a^-$, $P_{me}(EH^c) = \lambda_0\lambda_b^+$, $P_{me}(E^cH^c) = \lambda_0\lambda_b^-$, with a normalizing constant $\lambda_0 = (\alpha^{-a}(1-a)^{-1} + \beta^{-b}(1-b)^{-1})^{-1}$, and $\lambda_a^+ = \alpha^{1-a}$, $\lambda_a^- = \alpha^{-a}$, $\lambda_b^+ = \beta^{1-b}$, $\lambda_b^- = \beta^{-b}$. Now, it immediately follows that

$$P_{me}(H|E) = \frac{\alpha^{1-a}}{\alpha^{1-a} + \beta^{1-b}}$$

and $P_{me}(H) = \frac{a^{-a}(1-a)^{a-1}}{a^{-a}(1-a)^{a-1} + b^{-b}(1-b)^{b-1}}$.

The same results can be obtained in a more direct way by observing that every probability in $\mathcal{P}$ has the form $(ka, k(1-a), (1-k)b, (1-k)(1-b))$ with $k \in [0,1]$ and choosing $k$ such that entropy is maximized.

### 4.3 SHAPLEY VALUE

The Shapley value was first proposed in cooperative game theory ((Shapley 1953)), to extract from a set of weighted coalitions of agents (a non-additive set-function), an assessment of the individual power of each agent (a probability distribution). In the theory of belief functions, it is known as the "pignistic transformation" ((Smets and Kennes 1994)). Selecting the Shapley value comes down to assuming that all probabilities in $\mathcal{P}$ are equally probable so that by symmetry the center of mass of this polyhedron can be chosen by default as the best representative probability function in this set. This is similar as replacing a solid by its center of mass for studying its kinematics. As shown above, $\mathcal{P}$ is a segment on a straight line, bounded by the probabilities $(a, 1-a, 0, 0)$ and $(0, 0, b, 1-b)$. So the Shapley value is the midpoint of this segment, i.e. $(\frac{a}{2}, \frac{1-a}{2}, \frac{b}{2}, \frac{1-b}{2})$. Under this default probability,

$$P(H|E) = \frac{a}{a+b}$$

that is, the Shapley value supplies the same response as the Bayesian approach where a uniform prior is assumed! This is not too surprising as the Shapley value can be seen as assuming a uniform metaprobability over the probability set induced by the constraints, and considering the average probability resulting from this meta-assessment. The above result suggests that assigning a uniform prior to assumptions and assuming a uniform metaprobability over the probability polygon come down to the same result.

### 4.4 COMPARATIVE DISCUSSION

Contrary to the simple form, in some sense natural, of the Shapley value, the maximum entropy solution looks hard to interpret in the problem at hand, at first glance. But there is a similarity of form between them, except that the maxent solution distorts the influences of the probabilities $a$ and $b$ by the function

$$f(x) = \left(\frac{x}{1-x}\right)^{1-x}$$

so that the maxent solution for $P(H|E)$ takes the same form as the Shapley value, after distortion, namely, $\frac{f(a)}{f(a)+f(b)}$. Alternatively, one may see the maxent solution as attaching multiplicative weights to coefficients $a$ and $b$, of the form $w(x) = x^{-x}(1-x)^{x-1}$, so that $P(H|E)$ also takes the form $\frac{a\cdot w(a)}{a\cdot w(a)+a\cdot w(b)}$. These weights can be interpreted, up to a normalization, as a default prior, assuming $P(H) = \frac{w(a)}{w(a)+w(b)}$ in the maximum entropy approach.

This makes maximum entropy more cautious, i.e. returning in general probabilities which are closer to 0.5, according to the maxent philosophy of not introducing determinateness unnecessarily. As $a$ and $b$ approach the extreme probabilities 1 resp. 0, the maxent solution approaches the Shapley value.

In fact, we have $P_{Shapley}(H|E) = P_{me}(H|E)$, if and only if $a = b$, or $a = 1-b$. In the first case, $H$ and $E$ are statistically independent, in the second case, the influence of $H$ on $E$ is symmetrical – its presence makes $E$ probable to the same extent as its absence makes it improbable, which can be understood as a generalization of logical equivalence to the probabilistic case. This reflects a strong symmetric dependence between $E$ and $H$.

What makes Shapley value bolder in the scope of maxent is that both approaches coincide only when $E$ and

$H$ are either independent, or very strongly related. In fact, $a$ (the degree of the presence of $H$) has a positive effect throughout on the probability $P(H|E)$ whereas $b$ (the degree of the absence of $H$) has a negative effect. This means that increasing $a$ or decreasing $b$ always results in an increase of $P(H|E)$ which can be explained, e.g., by assuming $H$ to be an essential cause of $E$.

As opposed to this, the maximum entropy probability processes information in a more unbiased way, i.e. without assuming either strong dependence or independence in general. But note, that when such a relationship seems plausible (in the cases $a = b$ or $a = 1 - b$), then it coincides with the Shapley value.

A further difference between Shapley value and the maxent method becomes obvious when focusing on the prior probability $P(H)$. In the Shapley approach, it is an invariant ($P_{Shapley}(H) = 0.5$, independent of $a$ and $b$). In the maxent approach, we obtain $P_{me}(H) = \frac{w(a)}{w(a)+w(b)}$ where the value of weighting function $w$ is not altered by exchanging $a$ and $1 - a$, (and $b$ and $1 - b$), and taking on values in the interval $[\frac{1}{3}, \frac{2}{3}]$ with $P_{me}(H) = 0.5$ if and only if $a = b$ or $a = 1 - b$. In other words, this weighting function shrinks the [0, 1] scale symmetrically around 0.5.

Note that $\log w(x)$ is the entropy of the probability distribution $(x, 1-x)$. So $w(x)$ represents the distance between $x$ and 0.5. So the prior probability selected by the maxent approach basically reflects the relative distances from $P(E|H)$ to 0.5, and $P(E|\neg H)$ to 0.5, regardless of their being greater or less than 0.5. For instance the cases where $a = b = 0.9$ and where $a = b = 0.1$ are treated likewise, and yield the same default prior probability.

A general comparison between the inference process based on center of mass progagation (resulting in the Shapley value) and that by applying the maxent principle was made in (Paris 1994). Paris showed that center of mass inference violates some properties that reasonable probabilistic inference processes should obey. More precisely, in general, center of mass inference can not deal appropriately with irrelevant information and with (conditional) independencies. For the problem that we focus on in this paper, however, the Shapley value seems to be as good a candidate for reasonable inference as the maximum entropy value, regarding invariance with respect to irrelevant evidence.

Overall, it seems that the maximum entropy approach is syntactically similar to both the Shapley approach (since there exists similar implicit default priors in both approaches) and the maximum likelihood approach (posterior probabilities are proportional to likelihoods or some function thereof) for solving the abduction problem.

# 5 A RELAXED MAXIMAL LIKELIHOOD APPROACH

The reason why the maximum likelihood fails is that maximizing $P(E)$ on $\mathcal{P}$ enforces $P(H) = 1$. It may mean that the available knowledge is too rigidly modelled as precise conditional probability values.
As pointed out by (de Finetti 1936), $E|H$ stands as a three-valued entity, not a Boolean one as it distinguishes between examples $EH$, counterexamples $E^cH$ and irrelevant situations $H^c$. Authors like (Goodman, Nguyen, and Walker 1991) and (Dubois and Prade 1994) have claimed that $E|H$ can be identified with the pair $(EH, EH \vee H^c)$ of events (an interval in the Boolean algebra), or with the triple $(EH, E^cH, H^c)$ that forms a partition of the universal set. And indeed (provided that $P(H) > 0$) $P(E|H)$ is a function of $P(EH)$ and $P(E \vee H^c)$. Now, it is important to realize that $E|H$ is a kind of mid-term between $EH$ and $E \vee H^c$ since $P(E \vee H^c) \geq P(E|H) \geq P(EH)$. So it makes sense to interpret the knowledge as $P(E \vee H^c) \geq a \geq P(EH)$ and $P(E \vee H) \geq b \geq P(EH^c)$, respectively. This is consistent with the original data due to the above remarks, which also show that the new formulation is a relaxation of the previous one.

According to the maximum likelihood principle, the default probability function should now be chosen such that $P(E) = x + z$ is maximal, under constraints:

$$P(E \vee H^c) \geq a \geq P(EH);\ P(E \vee H) \geq b \geq P(EH^c)$$

and we assume here a positive likelihood function $a \geq b > 0$. The problem then reads: maximize $x + z$ such that : $1 - y \geq a \geq x$; $1 \geq x + y + z \geq b \geq z$. Since $a \geq x$, $b \geq z$, $x + z \leq a + b$, the maximal value of $P(E)$ is $P^*(E) = \min(1, a + b)$.

Now there may be more than one probability measure maximizing $P(E)$. In order to compute the posterior probability, $P(H|E)$, we are led to the problem of maximizing $P(EH) = x$ subject to $1 - y \geq a \geq x$, $1 \geq x + y + z \geq b \geq z$, $x + z = \min(1, a + b)$.

**Proposition 1** *Under the conditional event approach, the maximum likelihood posterior probability, $P(H|E)$, assuming a positive likelihood function $P(E|H^c) = b \leq a = P(E|H)$, is as follows:*

*1) if $a + b \geq 1$ then $P(H|E) \in [1 - b, a]$.*

*2) $P(H|E) = \dfrac{a}{a + b}$ otherwise.*

*Proof.* When $a + b \geq 1$ then $x + z = 1$, then $y = 0$ is enforced. Hence the problem reduces to: Maximize $x$ subject to $a \geq x$, $b \geq 1 - x$. Then $x = P(EH) = P(H|E) \in [1 - b, a]$. If $a + b < 1$, then $P(E) = x + z = a + b$. From this and $a \geq x, b \geq z$, it follows directly,

that $x = a, z = b$ must hold. So, the problem reads: Maximize $P(EH) = x$ subject to $1 - y \geq a \geq x, 1 \geq a + y + b \geq b \geq a + b - x$. The latter simplifies into $1 \geq a + y + b$ and $a \geq x \geq a$, which yields $x = a$ exactly.

These results are not so surprising, even if new to our knowledge. This approach, in opposition to the ones in the previous section does not necessarily enforce a default prior. When $P(E|H)$ and $P(E|H^c)$ are large, we only find upper probabilities $P^*(H|E) = a$ and $P^*(H^c|E) = b$ (since the lower probability $P_*(H|E) = 1 - b$), which gives a rigorous foundation to the interpretation of $P(E|H)$ and $P(E|H^c)$ as being the likelihoods $L(H)$ and $L(H^c)$ respectively. It is not surprising in the light of the interpretation of $L(H)$ and $L(H^c)$ as degrees of possibility (or upper probabilities). The larger they are the less information is available on the problem. In particular when $a = b = 1$, the likelihood function is a uniform possibility distribution on $\{H, H^c\}$ that provides no information (indeed $P(E|H) = P(E|H^c) = 1$ means that both $H$ and $H^c$ are possible). We do find that in this case the observation $E$ should not inform at all about $H$ in this case, that is, we find $P(H|E) \in [0, 1]$ (total ignorance) even assuming $P(E) = 1$. If $a = b$ increase to 1, our knowledge on the posterior evolves from equal probabilities on the hypothesis and its contrary to higher order uncertainty about them, ending up with total ignorance.

On the contrary, when $P(E|H)$ and $P(E|H^c)$ are small, the maximum likelihood solution in this case is a unique probability $P(H|E) = \frac{a}{a+b}$. This is the result obtained by the Bayesian approach under uniform priors and by the Shapley value of the probability sets induced by the likelihood functions. In this case the available knowledge, under maximum likelihood assumption, is rich enough to provide much information upon observing evidence, under the maximum likelihood principle.

When one of $P(E|H)$, $P(E|H^c)$ is small, the maximum likelihood principle enables hypothesis $H^c$ to be eliminated, if $b$ is much smaller than $a$. It supplies a unique probability measure proportional to $(a, b)$ if both values are small enough.

The previous results can be framed in the setting of coherence by the following reasoning, that encompasses the case of zero probabilities.
Given two quantities $a$ and $b$ in the interval $[0, 1]$, we consider the assessment $\mathbf{p} = (x, z, \alpha, \beta, \gamma, p)$, with $x, z, \alpha, \beta, \gamma, p$ unspecified quantities, on the family $\mathcal{F} = \{EH,\ EH^c,\ E \vee H^c,\ E \vee H,\ E,\ H|E\}$, with $P(E \vee H^c) \geq a \geq P(EH)$, $P(E \vee H) \geq b \geq P(EH^c)$. Notice that $EH = EH|\Omega$, and so on. We want to obtain all the coherent values of $p$ subject to the condition that $\gamma$ is maximum. Then, we obtain an extension of the above proposition, that takes into account all cases.

**Proposition 2** *Given the probability assessment* $\mathbf{p} = (x, z, \alpha, \beta, \gamma, p)$, *with* $a \geq b$, *on the family* $\mathcal{F} = \{EH,\ EH^c,\ E \vee H^c,\ E \vee H,\ E,\ H|E\}$, *let* $[p', p'']$ *be the set of coherent values* $p$ *such that* $\gamma$ *is maximum. We have:*
*1) if* $a = b = 0$, *then* $p' = 0, p'' = 1$;
*2) if* $a > 0$, $b = 0$, *then* $p' = p'' = 1$;
*3) if* $a > 0$, $b > 0$, $a + b \geq 1$, *then* $p' = 1 - b, p'' = a$;
*4)* $a > 0$, $b > 0$, $a + b < 1$, *then* $p' = p'' = \frac{a}{a+b}$.

While confirming the previous results, the coherence approach solves three cases with zero probabilities. When $a = 0$ and $b \neq 0$ or when $a \neq 0$ and $b = 0$, one of the assumptions $H$ or its contrary are eliminated. When $a = b$, we get either $P(H|E) = P(H^c|E) = \frac{1}{2}$ if $a \in (0, \frac{1}{2})$; equal upper probabilities $a$ on $H$ and its contrary if $a > 1/2$; and the same result (total ignorance) for $a = b = 0$ as for $a = b = 1$.

This new approach to handling abduction without priors has some advantages. It reconciliates the maximum likelihood principle (that failed in section 4.1 due to an overconstrained problem) and the ad hoc likelihood-based inference of non-Bayesian statistics. But it also recovers the Shapley value and the uniform prior Bayesian approach in some situations. It confirms the possibilistic behavior of likelihood functions, being all the more uninformative as the likelihood of the hypothesis and of its complement are both close to 1. When they are both low but positive, the uniform prior Bayesian approach is recovered. When one of $a$ and $b$ is zero, then the hypothesis with zero likelihood is unsurprisingly disqualified by observing $E$. However in the case when both likelihoods are zero or one, it comes down to total ignorance about the posterior probability of the hypothesis. This approach is at odds with the maximum entropy method which treats the cases $a = b < 0.5$ and $a = b > 0.5$ likewise.

## 6 CONCLUSION

One of the traditional disputes in probability theory opposes the Bayesian approach whereby any state of knowledge can be characterized by a single probability function on the suitable space, and classical statistics where likelihood functions are often empirically estimated but subjective prior probabilities are not considered to be relevant information. The Bayesian approach to abduction has the merit of offering a complete and harmonious solution, and the price paid is, as already stressed in the past, that a full data collec-

tion is needed. The classical statistics approach may look as lacking formal foundations despite the existing rationales for this pragmatic approach. This paper has tried to put together many tools proposed in additive and non-additive probability theories so as to sort out the issue of unknown priors. As a result some light is shed on the classical statistics approach and the maximum likelihood principle, by casting them in the framework of possibility and imprecise probability theories. It also show the agreement between the use of Shapley value and the classical Bayesian assumption of uniform priors under ignorance.

The maximum entropy approach is shown to differ from the Bayesian tradition of uniform priors and the non-Bayesian approach based on likelihoods. Indeed, the selected $P(H)$ depends on the relative distance between the likelihoods of $H$ and $H^c$ and 0.5. The farther $P(E|H)$ to 0.5 compared to $P(E|H^c)$ the more informative $H$ turns out to be.

It is also shown that applying the maximum likelihood principle for picking a default prior yields an unreasonable solution. To cope with this difficulty, our relaxation of the prior-free abduction problem provides an original solution that bridges the gap between the straightforward use of likelihood functions and the assumption of a uniform prior, being more informative than the pure sensitivity analysis approach but less precise then the Bayesian, Shapley and maxent solutions, as it maintains a family of possible posterior probabilities when the likelihood functions are to high to enable any hypothesis rejection.

More work is needed to fully interpret the obtained results. In particular, a systematic comparative study of first principles underlying the Shapley value and the maximal entropy approach is certainly in order. We should also compare our results with what the imprecise probability school has to say about this problem in a more careful way. Finally, another point to study is the influence of irrelevant information on the results of the various approaches.

## References


[1] V. Barnett (1973). *Comparative Statistical Inference* . New York: J. Wiley.

[2] B. Coletti and R. Scozzafava (2002). *Probabilistic logic in a coherent setting*. Kluwer Academic Publishers.

[3] B. de Finetti (1936). La logique de la probabilité. In *Actes du Congrès Inter. de Philosophie Scientifique*, volume 4. Paris: Hermann et Cie Editions.

[4] B. de Finetti (1974). Theory of Probability, Vol. 1, Wiley, London.

[5] D. Dubois, and H. Prade (1990). Modeling uncertain and vague knowledge in possibility and evidence theories. In Shachter, R.; Levitt, T.; Kanal, L.; and Lemmer, J., eds., *Uncertainty in Artificial Intelligence 4*. Amsterdam: North-Holland. 303–318.

[6] D. Dubois, and H. Prade (1994). Conditional objects as nonmonotonic consequence relationships. *IEEE Trans. on Systems, Man and Cybernetics* 24(12):1724–1740.

[7] D. Dubois, S. Moral, and H. Prade (1997). A semantics for possibility theory based on likelihoods. *J. of Mathematical Analysis and Applications* 205:359–380.

[8] W. Edwards (1972). *Likelihood*. Cambridge, U.K.: Cambridge University Press.

[9] A. Gilio (1990). Criterio di penalizzazione e condizioni di coerenza nella valutazione soggettiva della probabilità. In *Bollettino Unione Matematica Italiana*, volume 4-B of 7, 645–660.

[10] A. Gilio (1995). Algorithms for precise and imprecise conditional probability assessments. In Coletti, G.; D.Dubois; and Scozzafava, R., eds., *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*. New York: Plenum Publ. Co. 231–254.

[11] I. Goodman, H. Nguyen, and E. Walker (1991). *Conditional Inference and Logic for Intelligent Systems: A Theory of Measure-Free Conditioning*. Amsterdam: North-Holland.

[12] G. Kern-Isberner (2001). *Conditionals in nonmonotonic reasoning and belief revision*. Springer, Lecture Notes in Artificial Intelligence LNAI 2087.

[13] J. Paris (1994). *The uncertain reasoner's companion – A mathematical perspective*. Cambridge University Press.

[14] G. Shafer (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

[15] L.S. Shapley (1953). A value for n-person games. In Kuhn and Tucker, eds., *Contributions to the Theory of Games, II*, Princeton Univ. Press, 307-317.

[16] P. Smets, and R. Kennes (1994). The transferable belief model . *Artificial Intelligenc* 66:191–234.

[17] P. Walley (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall.